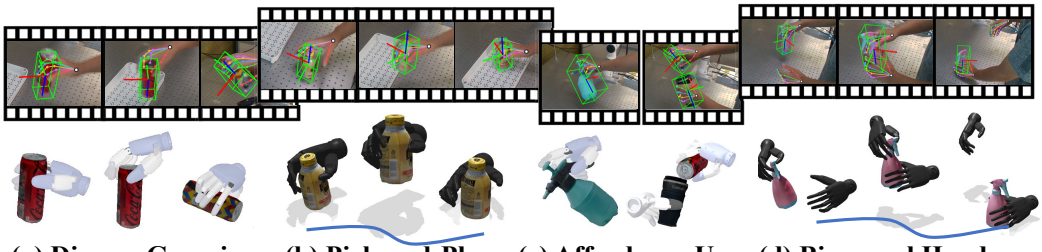
INTERACTION-CENTRIC HUMAN-TO-DEXTEROUS DEMONSTRATION CAPTURE SYSTEM AND DATASET

Anonymous authors
Paper under double-blind review



(a) Diverse Grasping **(b) Pick-and-Place** **(c) Affordance Use** **(d) Bimanual Handover**

Figure 1: **H2D** captures diverse human manipulation demonstrations, including pick-and-place, handover, and affordance-based use, with structured hand-object motion annotations. The captured demonstrations are further transformed into cross-embodiment dexterous manipulation trajectories through physics-based simulation, preserving the demonstrated hand-object interactions.

ABSTRACT

Large-scale human manipulation demonstrations offer a low-cost source of diverse data for dexterous robot manipulation, but the substantial embodiment gap between humans and robots limits their direct use for robot learning. We present an interaction-centric hand-object data capture pipeline that converts human demonstrations into dexterous manipulation data through cost-efficient acquisition and annotation. Our key insight is to leverage spatio-temporal hand-object interaction as an embodiment-transferable representation, using it to constrain motion estimation and preserve interaction relationships during human-to-robot transfer. Specifically, we introduce interaction-guided object motion estimation that exploits spatio-temporal contact plausibility to prune implausible pose hypotheses, complementing limited visual evidence under sparse-view capture and reducing the pose search space for efficient annotation. We further preserve hand-object contact relationships during simulation-based motion retargeting, transforming human demonstrations into physically plausible robot manipulation data. Experiments demonstrate several-fold improvements in object motion annotation efficiency over existing methods and sub-centimeter object pose estimation under sparse-view capture, while physics-based simulation validates interaction-preserving cross-embodiment transfer. Building on this pipeline, we construct **H2D**, comprising 5.6 hours of human demonstrations and 2,200 dexterous manipulation trajectories with structured 3D hand-object states.

1 INTRODUCTION

Large-scale human manipulation demonstrations contain rich real-world interaction knowledge and provide an attractive source for constructing dexterous robot manipulation data. Compared with robot demonstrations, which are often collected for specific tasks, environments, and robot embodiments, human demonstrations naturally cover a broader range of objects, scenes, and manipulation behaviors at low acquisition cost. However, human demonstrations cannot be directly used as data for robot manipulation learning because of the substantial differences between human and robot embodiments. A key question is therefore how to leverage human manipulation demonstrations as a source for scalable robot manipulation data construction.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Structured hand-object interaction provides an intermediate representation for bridging human demonstrations and robot manipulation data generation. Unlike visual observations, which primarily capture manipulation appearance, or human hand motions, which are specific to a particular embodiment, hand-object interactions explicitly characterize the spatial and physical coordination between the hand and object, including object motion, contact relationships, and their temporal evolution. These structured interaction states capture manipulation information beyond embodiment-specific hand motions and offer a more transferable representation across embodiments. Such structured interaction states can therefore serve as a bridge from human hand demonstrations to robot manipulation, enabling human manipulation to be adapted to a target robot while preserving physically plausible hand-object interactions.

Although human hand manipulation data can be collected at scale, the cost of acquiring and annotating their 3D interaction states can easily offset this advantage. Existing datasets based on motion capture, multi-view sensing, and automated vision-based reconstruction have reduced the reliance on manual annotation, but high-quality hand-object states often require dense visual observations and costly optimization or refinement. More cost-efficient sparse-view capture provides substantially less visual evidence for object pose estimation, making reliable object motion difficult to recover under severe hand-object occlusion, small object sizes, and weak textures. These challenges call for acquisition and annotation methods that can recover reliable interaction states from limited observations while remaining computationally efficient.

To address these challenges, we propose an interaction-centric hand-object data capture framework for human manipulation demonstrations. At its core, we develop a multi-faceted interaction representation that jointly captures contact plausibility across spatial, temporal, and semantic dimensions. These interaction constraints compensate for limited visual evidence under sparse-view capture, enabling reliable object motion recovery under severe hand-object occlusion, small object size, and weak textures. Meanwhile, contact feasibility progressively prunes implausible pose candidates, narrowing the motion search space and reducing annotation cost. Together, these designs enable cost-efficient data acquisition and efficient annotation of human manipulation demonstrations.

The recovered contact-centric representation further provides a natural interface for human-to-robot transfer. During simulation-based motion retargeting, we retain hand-object contact relationships to transform human demonstrations into robot manipulation data while preserving their essential interaction structure. Thus, hand-object interaction serves a dual role: constraining motion estimation and providing a structured representation for downstream manipulation data generation.

Experiments demonstrate several-fold higher annotation efficiency than existing methods, while achieving sub-centimeter motion estimation accuracy in cross-dataset evaluations under sparse-view settings. Physics-based simulation further validates interaction-preserving transfer across embodiments. Building on this pipeline, we construct a Human-To-Dexterous (**H2D**) dataset comprising 5.6 hours of human manipulation demonstrations and 2,200 corresponding dexterous manipulation trajectories with structured 3D hand-object states, providing a scalable resource for human-to-robot manipulation data generation.

2 Related Work

2.1 Hand-object Interaction Datasets

Existing hand-object interaction datasets can be broadly categorized by their data acquisition paradigms. **Motion capture-based datasets** provide accurate 3D hand and object states through dedicated motion capture systems (Fan et al., 2023; Banerjee et al., 2025; Zhan et al., 2024; Liu et al., 2024; Lv et al., 2024; Taheri et al., 2020; Garcia-Hernando et al., 2018), often relying on optical markers, but require specialized hardware and controlled environments, limiting their scalability and diversity. **Manually annotated datasets** (Chao et al., 2021; Liu et al., 2022) obtain 3D hand-object motion annotations without dedicated motion capture systems, but require substantial human effort and are difficult to scale. **Automated annotated datasets** (Hampali et al., 2020; Fu et al., 2025; Rim et al., 2026; Wang et al., 2025) reduce manual effort by estimating 3D hand-object motion from multi-view observations, but often require densely instrumented camera setups to provide sufficient visual coverage for occluded interactions. In contrast, **large-scale hand manipulation video datasets** (Grauman et al., 2024; Hoque et al., 2026; Ohkawa et al., 2023) provide broad cov-

erage of objects, scenes, and behaviors at low acquisition cost, but generally lack 3D hand-object motion annotations. Our work bridges this gap with a low-cost, sparse-view setup for scalable construction of 3D hand-object motion annotations.

2.2 Hand-Object Pose Estimation

Hand pose estimation has been extensively studied under full supervision, with parametric models such as MANO (Romero et al., 2017) widely used to represent articulated 3D hand motion (Huang et al., 2020; Park et al., 2022; Potamias et al., 2024; Pavlakos et al., 2024; Iqbal et al., 2018; Boukhayma et al., 2019; Moon & Lee, 2020; Lin et al., 2021a;b). Recent work explores limited-supervision settings using 2D pseudo-labels and multi-view consistency (Zheng et al., 2023; Han et al., 2026; Chen et al., 2021). Object pose estimation methods (Wen et al., 2024; 2023; Nguyen et al., 2024; Ren et al., 2026; Li et al., 2025; Labbé et al., 2022; Lin et al., 2024) recover 6D object motion from RGB or RGB-D observations, while hand-object interaction methods model their coupled motion and contact relationships (Qi et al., 2024; Chen et al., 2025; Hampali et al., 2022; Hasson et al., 2020; 2019; Liu et al., 2025; Hampali et al., 2020; Chen et al., 2023; 2022).

Building on these advances, automated hand-object annotation pipelines recover initial hand and object motion from visual observations and refine them through optimization, often using multi-view consistency and hand-object contact constraints (Fu et al., 2025; Rim et al., 2026; Wang et al., 2025). However, these methods typically rely on dense multi-view observations and sequence-level optimization, limiting annotation efficiency and scalability. In contrast, our method incorporates spatiotemporal hand-object interaction directly into a feed-forward motion annotation framework, enabling efficient annotation from sparse multi-view observations.

2.3 Human-to-Robot Hand Motion Retargeting

Human-to-robot hand motion retargeting aims to transfer human hand motions to dexterous robot hands despite differences in hand morphology, kinematic structure, and degrees of freedom. Existing approaches (Handa et al., 2020; Qin et al., 2023; Yin et al., 2025; Huang et al., 2025; Lin et al., 2025) primarily map human hand poses to robot joint configurations through kinematic correspondence, inverse kinematics, or optimization, while accounting for differences in joint limits and hand geometry. Some recent approaches (Wu et al., 2026; Pan et al., 2025; Yang et al., 2025) further consider hand-object interactions by incorporating object motion and contact relationships into the retargeting process. These interaction-preserving retargeting approaches seek to preserve the interaction between the robot hand and manipulated objects while adapting human motions to different robotic embodiments.

3 Method and Dataset

Given the collected human manipulation demonstrations, we develop an automated pipeline to recover 3D hand-object motions and transfer them across embodiments to a target dexterous robot. Our key idea is to use hand-object interaction as a transferable representation that connects motion recovery with cross-embodiment trajectory construction. Specifically, we first recover 3D hand-object motions and construct fine-grained interaction representations through hand pose estimation and interaction-aware object motion estimation (Section 3.1). We then incorporate the constructed interaction representation into a physics-based framework to validate the recovered interactions and construct physically feasible cross-embodiment trajectories (Section 3.2). Finally, based on the above pipeline, we construct H2D, a dataset for cross-embodiment dexterous manipulation, from diverse human demonstrations with hand-object motion annotations (Section 3.3).

3.1 Hand-Object Motion Annotation

Given the captured multi-view manipulation videos, we aim to recover temporally consistent 3D hand-object motions from sparse visual observations. The limited visual evidence makes object motion particularly ambiguous under hand-object occlusion, small object size, and weak textures. To resolve this ambiguity, we first recover the hand motion following USTHand (Han et al., 2026), a multi-view self-supervised hand pose estimation method, and use it to derive hand-semantic contact

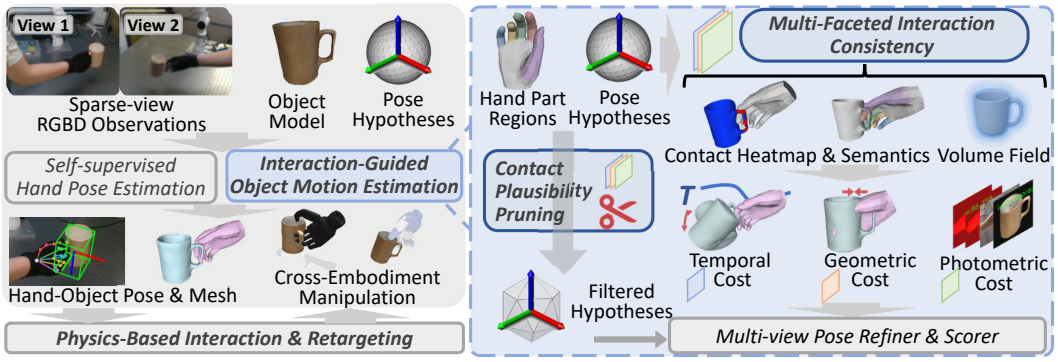


Figure 2: Overview of our interaction-centric pipeline for hand-object motion annotation and physics-based cross-embodiment manipulation trajectory construction. Given captured sparse multi-view manipulation videos, we first recover hand motion using self-supervised hand pose estimation, and then exploit multi-faceted hand-object interaction consistency to guide object motion estimation under limited visual evidence. Physics-based cross-embodiment transfer then preserves the recovered hand-object interactions.

cues for different object pose hypotheses. These interaction cues are then used to narrow the object pose hypothesis space and guide subsequent pose refinement, as shown in Fig. 2. In the following, we first introduce the construction of the hand-object interaction representation, followed by their incorporation into interaction-guided object pose estimation.

3.1.1 Interaction Representation Construction

Following existing hand-object interaction representations such as contact maps (Fan et al., 2023; Grady et al., 2021), which represent hand-object contact as a spatial heatmap over the object surface, we construct a multi-faceted interaction representation from the recovered hand and object geometry. Rather than representing interaction with a single global contact relation, our representation captures complementary cues from three perspectives: semantic correspondence, spatial contact distribution, and volumetric geometric consistency. First, we partition each hand into semantically meaningful regions, such as individual fingers and the palm, and associate each object surface region with its corresponding hand region, forming a semantic contact representation that preserves hand-object correspondence. Second, we construct a contact heatmap over the object surface to characterize the spatial distribution and strength of hand-object contact. Third, we construct an object volume field represented by a signed distance function (SDF), which characterizes the volumetric relationship between the hands and object.

3.1.2 Interaction-Guided Object Motion Estimation

We follow the candidate-based pose estimation and render-and-compare paradigm of FoundationPose (Wen et al., 2024) to generate and evaluate object pose hypotheses. Under sparse-view observations, severe hand-object occlusion, small object size, and weak textures make visual consistency insufficient to uniquely determine object motion. Multiple pose hypotheses may therefore obtain similar visual scores while implying different hand-object interactions.

We leverage the multi-faceted interaction representation constructed above to constrain the pose hypotheses further. Specifically, we compute contact consistency and SDF-based geometric costs for each candidate, use these costs to prune implausible hypotheses, and refine the remaining candidates with the pose refinement module of FoundationPose (Wen et al., 2024). The refined candidates are then jointly evaluated using multi-view visual and interaction consistency. For each frame t, we maintain a set of candidate object poses

$$\mathcal{H}_t = \{T_t^k\}_{k=1}^{N_t}, \tag{1}$$

where T_t^k denotes the k-th candidate object pose. For each candidate, we render the hand-object configuration from all N_v captured views. Let $\mathcal{R}(H_t, T_t^k, v)$ denote the rendered observation from

view v, and I_t^v the corresponding captured observation, where H_t denotes the recovered poses of both hands at frame t. The recovered hand masks are considered during render-and-compare to account for hand-induced occlusion. The multi-view visual score is computed using the pretrained visual scoring network of FoundationPose (Wen et al., 2024):

$$S_{\text{vis}}(T_t^k) = \frac{1}{N_v} \sum_{v=1}^{N_v} \mathcal{D}\left(\mathcal{R}(H_t, T_t^k, v), I_t^v\right), \tag{2}$$

where $\mathcal{D}(\cdot,\cdot)$ denotes the pretrained visual scoring network. Candidates with similar object poses are further removed before interaction-aware evaluation.

We next compute interaction consistency costs from the multi-faceted interaction representation. The semantic correspondence is used to partition the object contact region by hand side and hand part, while the contact heatmap characterizes the spatial distribution within each region. Let $\mathcal{S} = \{\text{L}, \text{R}\}$ denote the two hands, and let $\mathcal{F}_{t,s}$ denote the set of hand parts associated with hand s at frame t. For each hand part f, let $\mathcal{C}_{t-1,s,f}$ and $\mathcal{C}_{t,s,f}^k$ denote the corresponding contact regions at frames $t-1$ and t under candidate pose T_t^k, respectively. We measure their discrepancy using a bidirectional Chamfer distance:

$$E_{\text{con}}(T_t^k) = \frac{1}{2|\mathcal{S}|} \sum_{s\in\mathcal{S}} \frac{1}{|\mathcal{F}_{t,s}|} \sum_{f\in\mathcal{F}_{t,s}} \left[d(\mathcal{C}_{t-1,s,f}, \mathcal{C}_{t,s,f}^k) + d(\mathcal{C}_{t,s,f}^k, \mathcal{C}_{t-1,s,f})\right]. \tag{3}$$

where

$$d(A,B) = \frac{1}{|A|} \sum_{\mathbf{a}\in A} \min_{\mathbf{b}\in B} \|\mathbf{a}-\mathbf{b}\|_2 . \tag{4}$$

The resulting cost measures the temporal consistency of hand-object contact while preserving correspondence across both hands and their individual hand parts. We further impose geometric constraints using the object signed distance field. Let $\mathcal{V}_{H_s}$ denote the recovered vertices of hand s, and SDF_O the object signed distance field defined in its canonical coordinate space. For candidate T_t^k, each hand vertex $\mathbf{p}$ is transformed to the canonical object space as $T_t^k\mathbf{p}$, where its signed distance to the object surface is evaluated:

$$E_{\text{sdf}}(T_t^k) = \frac{1}{|\mathcal{S}|} \sum_{s\in\mathcal{S}} \frac{1}{|\mathcal{V}_{H_s}|} \sum_{\mathbf{p}\in\mathcal{V}_{H_s}} \phi\left(\text{SDF}_O\left(T_t^k\mathbf{p}\right)\right), \tag{5}$$

where

$$\phi(d) = \max(-d, 0, d-\delta). \tag{6}$$

Here, δ defines the desired contact proximity range. The overall interaction cost is

$$E_{\text{int}}(T_t^k) = \lambda_{\text{con}} E_{\text{con}}(T_t^k) + \lambda_{\text{sdf}} E_{\text{sdf}}(T_t^k). \tag{7}$$

We first use the interaction costs to prune implausible candidates. Candidates with large contact inconsistency or severe SDF violations are discarded according to E_{con} and E_{sdf}. The remaining candidates are refined using the pose refinement module of FoundationPose (Wen et al., 2024), yielding refined poses $\widetilde{T}_t^k$. For each refined pose $\widetilde{T}_t^k$, we recompute the multi-view visual score and interaction cost. The final score is

$$S(\widetilde{T}_t^k) = S_{\text{vis}}(\widetilde{T}_t^k) - \lambda_{\text{int}} E_{\text{int}}(\widetilde{T}_t^k), \tag{8}$$

where λ_{int} balances visual and interaction consistency. The refined candidates are ranked according to $S(\widetilde{T}_t^k)$, and the top-ranked hypotheses are retained:

$$\mathcal{H}_t^K = \text{TopK}\left(\left\{\widetilde{T}_t^k\right\}, S(\widetilde{T}_t^k)\right). \tag{9}$$

Overall, our approach incorporates bimanual hand-object interaction in a feed-forward manner into candidate-based object pose estimation, progressively pruning interaction-inconsistent hypotheses during candidate evaluation. Compared with conventional sequence-level optimization of object pose and contact, this formulation avoids costly iterative optimization over entire sequences, substantially reducing the computational cost of object pose annotation.

Table 1: Comparison of hand-object datasets. H/O/R: hand/object/robot trajectories.

Dataset	Seq.	Hours	Frames	Subj.	Obj.	Views	Content	Type
FPHA(Garcia-Hernando et al., 2018)	1,175	–	105K	6	4	Ego	H+O	Mocap
HO-3D(Hampali et al., 2020)	27	–	78K	10	10	1–5	H+O	Auto.
DexYCB(Chao et al., 2021)	1,000	–	582K	10	20	8	H+O	Manual
OakInk(Yang et al., 2022)	792	–	230K	12	100	4	H+O	Manual
HOI4D(Liu et al., 2022)	4,000	22.2	2.4M	4	800	Ego	H+O	Manual
ARCTIC(Fan et al., 2023)	399	2.0	2.1M	10	11	8+Ego	H+O	Mocap
OakInk2(Zhan et al., 2024)	627	9.3	4.01M	9	75	3+Ego	H+O	Mocap
HOT3D(Banerjee et al., 2025)	4.1K	13.9	3.7M	19	33	2–3	H+O	Auto.
GigaHands(Fu et al., 2025)	13.9K	33.9	3.7M	56	417	51	H+O	Auto.
SHOW3D(Rim et al., 2026)	–	–	4.3M	38	21	8+2 Ego	H+O	Auto.
H2D (Ours)	**2,200**	**5.6**	**1.1M**	**10**	**16**	**3–5**	**H+O+R**	**Auto.**

3.2 Physics-Based Interaction Validation and Cross-Embodiment Retargeting

The recovered human manipulation trajectories may contain physically implausible hand-object interactions due to estimation errors and limited visual evidence. To construct physically grounded cross-embodiment trajectories, we build upon the physics-based sampling framework of SPIDER (Pan et al., 2025). Given a reference manipulation trajectory, SPIDER samples a set of candidate dexterous trajectories and rolls them out in a physics simulator, where the candidates are assessed according to their consistency with the reference motion and physical constraints. The sampling process progressively favors physically consistent trajectories, allowing the target dexterous hand to adapt the demonstrated motion to its own embodiment while respecting object contact and collision constraints.

We refine its contact representation with our constructed interaction representation in Sec. 3.1.1. Specifically, we use the fine-grained contact representation established above, which characterizes the correspondence between semantic hand regions and object surface regions. This provides SPIDER with a more structured description of the demonstrated hand-object contacts while retaining its original physics-based sampling framework. The resulting trajectories are then replayed in simulation under gravity, contact, and collision constraints, yielding physically feasible cross-embodiment trajectories that preserve the demonstrated hand-object interactions.

3.3 H2D Dataset

Building upon the above hand-object annotation pipeline, we construct a Human-to-Dexterous (**H2D**) dataset of human dexterous manipulation demonstrations comprising 5.6 hours of recordings and 2,200 trajectories from 10 subjects, 16 objects, and 22 manipulation tasks, as summarized in Table 1. The tasks cover object pick-and-place, handover, and functional use, involving both unilateral and bimanual manipulation. The objects span different shapes, sizes, and interaction affordances, while the trajectories capture continuous hand and object motions with dynamic hand-object contacts. We deploy low-cost multi-view RGB-D capture systems, each consisting of 3–5 Intel RealSense D415 cameras sparsely arranged around the manipulation workspace. The cameras are hardware-synchronized to capture multi-view RGB-D streams at 15 fps, balancing temporal resolution and data scale to enable broader coverage of manipulation trajectories. For each object, we acquire a corresponding 3D model through mobile-device-based 3D scanning. Hand and object masks are semi-automatically annotated using SAM2 (Ravi et al., 2024) across the captured views. We then apply the fully automatic motion annotation pipeline to obtain 3D hand-object poses and interaction states for the captured sequences. Based on these annotations, robot manipulation trajectories can be further generated through the cross-embodiment retargeting procedure described in Section 3.2.

Table 2: Object annotation accuracy and efficiency on the DexYCB test set.

Method	Accuracy (mm)				Efficiency	
	OCE↓	MCE ↓	ADD-S ↓	ADD↓	Latency	GPU Time
HO-CAP (Wang et al., 2025)	17.92	21.60	12.77	35.21	18.26 s	59.12 h
AlignPose (Mikeštíková et al., 2026)	20.30	21.26	14.28	37.80	7.83 s	25.39 h
Ours	**7.86**	**7.63**	**3.79**	**22.12**	**0.99 s**	**3.20 h**

Table 3: Hand pose estimation errors (mm) on the DexYCB test set.

Method	MPJPE ↓	PA-MPJPE ↓	MPVPE ↓	PA-MPVPE ↓
HO-CAP (Wang et al., 2025)	18.78	7.64	18.30	7.34
Ours	**12.92**	**5.83**	**12.47**	**5.67**

4 Experiments

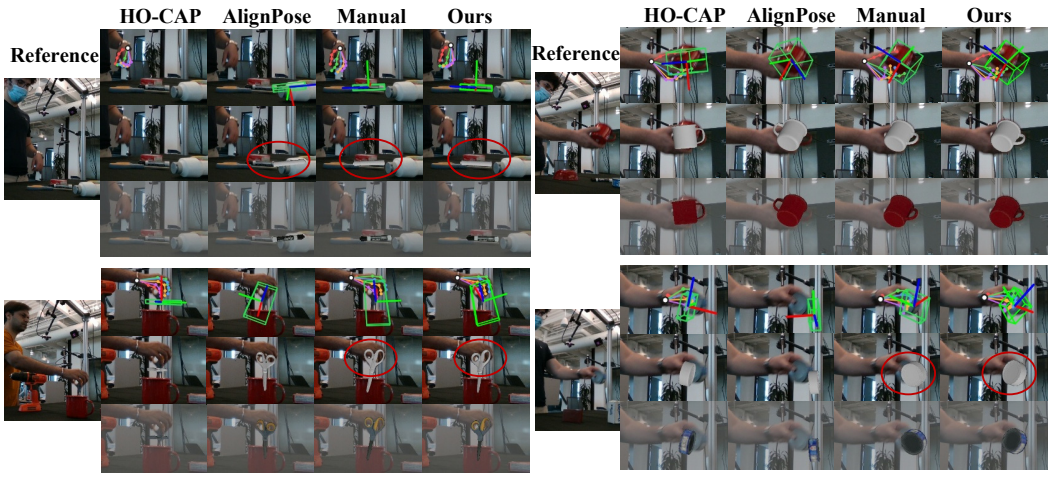
4.1 Experimental Setup

Cross-Dataset Evaluation Protocol. We evaluate on the DexYCB S0 test split, comprising 160 sequences and 11,658 frames, using four calibrated RGB-D views per sequence to assess sparse-view hand-object motion estimation. Experiments are conducted on 4 NVIDIA RTX 5090 GPUs. We measure hand accuracy using Mean Per Joint Position Error (MPJPE) and Mean Per Vertex Position Error (MPVPE) after wrist translation alignment, together with their Procrustes-aligned counterparts, PA-MPJPE and PA-MPVPE. Object accuracy is measured using Object Center Error (OCE), Mean Corner Error (MCE), Average Distance of Model Points (ADD), and Average Closest Point Distance (ADD-S). ADD measures the mean distance between corresponding mesh vertices, whereas ADD-S uses nearest-neighbor correspondences. All distances are reported in millimeters. The supplementary material includes the original evaluation results. The evaluation code will also be released with sufficient implementation details to facilitate reproducibility.

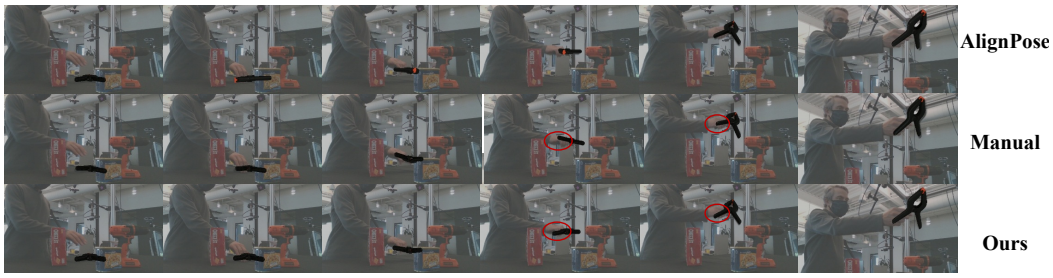
Baselines. We compare against HO-CAP (Wang et al., 2025) for hand-object motion annotation, and AlignPose (Mikeštíková et al., 2026) for multiview object pose estimation. For HO-CAP, we adapt the released pipeline to the DexYCB camera configuration and coordinate system. The dataset-provided segmentation masks are used in place of its segmentation stage, while the HoloLens-specific stage is omitted. We adopt frame-wise FoundationPose registrations for object pose initialization, as temporal tracking is prone to severe pose drift. For AlignPose, these registrations are further refined using its multi-view feature-metric optimization.

4.2 Hand-Object Annotation Quality

Pose Estimation Accuracy. Our annotation system leverages multi-view observations and interaction constraints to recover accurate object poses under severe hand-object occlusion. As shown in Table 2, our method reduces the OCE and MCE from the centimeter level with HO-CAP and AlignPose to the sub-centimeter level, achieving 7.86 mm and 7.63 mm, respectively. Notably, this accuracy is achieved with only sparse multi-view observations, where the available visual evidence is limited. The ADD-S of 3.79 mm indicates highly accurate alignment of the entire object. For the more challenging ADD metric, our method achieves substantially lower error than the baselines. On DexYCB, where many objects exhibit geometric symmetries and weak textures, ADD penalizes orientation errors that ADD-S ignores, making it a stronger indicator of hand-object interaction consistency and spurious in-hand rotations. Qualitative comparisons in Fig. 3 show greater robustness to challenging hand-object interactions than HO-CAP and AlignPose, with predictions exhibiting even tighter pixel-level alignment with the input images than the manually annotated ground truth. The temporal results further show consistent object motion over manipulation sequences. Table 3 additionally shows that our hand pose estimation pipeline achieves lower errors for both hand joints and vertices than HO-CAP.



(a) Comparisons of hand-object pose estimation under challenging interaction scenarios.



(b) Temporal comparisons of object motion estimation.

Figure 3: Qualitative comparisons of hand-object pose estimation on DexYCB.

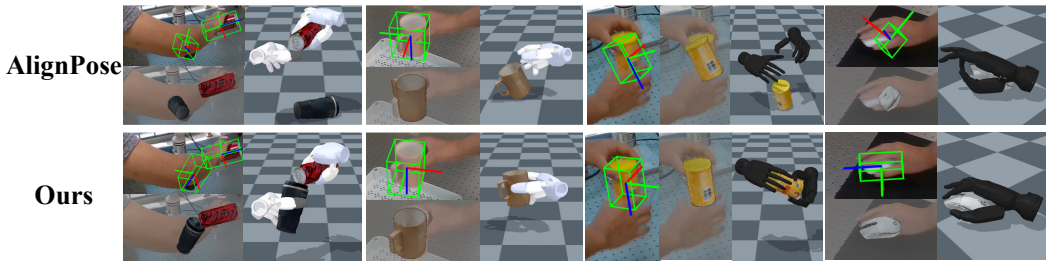
Table 4: Cross-embodiment simulation results on the H2D subset. ATE-Position (mm), ATE-Rotation (°), and Success (%). Rotation errors are converted to degrees for intuitive interpretation.

Method	Human Hand			XHand			Inspire		
	ATE-P ↓	ATE-R ↓	Succ. ↑	ATE-P ↓	ATE-R ↓	Succ. ↑	ATE-P ↓	ATE-R ↓	Succ. ↑
AlignPose	64.74	21.18	48.2	49.10	18.34	52.7	45.19	17.29	66.4
Ours	**61.93**	**18.80**	**52.7**	**46.85**	**12.12**	**60.9**	**39.39**	**11.70**	**73.6**

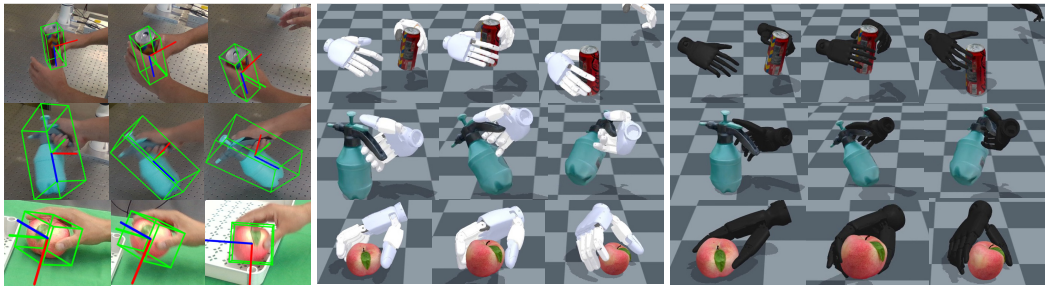
Annotation Efficiency. Object pose estimation and interaction optimization constitute the major computational costs in hand-object motion annotation. Our annotation system substantially reduces the object pose estimation cost through interaction-aware hypothesis pruning and a feed-forward, end-to-end design. As shown in Table 2, our method requires only 0.99 seconds per frame for object motion annotation. In contrast, the baselines rely on costly iterative pose optimization and refinement, resulting in substantially higher per-frame computation. For the 160 sequences in the DexYCB test split, this reduces the object pose annotation cost from days with the baselines to just a few GPU-hours, enabling efficient large-scale annotation of human manipulation demonstrations.

4.3 Cross-Embodiment Trajectory Validation

We further validate our interaction-guided object motion estimation through physics-based simulation with human and dexterous hands. To isolate its contribution, we compare our method with AlignPose (Mikeštíková et al., 2026) using the same hand pose annotations estimated by our method. Experiments are conducted on an H2D subset of 110 sequences from 22 tasks, with 5 demonstrations per task. We retarget the recovered hand-object trajectories to MANO-based human hands, XHand dexterous hands, and Inspire hands, and simulate them in MuJoCo (Todorov et al., 2012).



(a) Qualitative comparison of object pose estimation and physics-based execution.



(b) Temporal visualization of object pose estimation and physics-based execution.

Figure 4: Qualitative comparisons of object motion estimation and physics-based execution.

We evaluate simulated object trajectories using Absolute Trajectory Error (ATE) in position and rotation against the corresponding annotated trajectories. We also report the success rate, defined as the proportion of actions with position and rotation errors below the default thresholds used by SPIDER, 0.1 m and 0.3 rad, respectively. As shown in Table 4, our method achieves lower trajectory errors and higher success rates across all embodiments.

Qualitative results in Fig. 4 further illustrate the object motion estimation and cross-embodiment execution results. In Fig. 4(a), our method produces more accurate object poses under challenging conditions, including severe hand-object occlusion and weak object textures, supporting feasible manipulation across multiple embodiments. Fig. 4(b) shows the recovered trajectories over consecutive frames, where object poses remain temporally smooth and stable while hand-object contacts are consistently preserved without noticeable in-hand slipping.

5 Conclusion

We present an interaction-aware framework that leverages embodiment-transferable hand-object interaction priors to recover structured human manipulation states and transfer them to dexterous robots. By exploiting spatio-temporal contact plausibility, our method enables reliable object motion estimation from sparse-view observations and interaction-preserving motion retargeting. Extensive experiments demonstrate several-fold higher object annotation efficiency than existing methods and sub-centimeter object motion estimation under sparse-view capture, while physics-based simulation validates interaction-preserving cross-embodiment transfer. Building on this framework, we introduce **H2D**, a dataset comprising 5.6 hours of human manipulation demonstrations and 2,200 dexterous manipulation trajectories with structured 3D hand-object states, providing a scalable foundation for human-to-dexterous manipulation data generation.

Limitations. H2D provides structured dexterous manipulation trajectories with rich hand-object interaction information, while their integration into robot policy learning remains an open direction. Future work will explore how these trajectories and interaction priors can be incorporated into policy learning, and investigate their effects on learned policy performance and real-world transfer.

AI USE STATEMENT

In this work, we used generative AI tools for manuscript editing, including language polishing, grammar correction, and refinement of technical descriptions. We also used generative AI tools to assist with code editing and debugging during implementation. We have not used generative AI tools to generate experimental data, fabricate results, or perform mathematical or theoretical derivations. All AI-assisted code was reviewed and tested by the authors, and all AI-assisted text was manually reviewed for accuracy and consistency with the actual methods and experiments. We take responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves human manipulation demonstrations collected for research purposes. The collected data are used to study hand-object interaction and human-to-dexterous manipulation data generation. We take care to avoid including unnecessary personally identifiable information in the released data and associated materials. The dataset and its release will follow applicable requirements for privacy, consent, and responsible data sharing.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the data collection, hand-object motion annotation, interaction representation, and cross-embodiment motion retargeting pipeline in the main paper and supplementary material. We report the experimental settings, evaluation metrics, comparison methods, and implementation details needed to reproduce the reported results. The supplementary material also includes the raw experimental data used to obtain the reported results. Source code and more implementation details will be released to further facilitate reproduction of the experiments.

REFERENCES

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7061–7071. IEEE, 2025.

Adnane Boukhayma, Rodrigo de Bem, and Philip H.S. Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2019.

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*, pp. 9044–9053, 2021.

Yujin Chen, Zhigang Tu, Di Kang, Linchao Bao, Ying Zhang, Xuefei Zhe, Ruizhi Chen, and Junsong Yuan. Model-based 3d hand reconstruction via self-supervised learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10446–10455. IEEE, 2021.

Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In *European Conference on Computer Vision*, pp. 231–248. Springer, 2022.

Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. gsdf: Geometry-driven signed distance functions for 3d hand-object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12890–12900, 2023.

Zerui Chen, Rolandos Alexandros Potamias, Shizhe Chen, and Cordelia Schmid. Hort: Monocular hand-held objects reconstruction with transformers. *arXiv preprint arXiv:2503.21313*, 2025.

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*, 2023.

Rao Fu, Dingxi Zhang, Alex Jiang, Wanjia Fu, Austin Funk, Daniel Ritchie, and Srinath Sridhar. Gigahands: A massive annotated dataset of bimanual hand activities. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17461–17474. IEEE, 2025.

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 *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 409–419. IEEE, 2018.

Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1471–1481, 2021.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19383–19400, 2024.

Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020.

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*, pp. 11090–11100, 2022.

Tianhao Han, Haoyang Zhang, Liang Xie, Haochen Chang, Kun Gao, Yuan Cheng, Pengfei Ren, and Erwei Yin. Ust-hand: An uncertainty-aware spatiotemporal point cloud interaction network for 3d self-supervised hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8857–8867, June 2026.

Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan Ratliff, and Dieter Fox. Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system. In *2020 IEEE international conference on robotics and automation (ICRA)*, pp. 9164–9170. IEEE, 2020.

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*, pp. 11807–11816, 2019.

Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 571–580, 2020.

Ryan Hoque, Peide Huang, David Yoon, Jian Zhang, et al. Egodex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations*, volume 2026, pp. 4218–4237, 2026.

Linyi Huang, Hui Zhang, Zijian Wu, Sammy Christen, and Jie Song. Fungrasp: Functional grasping for diverse dexterous hands. *IEEE Robotics and Automation Letters*, 10(6):6175–6182, 2025.

Weiting Huang, Pengfei Ren, Jingyu Wang, Qi Qi, and Haifeng Sun. Awr: Adaptive weighting regression for 3d hand pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11061–11068, 2020.

Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5d heatmap regression. In *Proceedings of the European Conference on Computer Vision*, September 2018.

Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. *arXiv preprint arXiv:2212.06870*, 2022.

Ming-Feng Li, Xin Yang, Fu-En Wang, Hritam Basak, Yuyin Sun, Shreekant Gayaka, Min Sun, and Cheng-Hao Kuo. Ua-pose: Uncertainty-aware 6d object pose estimation and online object completion with partial references. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1180–1189. IEEE, 2025.

Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27906–27916. IEEE, 2024.

Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 1954–1963, 2021a.

Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12939–12948, 2021b.

Xiaoyi Lin, Kunpeng Yao, Lixin Xu, Xueqiang Wang, Xuetao Li, Yuchen Wang, and Miao Li. Dexflow: A unified approach for dexterous hand pose retargeting and interaction. *arXiv preprint arXiv:2505.01083*, 2025.

Xingyu Liu, Pengfei Ren, Jingyu Wang, Qi Qi, Haifeng Sun, Zirui Zhuang, and Jianxin Liao. Coarse-to-fine implicit representation learning for 3d hand-object reconstruction from a single rgb-d image. In *European Conference on Computer Vision*, pp. 74–92. Springer, 2025.

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21740–21751. IEEE, 2024.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20981–20990. IEEE, 2022.

Xintao Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, et al. Himo: A new benchmark for full-body human interacting with multiple objects. In *European Conference on Computer Vision*, pp. 300–318. Springer, 2024.

Anna Šárová Mikeštíková, Médéric Fourmy, Martin Cífka, Josef Sivic, and Vladimir Petrik. Alignpose: Generalizable 6d pose estimation via multi-view feature-metric alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14626–14636, 2026.

Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *European Conference on Computer Vision*, pp. 752–768. Springer, 2020.

Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent Lepetit. Gigapose: Fast and robust novel object pose estimation via one correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9903–9913, 2024.

Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12999–13008. IEEE, 2023.

Chaoyi Pan, Changhao Wang, Haozhi Qi, Zixi Liu, Homanga Bharadhwaj, Akash Sharma, Tingfan Wu, Guanya Shi, Jitendra Malik, and Francois Hogan. Spider: Scalable physics-informed dexterous retargeting. *arXiv preprint arXiv:2511.09484*, 2025.

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*, pp. 1496–1505, 2022.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild, 2024.

Haozhe Qi, Chen Zhao, Mathieu Salzmann, and Alexander Mathis. Hoisdf: Constraining 3d hand-object pose estimation with global signed distance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10392–10402, 2024.

Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system. *arXiv preprint arXiv:2307.04577*, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

Huan Ren, Yihan Chen, Chuxin Wang, Nailong Liu, Wenfei Yang, and Tianzhu Zhang. Compose: A unified completion-pose framework for robust category-level object pose estimation. *arXiv preprint arXiv:2605.25553*, 2026.

Patrick Rim, Kevin Harris, Braden Copple, Shangchen Han, Xu Xie, Ivan Shugurov, Sizhe An, He Wen, Alex Wong, Tomas Hodan, et al. Show3d: Capturing scenes of 3d hands and objects in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7111–7120, 2026.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017.

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*, pp. 581–600. Springer, 2020.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

Jikai Wang, Qifan Zhang, Yu-Wei Chao, Bowen Wen, Xiaohu Guo, and Yu Xiang. HO-cap: A capture system and dataset for 3d reconstruction and pose tracking of hand-object interaction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 606–617. IEEE, 2023.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17868–17879. IEEE, 2024.

Jielin Wu, Shenzhe Yao, Guanqi He, Xiaohan Liu, Zhaoqing Zeng, Xiangrui Jiang, Han Yang, Wentao Zhang, and Hang Zhao. Toporetarget: Interaction-preserving retargeting for dexterous manipulation. *arXiv preprint arXiv:2606.16272*, 2026.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20921–20930. IEEE, 2022.

Lujie Yang, Xiaoyu Huang, Zhen Wu, Angjoo Kanazawa, Pieter Abbeel, Carmelo Sferrazza, C Karen Liu, Rocky Duan, and Guanya Shi. Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. *arXiv preprint arXiv:2509.26633*, 2025.

Zhao-Heng Yin, Changhao Wang, Luis Pineda, Krishna Bodduluri, Tingfan Wu, Pieter Abbeel, and Mustafa Mukadam. Geometric retargeting: A principled, ultrafast neural hand retargeting algorithm. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 17376–17382. IEEE, 2025.

Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 445–456. IEEE, 2024.

Xiaozheng Zheng, Chao Wen, Zhou Xue, Pengfei Ren, and Jingyu Wang. Hamuco: Hand pose estimation via multiview collaborative self-supervised learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20706–20716. IEEE, 2023.