This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

RoboWheel: A Data Engine from Real-World Human Demonstrations for Cross-Embodiment Robotic Learning

Yuhong Zhang[1*] Zihan Gao[1*] Shengpeng Li[2] Ling-Hao Chen[1] Kaisheng Liu[2] Runqing Cheng[2]
Xiao Lin[2†] Junjia Liu[2,3] Zhuoheng Li[4] Jingyi Feng[5] Ziyan He[2] Jintian Lin[2]
Zheyan Huang[1] Zhifang Liu[1] Haoqian Wang[1†]
[1]Tsinghua University [2]Synapath [3]CUHK [4]HKU [5]PolyU

Project page: https://zhangyuhong01.github.io/Robowheel

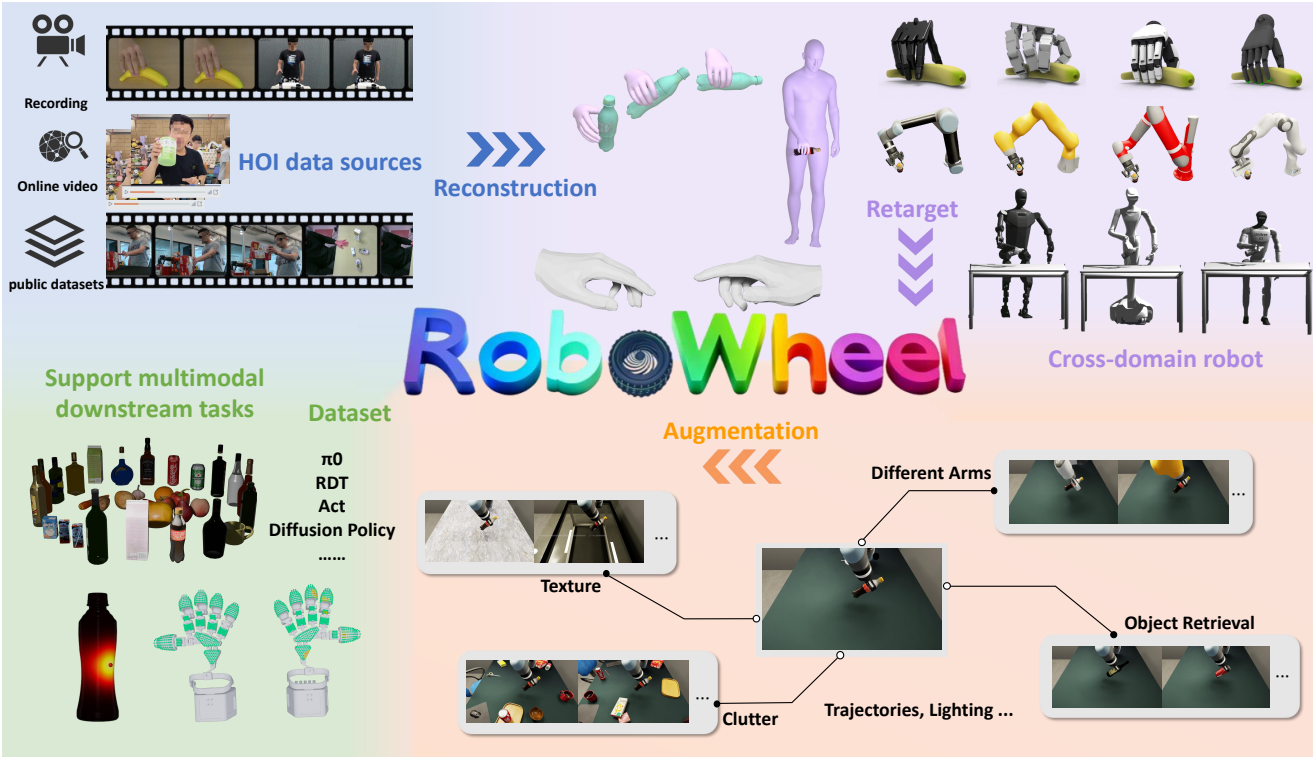


Figure 1. The *RoboWheel* data engine. Our pipeline could process hand-object interaction (HOI) videos from diverse sources (*e.g.*, recordings and public datasets) through high-fidelity reconstruction to recover physically consistent trajectories. The reconstructed motions are retargeted to cross-domain robotic embodiments (*e.g.*, arms, dexterous hands, humanoids), and enhanced with multi-modal data augmentations (*e.g.*, object retrieval, texture, trajectories). This generates a large-scale dataset with multimodal observations, supporting training for various vision-language-action (VLA) and imitation learning models (*e.g.*, ACT, Diffusion Policy).

Abstract

*We introduce **RoboWheel**, a data engine that converts human hand–object interaction (HOI) videos into training-ready supervision for cross-morphology robotic learning. From monocular RGB/RGB-D inputs, we perform high-precision HOI reconstruction and enforce physical plausibility via a reinforcement learning (RL) optimizer that refines hand–object relative poses under contact and penetration constraints. The reconstructed, contact-rich trajectories are then retargeted to cross-embodiments, robot arms with simple end-effectors, dexterous hands, and humanoids, yielding executable actions and rollouts. To scale coverage, we build a simulation-augmented framework on Isaac Sim with diverse domain randomization (embodiments, trajectories, object retrieval, background textures, hand motion mirroring), which enriches the distributions of trajectories and observations while preserving spatial relationships and*

*Equal contribution, †Corresponding author.

physical plausibility. The entire data pipeline forms an end-to-end pipeline from video → reconstruction → retargeting → augmentation → data acquisition. We validate the data on mainstream vision–language–action (VLA) and imitation learning architectures, demonstrating that trajectories produced by our pipeline are as stable as those from teleoperation and yield comparable continual performance gains. To our knowledge, this provides the first quantitative evidence that HOI modalities can serve as effective supervision for robotic learning. Compared with teleoperation, RoboWheel is lightweight: a single monocular RGB(D) camera is sufficient to extract a universal, embodiment-agnostic motion representation that could be flexibly retargeted across embodiments. We further assemble a large-scale multimodal dataset combining multi-camera captures, monocular videos, and public HOI corpora for training and evaluating embodied models.

1. Introduction

Embodied agents learn most effectively when their supervision matches how humans interact with the world. However, obtaining specific robot-usable supervision at scale is challenging. Existing data collection pipelines mainly rely on teleoperation or studio motion capture, which demands specialized hardware and careful curation, limiting the diversity of action behaviors and their transferability across robot embodiments and tasks. At the same time, multiple data sources provide vast amounts of hand–object interaction (HOI) information containing rich and real-world manipulation strategies, yet these signals are rarely converted into training-ready data for robots due to reconstruction noise, physical implausibility, and embodiment mismatch.

Recent advances in human- and object-centric perception suggest an opportunity to close this gap. Methods based on SMPL-H/MANO [45, 47, 49] and 6D object pose or mesh tracking [39, 40, 46] can recover detailed geometry and motion from monocular RGB/RGB-D inputs. Nevertheless, contact estimates are often inconsistent, hand–object interpenetrations arise under occlusion, trajectories lack temporal smoothness, and the recovered motions generally violate kinematic and dynamic constraints.

Bridging this gap requires a practical, scalable processing pipeline that satisfies three key requirements: (i) enabling large-scale, continuous acquisition of robot–object interaction trajectories in real-world operational spaces while enforcing physical plausibility; (ii) supporting flexible retargeting of these trajectories to diverse robot embodiments—even across domains—while preserving interaction semantics; and (iii) maintaining scalability by enabling effective composition of data augmentation strategies. Existing paradigms for robot data collection—teleoperation and simulated data generation—fail to meet these requirements jointly: teleoperation is costly and hardware-specific, whereas purely synthetic simulation data often does not reflect real-world perceptual and contact distributions.

We address these challenges with *RoboWheel*, a data engine that transforms real-world hand–object demonstrations into supervision for cross-embodiment robotic learning. Starting from monocular RGB(D) HOI videos, *RoboWheel* integrates state-of-the-art hand, body, and object motion estimators into a unified reconstruction framework. A multi-stage physical plausibility optimization then refines the trajectories: signed distance function (SDF)-based penalties discourage interpenetration and encourage contact, and a residual reinforcement learning (RL) policy further adjusts hand–object relative poses under reachability and stability priors. The resulting trajectories are mapped into a canonical action space and retargeted to heterogeneous robot morphologies—including 6/7-DoF arms with parallel grippers, dexterous hands, and humanoids, and output executable control sequences in both operational and joint spaces.

To scale coverage while preserving interaction semantics, we replay the retargeted trajectories in simulation across multiple robotic arms using GPU-accelerated parallel inverse kinematics. Within this environment, we apply domain randomization and diverse augmentations for observation and robotic action, including embodiment variations, trajectory distribution enrichment, object retrieval and replacement, background texture randomization, cluttered table configurations, hand mirroring, and more.

On top of this engine, we construct *HORA* (**H**and–**O**bject to **R**obot **A**ction dataset), a large-scale multimodal dataset from three sources: a custom multi-view motion capture system with tactile-sensor gloves, an RGB(D) HOI recording setup, and multiple public HOI datasets. As a multimodal dataset, *HORA* combines the data modalities of HOI corpora and embodied robot datasets and *HORA* supports both HOI-related analysis and downstream robotic learning.

In summary, our key contributions as follows.

- **Physically plausible HOI reconstruction and cross-domain retargeting.** A monocular RGB/RGB-D HOI reconstruction framework that integrates state-of-the-art hand, body, and object motion estimation with physics-based optimization. The framework supports flexible cross-embodiment retargeting, outputs executable motion trajectories, and provides scalable supervision signals for downstream learning.
- **Simulation-augmented data flywheel.** We implement rich augmentation and domain randomization in simulation driven by HOI data. This data flywheel is validated on mainstream VLA and imitation learning models, and we conduct a quantitative evaluation of the quality and effectiveness of both the HOI transformation and the augmentation strategy.
- **Large-scale multimodal dataset.** A multimodal dataset of over 150,000 sequences is constructed from multi-

view motion capture, public HOI datasets, and recorded videos. It includes embodiment trajectories, observations, assets, HOI annotations, and task descriptions, with a mocap-based subset also providing tactile signals, offering a rich and scalable resource for robotic learning and downstream HOI-related tasks.

2. Related Work

HOI datasets and monocular reconstruction. High-precision HOI annotations remain costly, as most public 3D HOI datasets rely on multiview rigs or motion capture (MoCap) systems for accurate hand-object geometry [5, 14, 15, 33, 37]. Large egocentric video corpora like [13] use head-mounted cameras to avoid MoCap but lack frame-accurate 3D HOI geometry for reconstruction [12]. Recent whole-body motion datasets such as [48] scale to millions of SMPL-X frames but are not dedicated HOI datasets and offer limited hand-object contact supervision. On the algorithmic side, [8] reconstructs objects by fusing pixel-aligned features with 3D hand geometry in a transformer-based coarse-to-fine point cloud decoder, yielding dense object geometry with high frame fidelity, while [11] jointly reconstructs articulated hands and objects using compositional SDF and contact constraints. These methods, however, are generally limited to single-frame or in-contact scenarios and struggle with approach/withdrawal phases, generalizability, occlusion, low video resolution, and varying hand movement speeds. Recently, more generalizable approaches [30, 31, 43] have used data-driven priors; for instance, [43] introduces diffusion-guided, per-video optimization to enhance robustness under occlusion, albeit at the cost of heavier computation and the need for short clips.

Robotic Learning from Human Demonstration Early work in vision-based programming by demonstration mapped human hand poses to robot grasps directly from images, establishing a pipeline from grasp recognition to example-based robot execution [20]. More recent systems leverage richer HOI signals: [52] extract binocular hand-motion cues from human videos, compress trajectories into keyframes with coordination masks, augment demonstrations geometrically, and train a bimanual diffusion policy that executes long-horizon dual-arm tasks and generalizes across scenes. Complementarily, [38] treats human gestures as structured priors, retrieving grasp affordances from HOI memories and transferring them to novel objects, yielding robust performance in single-object and cluttered settings. At scale, cross-embodiment corpora and models (Open X-Embodiment/RT-X) demonstrate positive transfer across heterogeneous robots, motivating retargetable supervision from human interactions [36]. Specialized transfer frameworks extend this idea to dexterous bimanual manipulation via residual learning [23], while whole-body humanoid control benefits from motion-tracking pipelines distilled into guided diffusion policies that enable versatile downstream behaviors [35]. Together, these threads indicate a viable route from human HOI video to robot-usable policies via reconstruction, retargeting, and augmentation.

Embodied models and scalable data for generalist manipulation. Generalist *vision–language–action* policies pretrained on large video and robot corpora to enable instruction following and out-of-distribution generalization across tasks and embodiments [2, 3, 19, 53]. In parallel, imitation- and diffusion-based visuomotor learning emphasize stable training and multimodal action distributions, from classic action-diffusion policies to large diffusion foundation models that scale to bimanual control [9, 26]. To reduce data and hardware barriers, low-cost bimanual teleoperation systems provide dense demonstrations for fine-grained skills [50], while object-/pose-centric representations and semantic flows improve cross-object generalization and pose awareness [7]. At the dataset/benchmark layer, dual-arm generators and domain-randomized platforms supply scalable supervision with unified evaluation [6, 27]; open-instruction rearrangement benchmarks probe 6-DoF reasoning under language guidance [10]; and video-driven pipelines synthesize long-horizon tasks directly from Internet videos [44]. Recent work on task-centric *latent actions* further mitigates embodiment mismatch by learning instruction-conditioned action spaces transferable across robots [4].

3. Method

3.1. System Overview

We build a systematic pipeline covering in-the-wild hand-object interaction(HOI) videos into robot-usable supervision data. Our pipeline overview is illustrated in Fig. 2.

3.2. HOI Reconstruction from RGB(D) videos

Problem setup. Given video frames $\{I_t\}_{t=1}^{T}$, our goal is to recover metrically consistent trajectories together with parametric representation of the hand and mesh of the manipulated object in the same world coordinate, while (i) preventing hand–object interpenetration and (ii) enforcing physically plausible, temporally stable contact. Concretely, the state of the hand pose at time t is,

$$\mathbf{h}_t = (\theta_h(t), \mathbf{R}_h^w(t), \mathbf{t}_h^w(t)), \tag{1}$$

where $\theta_h(t)$ is the hand pose, $\mathbf{R}_h^w(t)$ and $\mathbf{t}_h^w(t)$ are the global transform and wrist of hands in the world coordinate. The object state is the rigid 6D pose tied to its (scale-resolved) geometry, $\mathbf{p}_t = T_o^w(t) \in \mathtt{SE}(3)$, defines the location and rotation of the object.

Human and hand motion recovery. Our method initially determines whether a clip implies *hand-only* or *whole-body* motion. For the *hand-only* case, we estimate $\mathbf{h}_t$ per frame using [28]. Otherwise, we estimate the SMPL-H parameters via [48] and directly produce the world coordinate

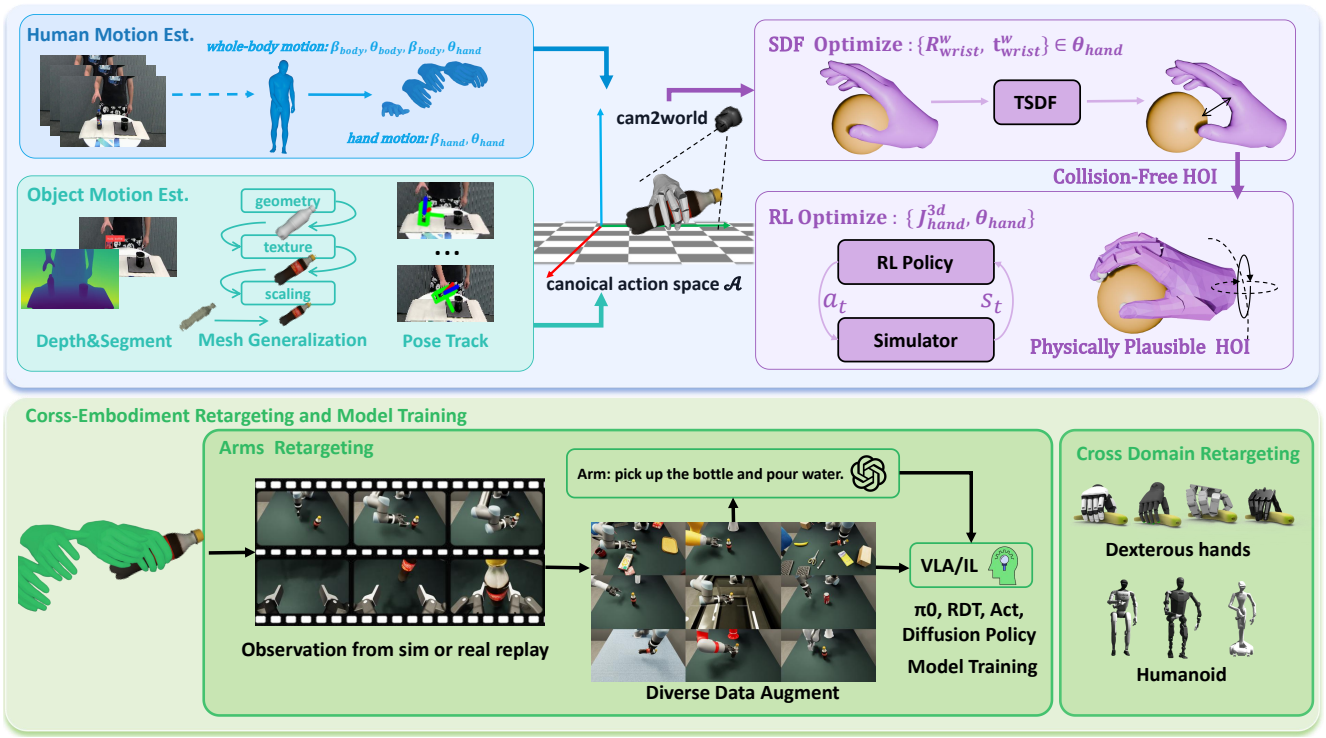


Figure 2. Given monocular RGB(-D) input, we first estimate the motion of the hand or wholebody and the manipulated object. We then perform a joint optimization, guided by TSDF and reinforcement learning, to improve physical plausibility and ensure robotic reachability. The resulting trajectories are retargeted to heterogeneous embodiments—including arms, dexterous hands, and humanoids. Finally, domain randomization for both observation and trajectories in Isaac Sim is applied to enrich observational diversity for robotic arms, and the generated embodied data is validated across both VLA and IL policy benchmarks.

body pose $\theta_b(t)$ and the shape β_b, equivalently extracting the hand state $\mathbf{h}_t$.

Object reconstruction and pose estimation. We ground the manipulated object, obtaining the per-frame mask m_t and depth D_t (predicted by [29] or RGB-D) in the video. Conditioned on semantic cues, we use a multiview 3D generator $\mathcal{G}$ [51] to produce an unscaled textured mesh $\hat{M}_o$. Then we recover the metric scale of the manipulated object by back-projecting the depth map inside the mask to a point set $\mathcal{P}_t = \{X_c(p) = D_t(p)\, K^{-1}\tilde{p} \mid p \in m_t\}$, aggregating as $\mathcal{P} = \bigcup_t \mathcal{P}_t$. We obtain the rescaled object mesh M_o via $M_o = s_o \hat{M}_o$, where s_o denotes the ratio between the diagonals of the bounding boxes of $\mathcal{P}$ and $\hat{M}_o$. With (M_o, M_t, D_t), a correspondence-driven tracker $\mathcal{F}(\cdot)$ [39] estimates the pose stream of the camera frame object $T^c_o(t)$.

To eliminate viewpoint-dependent inconsistencies in real-world HOI videos, we first estimate the camera intrinsics K and the camera-to-world transformation $T^w_c = (R_{wc}, t_{wc})$ using [34]. This allows us to transform all reconstructed hand-object interactions to the world coordinate system. We then align the resulting trajectories to a **canonical action space** $\mathcal{A}$ by constructing a reference frame based on body joint positions, ensuring consistency across heterogeneous sources. For detailed transformation steps, please refer to Appendix.

Optimization for physical plausibility. Let $\phi_o(\mathbf{x}; t)$ be a watertight object Truncated Signed Distance Function (TSDF, positive outside), and hand vertices V_h. First, we optimize the hand parameter $\mathbf{t}^w_h$ to avoid penetration between object and hand-palm by minimizing $\phi^2_o(V^{palm}_h, t)$.

As shown in the upper-right corner of Fig. 2, we then optimize the hand parameters $\mathbf{R}^{\mathbf{w}}_{\mathbf{wrist}}, \mathbf{t}^{\mathbf{w}}_{\mathbf{wrist}} \in \boldsymbol{\theta}_h$, to avoid hand–object interpenetration. Detailed collision optimization strategy is described in the supplementary material.

With collision-free HOI initialization, we introduce a residual RL policy [22] to obtain physically plausible hand–object poses while ensuring reachability on the robot. Specifically, a residual learning strategy is applied to refine the trajectories of both the hand and the object. The RL policy then encourages accurate tracking of these refined trajectories while promoting physically consistent contact once the hand–object distance drops below a threshold. The resulting state after RL is denoted as $s_t = (h_t, p_t, \dot{h}_t, \dot{p}_t, \mathcal{C}_t)$. Here,$h_t$, p_t denote the hand poses and object poses, respectively, along with their corresponding velocities and $\mathcal{C}_t$ denotes the contact force.Reward function r_t is defined as follows:

$$\begin{aligned} r_t = {} & \lambda_{\text{geo}}\Phi_{\text{geo}}\left(\|\Delta h_t\| + \|\Delta p_t\|\right) \\ & \lambda_{\text{dyn}}\Phi_{\text{dyn}}\left(\|\Delta \dot{h}_t\| + \|\Delta \dot{p}_t\|\right) + \lambda_{\text{con}}\Phi_{\text{con}}\left(\mathcal{C}_t\right), \end{aligned} \tag{2}$$

where Φ denotes the reward function and Δ denotes the er-

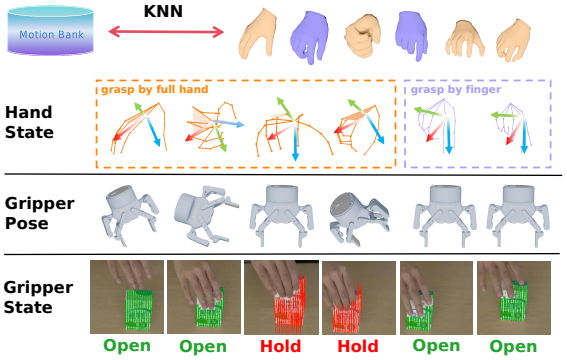


Figure 3. Map the hand's joints to the gripper's end-effector pose, including the corresponding mapping of the gripper's opening and closing state.

3.3. Cross-Embodiment Retargeting

Based on the physically plausible joint HOI reconstruction in Section 3.2, we obtained physically plausible trajectories $\{h_t, p_t\}_{t=1}^{T}$ and ensured stable hand-object contacts. We aim to retarget these to heterogeneous robot embodiments—industrial arms, dexterous hands, and humanoids.

Robot arms. Given accurate 3D hand joints, we retarget hand poses into executable end-effector poses $\{T_g(t), g(t)\}_{t=1}^{T}$ for a parallel-jaw gripper (Fig. 3). Inspired by [21], we implement two complementary orientation constructions depending on whether the *whole hand* (palm-involved) or *only finger tips* dominate the contact geometry. We use a kNN classifier to determine the gesture category.*Whole-hand* retargeting builds a stable palm frame from MCP joints to suppress fingertip jitter; *finger-only* mapping aligns to a hand-intrinsic frame and uses the index–thumb chord to define the gripper axis. For the detailed retargeting algorithm, please refer to the Appendix.To assess the state of the gripper, we employ CoTracker [17] to track the motion trajectories of key points on the manipulated object. The gripper state is determined based on the displacement of these key points.If these points remain stationary, the gripper is classified as open; otherwise, it is classified as closed. By focusing on keypoints rather than the object's mask, this approach achieves robustness to the visual ambiguity induced by severe occlusions during manipulation. We subsequently conducted real-robot experiments on a UR5 robotic arm equipped with a parallel gripper to validate the effectiveness of this retargeting method.

Dexterous hands and humanoids. Beyond retargeting to simple gripper-based arms, our high-fidelity HOI reconstruction offers a glimpse into transferring human demonstrations to more complex embodiments, including dexterous hands and humanoid robots. For dexterous hands, we retarget the reconstructed hand motions to the joint space of target robotic hands using kinematic similarity and contact-preserving constraints. For whole-body human demonstrations, we extend retargeting to humanoid platforms by leveraging full-body SMPL-H estimates. The resulting motion sequences are adapted to humanoid joint trees through inverse kinematics and dynamics-aware optimization, ensuring physical plausibility and intent preservation. Preliminary validations are provided in the Appendix.

Replay Assessment and Language Caption When replaying trajectories in simulation, executions may fail due to precision or non-physical collisions in the physics engine. To mitigate this, we employ Qwen2.5 VL [1] for automatic binary task evaluation (success/failure). The model takes as input the task description along with both wrist- and third-view simulated observations to determine whether the action has been successfully completed. For trajectories deemed successful, we further derive and refine fine-grained language instructions using [16].

3.4. Data Augmentation in Simulation

While preserving contact semantics and physical plausibility that are critical for embodiment control, we perform extensive simulation-based data augmentation to enrich the diversity of visual observations and motion trajectories. All augmentation strategies are applied within the canonical action space $\mathcal{A}$ to handle different manipulation orientations. More visualizations are provided in the Appendix.

Different types of arm retargeting. Given an executable end-effector (EE) trajectory $\{T_g(t)\}$ produced by our retargeting method, we generate observations for heterogeneous arms, as illustrated in Fig. 4. In Isaac Sim, we instantiate five widely used $6/7$-DoF robotic arms as simulation assets, including *UR5/UR5e*, *Franka Emika Panda*, *KUKA LBR iiwa 7*, *Kinova Gen3*, and *Rethink Robotics Sawyer*. HOI-derived 6D EE trajectories $T_g(t) \in \mathtt{SE}(3)$, $t = 1, \ldots, T$, are mapped into feasible joint trajectories using cuRobo's GPU-accelerated inverse kinematic (IK) backend [32]. For each robotic arm, at every timestep we invoke IK solver with the target pose $T_g(t)$. The solver returns a feasible joint configuration:

$$\begin{aligned} q_t \;&=\; \arg\min_{q} \mathcal{C}_{\text{goal}}(T_g(t), q) \\ \text{s.t.} q_{\min} &\preceq q \preceq q_{\max},\; \mathcal{C}_{\text{coll}}(q) \le 0, \end{aligned} \tag{3}$$

where $\mathcal{C}_{\text{goal}}$ is cuRobo's pose reaching cost and $\mathcal{C}_{\text{coll}}$ is the self-collision constraint. To encourage temporal consistency, we use the previous solution q_{t-1} as the IK seed when invoking the solver.

Episodes that pass the replay check preserve the original HOI intent (*e.g.*, grasp/place/pour) while providing embodiment diversity in joint space. We export both the joint-space commands $\{q_t\}_{t=1}^{T}$ (arm and gripper included) and aligned operational-space labels per robot, enabling multi-morphology policy training from the same HOI source.

Object retrieval and replacement. We build a large object library by combining [51] generations with in-house scans; each asset includes a watertight mesh, texture, category tag, and a canonical pose. For a source episode with

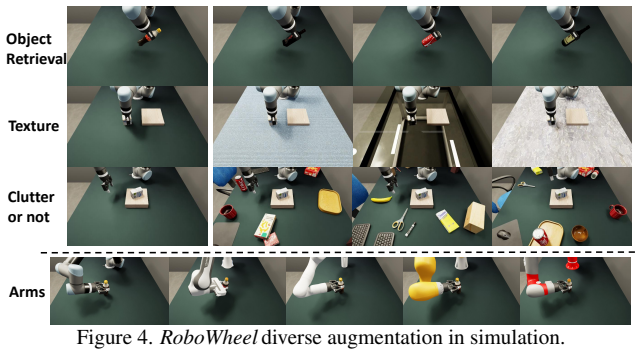


Figure 4. *RoboWheel* diverse augmentation in simulation.

object mesh M_o and object pose stream $\{T_o(t)\}$, we retrieve top-K substitutes $\tilde{M} = \{\tilde{M}_k\}$ using a fused similarity,

$$\mathcal{S}(M_o, \tilde{M}) = \alpha\, \mathrm{CD}\big(\hat{M}_o, \hat{\tilde{M}}\big) + \beta\,\big(1 - \mathrm{IoU}_{\mathrm{AABB}}\big) + \gamma\,\big\langle \phi_{\mathrm{sem}}(M_o),\, \phi_{\mathrm{sem}}(\tilde{M})\big\rangle, \tag{4}$$

where $\hat{\cdot}$ denotes unit AABB normalization, CD is the symmetric Chamfer distance on surface samples, $\mathrm{IoU}_{\mathrm{AABB}}$ measures coarse shape compatibility, and ϕ_{sem} are text–shape embeddings.

To ensure replay compatibility on a retrieved substitute, we align principal axes and bind the same maximum AABB and canonical pose definition as the source M_o. Under this binding, the original EE motion plan and the hand–object interaction geometry remain consistent, so the control trajectory can be directly replayed on geometrically/semantically matched novel objects (*e.g.*, mug↔cup-with-handle, box↔carton).

Trajectory augmentation. Informed by [41] and tailored to our setting, we represent each demonstration as a trajectory $\tau = \{(T_g(t), g(t))\}_{t=1}^{T}$, where $T_g(t) = (R(t), p(t)) \in \mathtt{SE}(3)$ denotes the EE pose with orientation $R(t)$ and translation $p(t)$, and $g(t)$ is the gripper command. The trajectory is partitioned into object-centric segments $\{\tau^{(k)}\}$, each labeled by a contact state $c^{(k)} \in \{\mathtt{hold}, \mathtt{open}\}$. Instead of re-planning trajectories, we augment them as follows.

(i) For interaction segments ($c^{(k)} = \mathtt{hold}$), we apply an object-frame rigid transform $T_o \in \mathtt{SE}(3)$ to each waypoint:

$$\tilde{T}_g(t) = T_o T_g(t), \qquad \tilde{g}(t) = g(t).$$

Let $R_\Delta := \mathtt{Rot}(T_o)$. To maintain continuity without motion-plan regeneration, the same EE orientation change is applied to non-interaction segments (see (ii)), and the orientation change induced by R_Δ is kept small for IK feasibility and repeatable execution.

(ii) For each non-interaction segment ($c^{(k)} = \mathtt{open}$), we linearly remap the translational path and set the EE orientation as $\tilde{R}(t) = R_\Delta R(t)$. Let p_s, p_e be the original endpoints and $\hat{p}_s, \hat{p}_e$ the remapped anchors: the anchor adjacent to a transformed interaction segment is fixed by that segment, while the opposite anchor is chosen within a predefined reachable set. With $\alpha_t \in [0,1]$ denoting the normalized progress along the original segment from p_s to p_e,

$$\tilde{p}_t = \hat{p}_s + \alpha_t(\hat{p}_e - \hat{p}_s) + \Big[p_t - \big(p_s + \alpha_t(p_e - p_s)\big)\Big]. \tag{5}$$

4. Dataset

Using the *RoboWheel* pipeline, we build a large-scale multimodal HOI-robotic dataset ***HORA***. The dataset integrates multiple data sources and leverages successive stages of the *RoboWheel* to convert heterogeneous inputs into a unified hoi representation, followed by cross-embodiment retargeting and diversified data augmentation. Data sources include: (i) a custom multi-view mocap system equipped with tactile sensor–instrumented gloves, (ii) multiple public HOI datasets, and (iii) a custom RGB(D) video capture setup. For source (i), high-precision HOI signals are obtained via triangulation and FoundationStereo [40], while the remaining robot-relevant modalities are generated using the *RoboWheel* retargeting and augmentation pipeline. For source (ii), existing HOI modalities in public datasets are directly convert to the canonical action space $\mathcal{A}$, then retargeted and augmented in the same manner. For source (iii), the full *RoboWheel* pipeline processes raw videos to extract both HOI and embodied modalities. For detailed information on the tasks included in each subset, please refer to our Appendix. The following sections detail the modality composition and statistics of *HORA* , as well as the acquisition system for the **mocap subset**.

4.1. Modalities Composition and Statistics.

As summarized in Tab. 1, *RoboWheel* produces a unified multimodal representation across the three subsets of *HORA*. These modalities can be broadly grouped into two categories: those for HOI downstream tasks and those for robot-related tasks. The HOI modalities include sequences of hand MANO parameters (pose/shape with global orientation and translation) in world frame from the original data source, object 6dof pose, object assets, and hand–object contact annotations. The robot-related modalities include observations from the robot wrist and third-person viewpoints, as well as end-effector pose trajectories for the robotic arm. All trajectories are transformed into the canonical action space $\mathcal{A}$.

The **mocap subset** provides tactile signals, robotic observations, and HOI annotations. Concretely, we release the 6-DoF object pose, object assets, and a dense tactile map for both the hand and the object. The **recording subset** provides the same HOI and robotic modalities, except for tactile signals, because the collection setup did not include the hardware. The **public HOI subset** is constructed from existing HOI datasets [5, 15, 24, 33, 37] and currently includes the cross-embodiments modalities retargeted to different arms (6/7-DoF). The three subsets together constitute roughly 150k trajectories, distributed shown in Fig. 5.

4.2. Custom-built Mocap system

Acquisition System Our high-precision mocap subset is collected with a synchronized, multi-sensor capture rig, the data acquisition setup and tactile glove are shown in Fig. 6.

Dataset	Tactile	Robotic data	HOI data	# Trajectories	Object Info Granularity
GRAB [33]	✗	✗	✓	1334	Body & 6-DoF pose & Contact maps
HO3D (v3) [15]	✗	✗	✓	68	6-DoF pose & Assets
DexYCB [5]	✗	✗	✓	1,000	6-DoF pose & Assets
HO-Cap [37]	✗	✗	✓	∼64	6-DoF pose & Assets
TACO [24]	✗	✗	✓	2,500	Assets
DROID [18]	✗	✓	✓	76,000	Assets
LIBERO [25]	✗	✓	✗	366	Assets
UCSD Kitchen [42]	✗	✓	✗	150	■
***HORA*(Recordings)**	✗	✓	✓	23560	6-DoF object pose & Assets
***HORA*(Mocap)**	✓	✓	✓	63141	6-DoF object pose & Assets & Tactile map
***HORA*(Public Dataset)**	✗	✓	✓	66924	■

Table 1. Modalities and scale comparison. For *HORA*, the first three rows correspond to our mocap, recorded RGB(D), and public HOI subsets. "✗" and "✓" denote absence and presence, respectively.

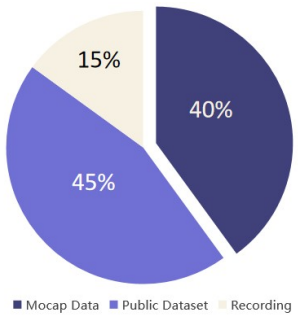


Figure 5. Composition of the three subsets in *HORA*.



Figure 6. Data collection setup and tactile information.

We use three Intel RealSense D455 RGB-D cameras connected via hardware sync cables to obtain time-aligned, metric depth from multiple viewpoints, and eight additional RGB cameras, yielding up to eleven views for each interaction. All cameras are extrinsically calibrated into a common world coordinate frame and temporally synchronized with the hand instrumentation.

Hand motion and tactile sensing are recorded using **EVT2** developed by **Paxini**. The glove integrates

- **29 magnetic encoders** distributed over the hand kinematic chain (wrist, palm, and phalanges), which provide joint-level pose measurements.
- **16 Gen3 tactile (haptic) sensors** mounted on key contact regions (fingertips, finger pads, and palm), which measure local normal forces.

These signals are used both for fitting MANO parameters and for deriving tactile and contact annotations that we release as part of the dataset.

Fitting to MANO with tactile signals We develop a tactile-guided, multi-constraint optimization scheme for motion retargeting. By registering glove measurements to the MANO hand model (or a target robot URDF) and jointly enforcing constraints on tactile contact states, wrist pose calibration, joint smoothness, and anatomical priors, the system optimizes all frames in parallel to produce spatiotemporally consistent, tactile-synchronized, and physically plausible motion trajectories. See Appendix for implementation details of the retargeting algorithm.

5. Experiment

5.1. HOI Reconstruction Quality

We evaluate HOI reconstruction quality on [37] with common metrics in Tab. 2 and Fig. 7. We compare *RoboWheel* with HOLD [50]. Both methods receive the same camera parameters and object meshes for a fair comparison.

Method	Object			Hand		Rel. pose cons.↓		Replay↑
	CD↓	F5↑	F10↑	jitter↓	W-MPJPE↓	Trans.↓	Rot↓	SR↑
HOLD	7.5	53.2	77.9	3.47	20.59	2.44	-	-
Ours (w/o TSDF, w/o RL)	5.4	59.2	83.2	3.34	7.74	2.05	5.1	78.7
Ours (TSDF only)	5.4	59.2	83.2	0.84	10.15	0.67	3.2	85.2
Ours (RL only)	6.2	53.1	75.9	4.13	7.89	0.46	3.1	61.7
Ours (TSDF + RL)	**5.1**	**63.4**	**89.1**	**0.92**	**7.81**	**0.26**	**1.9**	**93.6** ↑14.9

Table 2. Module ablations.

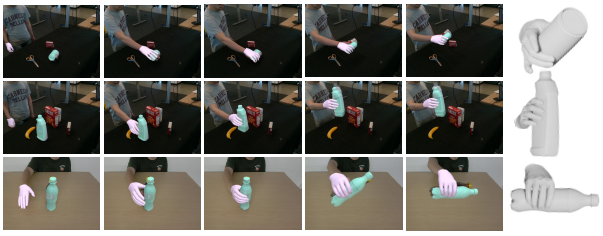


Figure 7. HOI reconstruction results of *RoboWheel* .Whether the data comes from public HOI datasets (*e.g.*, DexYCB) or not, *RoboWheel* can achieve high-precision HOI reconstruction.

5.2. Validation of *HORA* Data

Here we show how *HORA* data support downstream tasks and how well learned skills transfer with data augmentation.

Performance on different VLA/IL models. To study how Robowheel reconstructions translate to downstream control, we benchmark eight household tasks grouped by difficulty (Easy/Hard) and evaluate several VLA/IL algorithms (ACT, DP, RDT, Pi0). We evaluate each algorithm under three training regimes: (i) fine-tuning on 10 teleoperation demonstrations (tele.), (ii) fine-tuning on 10 *HORA* trajectories (*HORA*), and (iii) a two-stage curriculum that first pre-trains on 5k *HORA* trajectories and then fine-tunes on 10 *HORA* trajectories (reported as RDT+5k*HORA* and Pi0+5k*HORA*).The third training regime is applied only to RDT and Pi0, as the other methods do not support pre-training. Per-task success rate(%) is reported in the same real setup in Tab. 3. Macro averages within each difficulty group are reported. See Appendix for more results.

As shown in Tab. 3 and Fig. 8, VLA policies (Pi0 & RDT) pretrained with 5k *HORA* data achieved the highest success rates, demonstrating a remarkable performance

Real-world Tasks	Unseen Object		Clutter Objects		Unseen Background	
	HORA	***HORA*-aug**	***HORA***	***HORA*-aug**	***HORA***	***HORA*-aug**
Lift wooden cup	4/10	**5/10**	4/10	**4/10**	0/10	**4/10**
Place milk	5/10	**6/10**	5/10	**6/10**	1/10	**3/10**
Upright milk	3/10	**3/10**	4/10	**4/10**	3/10	**5/10**
Pour cola	3/10	**3/10**	3/10	**4/10**	2/10	**4/10**
Average	3.75/10	**4.25/10(+0.5)**	4.00/10	**4.50/10(+0.5)**	1.50/10	**4.00/10(+2.5)**

Table 4. Task performance in unseen scenarios with RDT.

Task	Ours	GAT-Grasp	yoto
Lift wooden cup	**87.5**	25.0	75.0
Place bowl	**100**	62.5	100
Upritght milk	**100**	50.0	62.5
Pour cola	**100**	62.5	75.0
move banana	**87.5**	62.5	87.5
Tip teacup	**87.5**	37.5	00.0
Macro avg	**91.7**	50.0	66.7

Table 5. Robot **direct replay** success rate ↑ (%).

Real-world Tasks	Diff.	ACT tele. | *HORA*	DP tele. | *HORA*	RDT tele. | *HORA*	Pi0 tele. | *HORA*	RDT+5k*HORA* *HORA*	Pi0+5k*HORA* *HORA*
Pick up milk	Easy	15.0 | 0.0	20.0 | 15.0	55.0 | 30.0	70.0 | 65.0	**70.0**	**80.0**
Lift wooden cup		0.0 | 0.0	45.0 | 25.0	75.0 | 45.0	65.0 | 55.0	**70.0**	**70.0**
Place milk		35.0 | 0.0	50.0 | 35.0	70.0 | 50.0	80.0 | 55.0	**85.0**	**80.0**
Restore bowl		0.0 | 0.0	5.0 | 0.0	65.0 | 65.0	60.0 | 60.0	**75.0**	**75.0**
Average		12.5 | 0.0	30.0 | 18.8	66.3 | 47.5	68.8 | 58.8	**75.0**	**76.3**
Move banana	Hard	0.0 | 0.0	5.0 | 0.0	55.0 | 20.0	40.0 | 15.0	**65.0**	**60.0**
Upright milk		0.0 | 0.0	0.0 | 15.0	45.0 | 30.0	60.0 | 50.0	**60.0**	**75.0**
Pour cola		0.0 | 0.0	0.0 | 10.0	35.0 | 35.0	25.0 | 35.0	**40.0**	**55.0**
Tip teacup		0.0 | 0.0	0.0 | 0.0	5.0 | 15.0	35.0 | 25.0	**25.0**	**45.0**
Average		0.0 | 0.0	1.3 | 6.3	35.0 | 25.0	40.0 | 31.3	**47.5**	**58.8**

Table 3. Real-world task performance (%) grouped by difficulty.

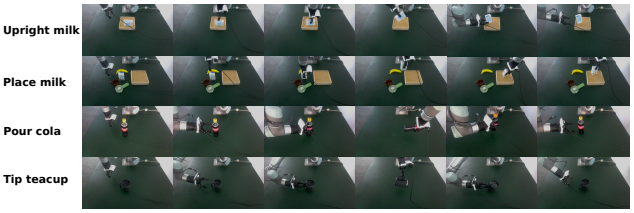


Figure 8. Real-world Performance on 4 Tasks.

improvement over counterparts without pretraining. The impact of this improvement is more evident in tasks of higher complexity. Notably, when training or fine-tuning these VLA/IL methods with an identical number of training episodes from teleoperation or *HORA*, policies trained solely on *HORA* achieve performance comparable to those trained on teleoperation data, despite the sim-to-real gap. The underlying reason for this is twofold. First, *RoboWheel* provides precise HOI reconstruction, enabling the generation of trajectories with accuracy approaching that of teleoperation, ensuring effective transfer from simulation to real-world environments. Second, data augmentation is applied to *HORA*, and its broader data distribution helps the policy mitigate the negative impact of the visual domain gap, endowing the policies with enhanced robustness.

Impact of *HORA* on Model Robustness To assess whether the data augmentation strategies in *HORA* enhance the robustness of VLAs,we compare RDT fine-tuned on *HORA* only (*HORA*) versus *HORA* with augmentation (*HORA*-aug). Specifically,we select four tasks and conduct real-robot experiments under three distribution shifts: *unseen objects*, *unseen backgrounds*, and *cluttered scenes*. Each cell shows *successes/trials* for two independent runs, as shown in Tab. 4. When trained on the *HORA* without augmentation, the fine-tuned RDT still manages to achieve some successful trials when encountering unseen objects or inferring in cluttered environments. However, when the background is altered, inducing substantial changes in the observations, the model exhibits catastrophic degradation in performance. In contrast, the fine-tuned RDT with augmented *HORA* shows a significant improvement in handling new observations, particularly in the unseen background setting, where the success rate increased by 25%. These results demonstrate that training on *HORA*, particularly with augmentation, substantially enhances the robustness of VLA models to visual variations.

5.3. Real-robot replay performance comparison

Under the same hand joints sequence input, we retarget the motions to a two-finger gripper using both existing methods and our proposed approach as discussed in Sec. 3.3, executing an identical set of tasks and measuring success rates to quantitatively evaluate the robustness of each mapping scheme. As shown in Tab. 5, our retargeting method consistently achieves higher success rates across all manipulation tasks. Its refined formulation of gripper orientation proves more robust, enabling stable yet flexible execution under diverse and dynamic hand-gesture regimes.

6. Conclusion and Discussion

Conclusion. Our data engine, *RoboWheel*, transforms real-world HOI videos into cross-embodiment robotic learning supervision by integrating high-fidelity hand–object reconstruction, physically plausible trajectory optimization, and cross-embodiment retargeting. Leveraging a unified HOI representation that is inherently robust to existing embodiment variations, *RoboWheel* provides a scalable framework for generating robotic training data. The framework is further strengthened with a rich suite of simulation-based data augmentation strategies. Based on two self-developed hardware platforms and existing public HOI datasets, we have constructed a large-scale HOI-to-Robot dataset *HORA*, which supports VLA models and imitation-learning approaches. Furthermore, our study establishes the first quantitatively validated evidence that HOI data can serve as an effective upstream modality for training models across diverse downstream, embodiment-specific tasks.

Limitations and Future Work. The scope of our real-world cross-embodiment experiments—particularly those involving dexterous hands and humanoid remains limited. Expanding cross-domain validation in both real-world and simulated settings will be an important direction in future.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5

[2] Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y Galliker, et al. π0. 5: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. 3

[3] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 3

[4] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025. 3

[5] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9044–9053, 2021. 3, 6, 7

[6] Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Qiwei Liang, Zixuan Li, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025. 3

[7] Tianxing Chen, Yao Mu, Zhixuan Liang, Zanxin Chen, Shijia Peng, Qiangyu Chen, Mingkun Xu, Ruizhen Hu, Hongyuan Zhang, Xuelong Li, et al. G3flow: Generative 3d semantic flow for pose-aware and generalizable object manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1735–1744, 2025. 3

[8] Zerui Chen, Rolandos Alexandros Potamias, Shizhe Chen, and Cordelia Schmid. Hort: Monocular hand-held objects reconstruction with transformers. *arXiv preprint arXiv:2503.21313*, 2025. 3

[9] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023. 3

[10] Yufei Ding, Haoran Geng, Chaoyi Xu, Xiaomeng Fang, Jiazhao Zhang, Songlin Wei, Qiyu Dai, Zhizheng Zhang, and He Wang. Open6dor: Benchmarking open-instruction 6-dof object rearrangement and a vlm-based approach. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7359–7366. IEEE, 2024. 3

[11] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J Black, and Otmar Hilliges. Hold: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 494–504, 2024. 3

[12] Kristen Grauman, Michael Wray, Adriano Fragomeni, Jonathan P. N. Munro, Will Price, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, C. V. Jawahar, Kris M. Kitani, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Mike Zheng Shou, Antonio Torralba, Jitendra Malik, 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 (CVPR)*, pages 18973–18990, 2022. 3

[13] 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*, pages 19383–19400, 2024. 3

[14] 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 (CVPR)*, 2020. 3

[15] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Ho-3d_v3: Improving the accuracy of hand-object poses in the ho-3d dataset. *arXiv preprint arXiv:2107.00887*, 2021. 3, 6, 7

[16] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 5

[17] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024. 5

[18] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024. 7

[19] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 3

[20] Hedvig Kjellstrom, Javier Romero, and Danica Kragic. Visual recognition of grasps for human-to-robot mapping. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3192–3199. IEEE, 2008. 3

[21] Hedvig Kjellstrom, Javier Romero, and Danica Kragic. Visual recognition of grasps for human-to-robot mapping. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3192–3199. IEEE, 2008. 5

[22] Kailin Li, Puhao Li, Tengyu Liu, Yuyang Li, and Siyuan Huang. Maniptrans: Efficient dexterous bimanual manip-

ulation transfer via residual learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6991–7003, 2025. 4

[23] Kailin Li, Puhao Li, Tengyu Liu, Yuyang Li, and Siyuan Huang. Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6991–7003, 2025. 3

[24] Rongao Li, Jie Fu, Bo-Wen Zhang, Tao Huang, Zhihong Sun, Chen Lyu, Guang Liu, Zhi Jin, and Ge Li. Taco: Topics in algorithmic code generation dataset. *arXiv preprint arXiv:2312.14852*, 2023. 6, 7

[25] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023. 7

[26] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024. 3

[27] Yao Mu, Tianxing Chen, Shijia Peng, Zanxin Chen, Zeyu Gao, Yude Zou, Lunkai Lin, Zhiqiang Xie, and Ping Luo. Robotwin: Dual-arm robot benchmark with generative digital twins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[28] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024. 3

[29] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *arXiv preprint arXiv:2502.20110*, 2025. 4

[30] Aditya Prakash, Matthew Chang, Matthew Jin, and Saurabh Gupta. Learning hand-held object reconstruction from in-the-wild videos. *arXiv preprint*, 2023. 3

[31] Wentian Qu, Zhaopeng Cui, Yinda Zhang, Chenyu Meng, Cuixia Ma, Xiaoming Deng, and Hongan Wang. Novel-view synthesis and pose estimation for hand-object interaction from sparse views. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15100–15111, 2023. 3

[32] Balakumar Sundaralingam, Siva Kumar Sastry Hari, Adam Fishman, Caelan Garrett, Karl Van Wyk, Valts Blukis, Alexander Millane, Helen Oleynikova, Ankur Handa, Fabio Ramos, et al. Curobo: Parallelized collision-free robot motion generation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8112–8119. IEEE, 2023. 5

[33] 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*, pages 581–600. Springer, 2020. 3, 6, 7

[34] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021. 4

[35] Takara E. Truong, Qiayuan Liao, Xiaoyu Huang, Guy Tevet, C. Karen Liu, and Koushil Sreenath. Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. *arXiv preprint*, 2025. 3

[36] Quan Vuong, Sergey Levine, Homer Rich Walke, Karl Pertsch, Anikait Singh, Ria Doshi, Charles Xu, Jianlan Luo, Liam Tan, Dhruv Shah, et al. Open x-embodiment: Robotic learning datasets and rt-x models. In *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023*, 2023. 3

[37] 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. *arXiv preprint arXiv:2406.06843*, 2024. 3, 6, 7

[38] Ruixiang Wang, Huayi Zhou, Xinyue Yao, Guiliang Liu, and Kui Jia. Gat-grasp: Gesture-driven affordance transfer for task-aware robotic grasping. *arXiv preprint arXiv:2503.06227*, 2025. 3

[39] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17868–17879, 2024. 2, 4

[40] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5249–5260, 2025. 2, 6

[41] Zhengrong Xue, Shuying Deng, Zhenyang Chen, Yixuan Wang, Zhecheng Yuan, and Huazhe Xu. Demogen: Synthetic demonstration generation for data-efficient visuomotor policy learning. *arXiv preprint arXiv:2502.16932*, 2025. 6

[42] Ge Yan, Kris Wu, and Xiaolong Wang. ucsd kitchens dataset. 2023. 7

[43] Jie Yang, Bingliang Li, Fengyu Yang, Ailing Zeng, Lei Zhang, and Ruimao Zhang. Boosting human-object interaction detection with text-to-image diffusion model. *arXiv preprint arXiv:2305.12252*, 2023. 3

[44] Weirui Ye, Fangchen Liu, Zheng Ding, Yang Gao, Oleh Rybkin, and Pieter Abbeel. Video2policy: Scaling up manipulation tasks in simulation through internet videos. *arXiv preprint arXiv:2502.09886*, 2025. 3

[45] Wanqi Yin, Zhongang Cai, Ruisi Wang, Fanzhou Wang, Chen Wei, Haiyi Mei, Weiye Xiao, Zhitao Yang, Qingping Sun, Atsushi Yamashita, et al. Whac: World-grounded humans and cameras. In *European Conference on Computer Vision*, pages 20–37. Springer, 2024. 2

[46] Jiyao Zhang, Weiyao Huang, Bo Peng, Mingdong Wu, Fei Hu, Zijian Chen, Bo Zhao, and Hao Dong. Omni6dpose: A benchmark and model for universal 6d object pose estimation and tracking. In *European Conference on Computer Vision*, pages 199–216. Springer, 2024. 2

[47] Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. In *Proceedings of the*

Computer Vision and Pattern Recognition Conference, pages 1805–1815, 2025. 2

[48] Yuhong Zhang, Jing Lin, Ailing Zeng, Guanlin Wu, Shunlin Lu, Yurong Fu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x++: A large-scale multimodal 3d whole-body human motion dataset. *arXiv preprint arXiv:2501.05098*, 2025. 3

[49] Yuhong Zhang, Guanlin Wu, Ling-Hao Chen, Zhuokai Zhao, Jing Lin, Xiaoke Jiang, Jiamin Wu, Zhuoheng Li, Hao Frank Yang, Haoqian Wang, et al. Humanmm: Global human motion recovery from multi-shot videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1973–1983, 2025. 2

[50] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023. 3, 7

[51] Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025. 4, 5

[52] Huayi Zhou, Ruixiang Wang, Yunxin Tai, Yueci Deng, Guiliang Liu, and Kui Jia. You only teach once: Learn one-shot bimanual robotic manipulation from video demonstrations. *arXiv preprint arXiv:2501.14208*, 2025. 3

[53] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023. 3