HRDexDB: A Paired Human-Robot Dataset for Cross-Embodiment Dexterous Grasping

Jongbin Lim[1*], **Taeyun Ha**[1*], **Mingi Choi**[1], **Jisoo Kim**[1],
Byungjun Kim[1], **Subin Jeon**[1], **Hanbyul Joo**[1,2]
[1]Seoul National University [2]RLWRLD
{whdqls0534,taeyun012,willi19,jlogkim,byungjun.kim,subinjeon,hbjoo}@snu.ac.kr

https://snuvclab.github.io/HRDexDB/

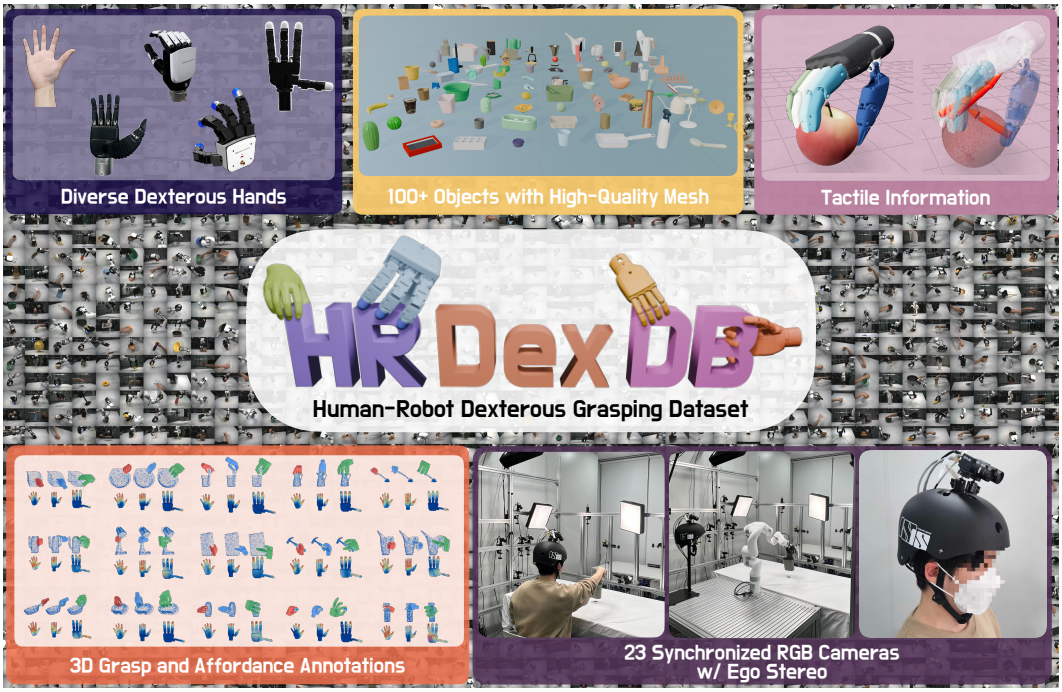


Figure 1: **Overview of HRDexDB.** HRDexDB contains paired human and robotic dexterous grasping episodes across 100 objects and multiple hand embodiments. Using a synchronized multi-view capture system, we record comparable human demonstrations and robotic executions with reconstructed 3D hand and robot trajectories, object 6D poses, egocentric observations, contact force signals from robotic hands equipped with tactile sensors, and success/failure annotations.

Abstract: We present **HRDexDB**, a paired cross-embodiment dexterous grasping dataset of high-fidelity dexterous grasping sequences featuring both human and diverse robotic hands. Unlike existing datasets, HRDexDB provides a comprehensive collection of grasping trajectories across human hands and multiple robot hand embodiments, spanning 100 diverse objects. Leveraging state-of-the-art vision methods and a dedicated multi-camera system, HRDexDB offers high-precision spatiotemporal 3D ground-truth motion for both the agent and the manipulated object. The dataset comprises 2.1K grasping trials, each enriched with synchronized visual and kinematic modalities, with contact-force signals available for tactile-enabled robotic hands. By providing closely aligned captures of human dexterity and robotic execution on the same target objects under comparable grasping motions, HRDexDB serves as a foundational benchmark for cross-embodiment dexterous manipulation.

Keywords: Cross-Embodiment, Human-to-Robot Learning, Dexterous Manipulation

1 Introduction

Enabling robots to achieve human-level dexterity is one of the central goals in robotics. Since many tools and objects in human environments are designed for the human hand, recent studies have explored anthropomorphic robotic hands that go beyond simple parallel grippers. Human manipulation therefore provides a natural source of demonstrations for robot learning, but transferring these demonstrations requires more than direct imitation. Human and robotic hands differ in morphology, kinematics, and actuation, and this embodiment gap also extends across robotic hands themselves: different dexterous hands impose embodiment-specific physical and kinematic constraints, resulting in distinct feasible contact patterns and grasp strategies. Determining how robots should learn from human manipulation and transfer grasp strategies across diverse hand embodiments therefore remains an open problem.

Despite substantial progress, existing datasets rarely provide paired, multi-embodiment captures of comparable grasping behavior over shared objects, limiting the study of how human dexterity transfers to diverse robotic hands. Most computer-vision datasets focus primarily on human hands, providing either isolated hand motion without objects [1], human hand-object interactions [2, 3, 4, 5, 6, 7, 8, 9], or large-scale egocentric RGB videos without 3D information [10, 11]. Robotic hand datasets [12, 13, 14, 15, 16], in contrast, focus on the robot side and often leave object motion only partially tracked. A few recent approaches attempt to collect paired human–robot data [17, 18, 19], but they do not provide paired captures across multiple dexterous robotic embodiments over shared objects, and often lack markerless RGB observations or tactile signals.

To address these limitations, we introduce **HRDexDB**, a paired cross-embodiment dexterous manipulation dataset that captures human hands and four dexterous robotic hand embodiments manipulating a shared set of 100 diverse objects. HRDexDB provides paired human and robotic grasping sequences with 21 synchronized exocentric RGB views, 2 egocentric views, scanned 3D object models, 3D human hand motion, robot states, object 6D pose trajectories, and tactile signals for tactile-enabled robotic hands. Acquiring such data is challenging because dexterous manipulation induces severe occlusions, making markerless hand reconstruction and object tracking difficult. We therefore build a synchronized capture and reconstruction pipeline with 21 calibrated exocentric cameras and 2 egocentric cameras. To the best of our knowledge, HRDexDB is the first dataset to provide paired human and multi-robot dexterous manipulation captures over shared objects with markerless multi-view RGB observations in a unified and paired manner.

We further demonstrate the value of HRDexDB through benchmarks in two directions. First, for human-to-robot transfer, we study *contact map transfer*, which converts human contact patterns into robot-specific contact maps, and *cross-embodiment grasp retrieval*, which learns a shared latent space for retrieving feasible robot grasp priors from human grasps. Second, we evaluate HRDexDB as a benchmark for perception under dexterous interaction. We test state-of-the-art 3D hand pose estimation and object 6D pose estimation methods on our captured sequences, where severe hand–object and robot–object occlusions make perception substantially more challenging. These experiments show that HRDexDB not only supports cross-embodiment transfer, but also provides useful training and evaluation signals for markerless hand–object perception.

In summary, our contributions are threefold: (1) We introduce HRDexDB, the first markerless paired human-robot dexterous manipulation dataset, capturing 100 objects with multi-view observations, 3D hand and object annotations, and tactile signals for tactile-enabled robotic hands. (2) We present a novel multi-camera capture system and integrated hardware and software solutions to address the substantial challenges of synchronized 3D hand-robot-object tracking and tactile acquisition. (3) We establish downstream benchmarks on HRDexDB, including human-to-robot contact map transfer, cross-embodiment grasp retrieval, 3D hand pose estimation, and object 6D pose estimation under dexterous grasping. At the time of submission, HRDexDB includes over 100 captured objects, with

Table 1: Comparison of Human-Object Interaction (HOI) and Robotics Datasets with HRDexDB

Dataset	Type	#Emb.	Dex Robot Hand	Modality	Views	Objs	Resolution	Seqs	Frames	Tactile	M-less	3D Hand	Obj 6D
FPHA[2]	HOI	1	–	RGB-D	1	26	1920×1080	1.2K	105K	×	×	✓	✓
ContactDB[3]	HOI	1	–	RGB-D	9	50	1920×1080	3.7K	375K	×	✓	×	✓
FreiHAND [4]	HOI	1	–	RGB	8	25	1280×1024	–	37K	×	✓	✓	×
Ho-3D[9]	HOI	1	–	RGB-D	5	10	640×480	68	103K	×	✓	✓	✓
DexYCB[5]	HOI	1	–	RGB-D	8	20	640×480	1K	582K	×	✓	✓	✓
HOI4D[6]	HOI	1	–	RGB-D	1	800	1280×720	4K	2.4M	×	✓	✓	✓
ARCTIC[7]	HOI	1	–	RGB	9	11	2800×2000	339	2.1M	×	×	✓	✓
TACO [20]	HOI	1	–	RGB/RGB-D	13	196	4096×3000	2.5K	5.2M	×	×	✓	✓
OakInk2[8]	HOI	1	–	RGB-D	16	75	840×480	627	1.34M	×	×	✓	✓
Contact4D[21]	HOI	1	–	RGB	19	N.A.	3840×2160	375	2M	✓	×	✓	✓
HOT3D[22]	HOI	1	–	RGB	2	33	1408×1408	425	3.7M	×	×	✓	✓
GigaHands[23]	HOI	1	–	RGB	51	417	1280×720	14K	183M	×	✓	✓	✓
RealDex[24]	ROI	1	✓	RGB-D	4	52	–	2.6K	955K	×	✓	✓	✓
RoboCOIN[14]	ROI	15	✓	RGB-D	3	432	–	180K	–	×	✓	×	×
AgiBotWorld[12]	ROI	1	✓	RGB-D	8	3000	–	1M	–	✓	✓	×	×
OXE[13]	ROI	22	×	RGB-D	3	–	–	1M	130M	×	✓	×	×
DROID[15]	ROI	1	×	RGB-D	3	–	1280×720	76K	56.7M	×	✓	×	×
RoboMIND[16]	ROI	4	✓	RGB-D	3	96	480×640	107K	–	×	✓	×	×
RH20T[17]	HROI	7	×	RGB-D	7	–	1280×720	220K	50M	✓	✓	×	×
DexWild[18]	HROI	2	✓	RGB-D	6	180	224×224	10K	–	×	×	×	×
H&R[19]	HROI	2	×	RGB-D	1	-	240×424	2.6K	1M	×	✓	×	×
HRDexDB (Ours)	**HROI**	5	✓	**RGB**	**23**	**100**	**2048×1536**	**2.1K**	**24M**	✓	✓	✓	✓

ongoing expansion toward 1,000 objects. The full dataset will be publicly released to facilitate future research in dexterous manipulation and robot learning.

2 Related Work

Human-Object Interaction Dataset. Human-object interaction datasets have enabled substantial progress in modeling hand articulation, object motion, and contact-rich manipulation. Early benchmarks such as FreiHAND [4], HO-3D [9], and DexYCB [5] focused on 3D hand-object pose estimation, while later datasets such as ARCTIC [7], HOT3D [22], HOI4D [6], GigaHands [23], TACO [20], Contact4D [21], and OakInk2 [8] expanded the scale, viewpoints, object diversity, contact annotations, and task complexity of hand-object interaction capture. Despite these advances, such datasets remain primarily human-centric and do not provide paired correspondence with robotic dexterous embodiments. In contrast, **HRDexDB** is designed to bridge human and robot dexterous grasping through paired captures over shared objects.

Robot-Object Interaction Dataset. Robot manipulation datasets have grown substantially in scale and diversity for robot learning. Large-scale efforts such as Open X-Embodiment [13] and DROID [15] aggregate diverse demonstrations across many tasks and environments, but much of this data is collected with relatively low-DoF grippers. More recent datasets such as AgiBot World [12], RoboMIND [16], and RoboCOIN [14] further expand scale and task diversity, supporting bimanual manipulation and demonstrations collected with both grippers and dexterous hands. However, existing datasets rarely provide direct correspondence between human grasp motion and robotic dexterous manipulation, especially across multiple robotic hand embodiments. RealDex [24] collects real-world dexterous robot grasping trajectories through teleoperation, but focuses on a single robotic hand and does not include separately captured human demonstrations. **HRDexDB** addresses this gap by providing paired human and robot grasping data over shared objects across multiple dexterous hand embodiments.

Human-Robot Correspondence Dataset. Recent datasets aim to associate human demonstrations with robotic executions to bridge the embodiment gap. RH20T [17] and H&R [19] provide task- or frame-level human-robot alignment, but are primarily based on parallel-jaw grippers, limiting their applicability to dexterous manipulation. DexWild [18] collects task-level aligned human and multi-finger robot demonstrations using a portable glove and camera system, but its tracking setup is vulnerable to occlusion and does not provide dense episode-wise behavioral alignment.

Synthetic Dexterous Grasp Datasets. Large-scale synthetic grasp datasets provide a complementary direction for learning dexterous manipulation priors. DexGraspNet [25] generates stable dexterous grasps in simulation, while GenDexGrasp/MultiDex [26] extends grasp synthesis to multiple robotic hand embodiments. GraspXL [27] scales grasp motion generation across human and robotic hands. However, these datasets are synthetic and primarily provide grasp priors rather than real

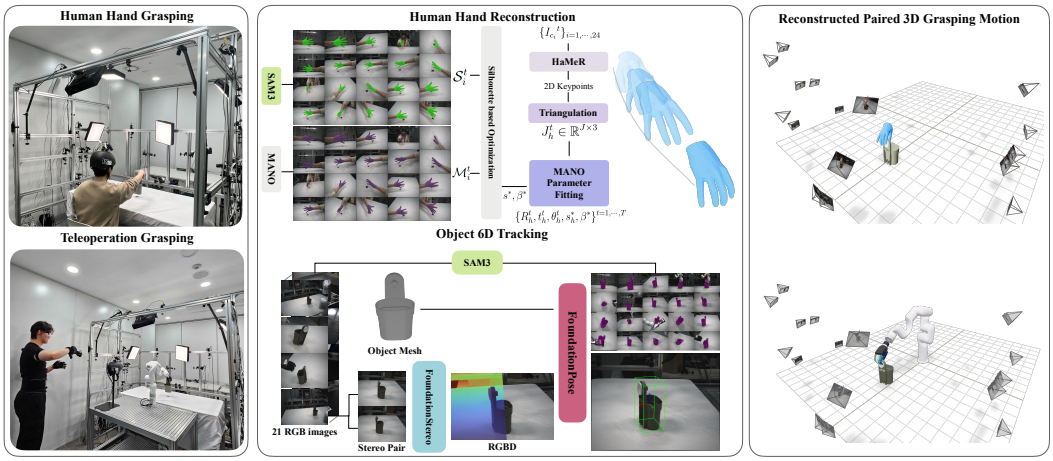


Figure 2: **Capture and Reconstruction Pipeline**. Multi-view recordings are processed to reconstruct hand motion and object 6D trajectories, producing aligned human and robot grasps.

human- and robot-object interaction trajectories. In contrast, **HRDexDB** captures paired real-world human and dexterous robot grasping data with synchronized multi-view observations, 3D annotations, and object 6D poses, and tactile signals.

3 Constructing HRDexDB

3.1 Dataset Overview

HRDexDB is constructed as a paired multi-modal dataset of dexterous grasping sequences performed by both human subjects and four robotic embodiments, spanning Allegro Hand V4 and V5 Plus, and Inspire Hand RH56DFTP and RH56F1. The dataset contains 24M frames and 2.1K sequences over 100 objects, including synchronized visual observations, kinematic states, reconstructed geometry, object 6D poses, and tactile signals when available. All spatial quantities are aligned in a unified world coordinate system defined by our calibrated multi-camera platform. A robotic grasping trial is represented as a time-indexed sequence

$$\mathcal{T}^{\text{robot}} = \left\{ \{\mathbf{I}_t^{c_i}\}_{c_i=1}^{21}, \mathbf{I}_t^{\text{ego}}, \boldsymbol{q}_t^{\text{robot}}, \boldsymbol{T}_t^{\text{object}}, \boldsymbol{F}_t^{\text{tactile}}, y \right\}_{t=1}^{T_r}. \tag{1}$$

Here, $\mathbf{I}_t^{1..21}$ and $\mathbf{I}_t^{\text{ego}}$ denote synchronized exocentric and egocentric RGB observations, $\boldsymbol{q}_t^{\text{robot}}$ denotes the robot state, $\boldsymbol{T}_t^{\text{object}} \in \text{SE}(3)$ represents the object 6D pose. Tactile signals $\boldsymbol{F}_t^{\text{tactile}}$ are measured from tactile-enabled robot fingertips, and $y \in \{0, 1\}$ indicates whether the grasp was successful. The total sequence length is denoted by T_r. Similarly, a human grasping trial is represented as

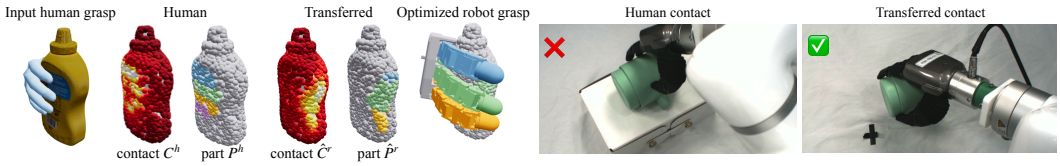
$$\mathcal{T}^{\text{human}} = \left\{ \{\mathbf{I}_t^{c_i}\}_{c_i=1}^{21}, \mathbf{I}_t^{\text{ego}}, \boldsymbol{\theta}_t^{\text{human}}, \boldsymbol{T}_t^{\text{object}}, y \right\}_{t=1}^{T_h}, \tag{2}$$

where $\boldsymbol{\theta}_t^{\text{human}} \in \mathbb{R}^{51}$ denotes MANO pose parameters and T_h is the human sequence length.

3.2 Multi-Modal Capture System and Paired Data Collection

Capture System. Our capture platform (Fig. 2) consists of a 21-camera RGB rig on a three-sided metal frame surrounding the workspace, enabling dense multi-view capture under severe hand–object occlusions, plus stereo egocentric views from an over-the-shoulder rig for robotic trials and a custom stereo helmet for human demonstrations. The robot is teleoperated using an Xsens inertial motion-capture suit and MANUS gloves, which map the operator's wrist and finger motions to the robot arm and robot hand.

Paired Acquisition Protocol. We collect paired human–robot grasps using a two-stage protocol under the same object and workspace conditions. A human subject first performs a natural grasp



(a) Human-to-robot contact transfer (b) Real-world execution

Figure 3: **Human-to-robot contact transfer and real-world grasping.** (a) Our model transfers the human contact and part maps (C^h, P^h) to robot-specific maps $(\hat{C}^r, \hat{P}^r)$, which serve as the optimization objective for grasp synthesis. (b) On the same object and input human grasp, the grasp optimized from the transferred map succeeds while the one from the human map fails.

on the target object, and the multi-view recordings are used to reconstruct the human hand motion and object trajectory. A teleoperator then observes the demonstration and performs a semantically corresponding grasp with the robotic embodiment, preserving the grasp intent while allowing embodiment-specific differences in morphology, kinematics, and timing.

3.3 Multi-Modal State Reconstruction

We process the synchronized recordings to reconstruct human hand motion, object 6D pose, and robot alignment within the unified world coordinate system.

Human Hand Reconstruction. To reconstruct 3D human hand motion, we employ the MANO parametric hand model [28]. Following the multi-view fitting strategy of GigaHands [23], we detect 2D hand keypoints in each calibrated view using HaMeR [29], triangulate 3D joints, and optimize MANO pose parameters for each frame. Subject-specific hand shape is calibrated using silhouette alignment with SAM3-generated masks [30], and temporal filtering is applied to reduce jitter.

Object 6D Tracking. To obtain accurate object poses $\boldsymbol{T}_t^{\text{object}} \in \text{SE}(3)$, we develop a model-based 6D tracking pipeline within the synchronized multi-view system. A designated calibrated stereo pair estimates dense depth maps using FoundationStereo [31], while SAM3 [30] provides object masks to localize the manipulated object. Given RGB-D observations and object CAD models, we perform 6D pose estimation using FoundationPose [32], initializing the pose in the first frame through global registration and refining subsequent frames via temporal tracking to ensure consistency. Since stereo-based tracking relies on a single viewpoint, we further enforce cross-view geometric consistency by rendering the object mesh into all calibrated camera views and minimizing silhouette misalignment across views, reducing drift during long-horizon manipulation.

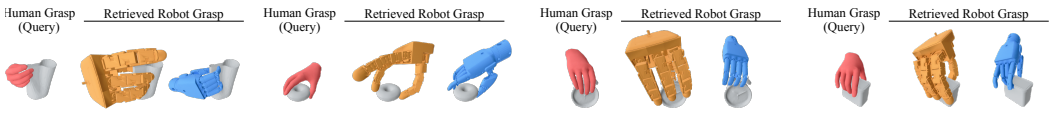
4 Applications & Experiments

We introduce two learning-based human-to-robot transfer baselines that leverage HRDexDB's paired human–robot grasp data: contact map transfer and cross-embodiment grasp retrieval. We further demonstrate the utility of HRDexDB as a challenging perception benchmark by evaluating 3D hand pose estimation and object 6D pose estimation under severe hand–object occlusions.

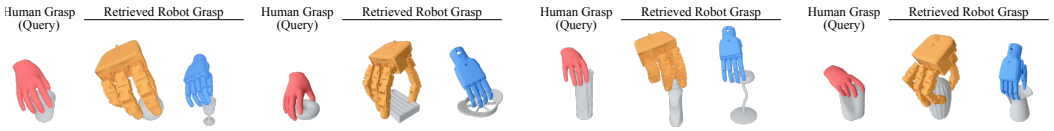
4.1 Human-to-Robot Contact Map Transfer

Task definition. Dexterous robotic hands often resemble human hands, yet directly imitating human contact patterns can be suboptimal due to differences in morphology and kinematics. Prior contact-map-based methods such as CEDex [33] rely on predefined human-to-robot contact correspondences for grasp synthesis. In contrast, HRDexDB enables a data-driven alternative: learning robot-specific contact maps directly from paired human–robot grasps. Given a human contact on an object, the goal is to predict a robot-specific contact representation that captures how successful contact strategies adapt across embodiments.

Experimental setup. We represent a grasp on an object point cloud $O \in \mathbb{R}^{N\times 3}$ with a contact map $C \in [0,1]^N$ of per-point contact probabilities and a part map P assigning contacted points to hand



(a) Same-Object Candidate Retrieval



(b) Cross-Object Candidate Retrieval

Figure 4: **Qualitative examples of human-conditioned robot grasp retrieval.** Given a human hand–object grasp query, the model retrieves robot grasp candidates from the learned embedding space. The same-object setting restricts candidates to the query object, while the cross-object setting retrieves candidates from training objects to evaluate whether the model can find compatible grasp priors for unseen test queries.

parts. The human part map is $P^h \in \mathbb{R}^{N\times 6}$, while the robot part map is $P^r \in \mathbb{R}^{N\times B}$, with $B = 6$ for the Inspire hand and $B = 5$ for the Allegro hand. Conditioned on the human representation $[C^h, P^h]$ and PointNet++ [34] object features, our model predicts the robot representation $[C^r, P^r]$, supervised with a contact-weighted L_1 loss on C^r and a cross-entropy loss on P^r over contacted points. We train a separate model for each target hand, using the Inspire RH56F1 and Allegro Hand V5 Plus.

Given the predicted robot contact representation, we synthesize grasps using the physics-aware optimization pipeline of CEDex [33], which combines contact, penetration, and self-collision terms. We compare *Human-Contact*, which optimizes against the captured human contact map, with *Transferred-Contact*, which optimizes against our predicted robot-specific contact map. The optimizer is fixed across both settings, isolating the effect of the contact objective. For the four-fingered Allegro hand, the baseline maps the fourth finger to the union of the human ring and little fingers.

We report grasp success rates in simulation and real-world experiments on unseen poses of seen objects. In simulation, we follow CEDex and apply forces along six orthogonal axes in Isaac Gym [35]. In the real world, a grasp succeeds if the object is lifted and held for ten seconds. We generate pre-grasp and squeeze motions with BODex [36] and compute execution trajectories with CuRobo [37].

Results. Table 2 summarizes grasp success rates under the two contact objectives. Transferred-Contact improves success over directly using human contact maps in both simulation and real hardware, showing that HRDexDB enables learning robot-specific contact strategies from paired human–robot grasps. Figure 3 further shows that the transferred maps preserve the functional intent of the human grasp while adapting the contact distribution to the target robot morphology.

Table 2: Grasp success rate in simulation and on real hardware. Both conditions share the same grasp optimizer and differ only in the source of the contact term. Sim trials: 1000/1000; real trials: 60/30 (Inspire/Allegro). Values in %.

Method	**Inspire**		**Allegro**	
	Sim ↑	**Real ↑**	**Sim ↑**	**Real ↑**
Human-Contact	54.6	66.7	60.2	63.3
Transferred (Ours)	**55.6**	**73.3**	**65.8**	**80.0**

4.2 Latent-Space Robot Grasp Retrieval

Task definition. Given paired human and robot grasp demonstrations, the goal is to learn a shared latent representation that aligns geometrically and functionally corresponding grasps across embodiments. At inference time, a human hand–object grasp and the corresponding object geometry are used as a query, and robot grasp candidates from HRDexDB are ranked by similarity in the learned embedding space. This retrieval formulation evaluates whether paired data can induce an embodiment-aware grasp representation, while selecting robot grasp priors that reflect feasible grasping patterns for the target embodiment.

Table 3: Cross-embodiment grasp retrieval performance over 33 candidate grasps.

Retrieval Direction	R@1	R@3	R@5
Human → Inspire	**36.36%**	**81.82%**	**100.00%**
Human → Allegro	24.24%	63.64%	72.73%
Inspire → Allegro	8.18%	57.58%	72.73%

Table 4: BODex refinement success under different initialization strategies.

Initialization Method	Seed-level (%) ↑		Episode-level (%) ↑	
	Inspire-F1	Allegro-v5	Inspire-F1	Allegro-v5
Vanilla	3.39	16.24	69.70	84.85
Kinematic Retargeting	3.52	1.21	42.42	30.30
Retrieval-top5	10.79	17.09	**75.76**	**93.94**
Retrieval-top1	**12.24**	**21.33**	57.58	75.76

Experimental setup. We implement this task with a CLIP-style multi-branch retrieval model [38]. The model consists of separate point-cloud encoders for the human hand, Inspire-F1 hand, and Allegro-V5 hand, together with a shared object encoder. For each retrieval direction, the corresponding query and candidate branches are projected into a shared embedding space and trained with a symmetric contrastive loss so that paired cross-embodiment grasps are close.

We evaluate the learned representation in two ways. First, we measure retrieval accuracy by ranking robot grasp candidates according to their similarity to a human hand–object query. Second, we test whether the retrieved robot grasps provide useful initialization for downstream grasp optimization. For this evaluation, we initialize the fine stage of BODex [36] with retrieved grasps and compare against an AnyTeleop-style kinematic retargeting baseline [39], using the same BODex refinement backend and MuJoCo evaluation protocol. We report success on 33 episodes across 7 unseen objects, with 50 optimization seeds per episode. Retrieval-top1 initializes all seeds from the highest-ranked grasp, while Retrieval-top5 distributes the same seed budget across the five highest-ranked grasps.

Results. Table 3 reports retrieval accuracy over 33 candidate grasps. The learned embedding retrieves paired robot grasps substantially above random, indicating that HRDexDB supports learning an embodiment-aware latent representation from paired human–robot demonstrations. Figure 4 provides qualitative examples in both same-object and cross-object retrieval settings.

Table 4 summarizes the downstream BODex refinement results. Retrieval-based initialization improves over Vanilla BODex by providing stronger local grasp priors, and outperforms kinematic retargeting by avoiding direct human-to-robot pose transfer under embodiment mismatch. Retrieval-top1 yields the highest seed success, whereas Retrieval-top5 yields the highest episode success, reflecting a precision–coverage trade-off between the best single prior and multiple candidate priors.

Table 5: Hand pose estimation accuracy on our dataset vs. FreiHAND [4]. All metrics are in mm. Δ = Ours − FreiHAND; positive values indicate that our benchmark is more challenging.

	Our Dataset		FreiHAND [4]		Δ	
Model	PA-MPJPE ↓	PA-MPVPE ↓	PA-MPJPE ↓	PA-MPVPE ↓	PA-MPJPE	PA-MPVPE
WiLoR [40]	5.94	6.09	5.71	5.27	**+0.23**	**+0.82**
HaMeR [29]	6.15	6.16	6.11	5.72	**+0.04**	**+0.44**
Hamba [41]	6.11	6.10	6.14	5.84	−0.03	**+0.26**
MeshGraphormer [42]	8.31	8.10	6.64	6.78	**+1.67**	**+1.32**
FrankMocap [43]	10.61	12.48	9.52	11.64	**+1.09**	**+0.84**

Table 6: Effect of mixing HRDexDB hand pose data into finetuning, evaluated on FreiHAND [4].

	PA-MPJPE		PA-MPVPE	
Method	Baseline	+ Ours	Baseline	+ Ours
HaMeR	6.108	**6.027**	5.718	**5.679**
WiLoR	5.711	**5.677**	5.273	**5.260**

4.3 Benchmarking 3D Hand Pose Estimation

Accurate 3D hand pose and mesh estimation is central to dexterous manipulation and learning from human demonstrations. To assess this capability, we evaluate state-of-the-art hand reconstruction methods on our dataset, which provides synchronized multi-view RGB of hands manipulating objects with accurate 3D supervision.

Table 5 shows that all evaluated models incur consistently higher errors on our dataset than on FreiHAND [4], confirming that our benchmark poses a more challenging setting. A natural question is whether our data is not only harder to fit but also useful as a training signal.

To test if our data provides complementary signal at scale, we add 6k of our samples into the finetuning set and re-train two state-of-the-art hand reconstruction models. The combined set aggregates

Table 7: Object 6D pose estimation performance on paired human- and robot-grasp frames in HRDexDB. ADD is in cm, AR_{MSSD} in %, and Δ denotes Robot $-$ Human.

Method	Human		Robot		Δ	
	ADD (cm) ↓	AR_{MSSD} ↑	ADD (cm) ↓	AR_{MSSD} ↑	ADD (cm)	AR_{MSSD}
FoundPose	6.91	44.10	8.74	33.30	+1.83	-10.80
FoundPose + MegaPose	3.35	70.00	4.40	64.10	+1.05	-5.90
GigaPose	13.10	19.70	13.80	17.30	+0.70	-2.40
GigaPose + MegaPose	5.99	54.10	8.02	49.20	+2.03	-4.90
PicoPose	6.31	48.40	8.39	38.80	+2.08	-9.60

ten hand datasets [1, 4, 5, 9, 44, 45, 46, 47, 48], totaling 2.7M samples. As shown in Table 6, both HaMeR [29] and WiLoR [40] improve over baselines on FreiHAND in PA-MPJPE and PA-MPVPE, indicating that our data contributes complementary information rather than redundant samples.

4.4 Benchmarking Object 6D Pose Estimation Methods under Human, Robot Grasping

Object 6D pose estimation is fundamental for robotic manipulation, but standard benchmarks mostly focus on object-centric tabletop or cluttered scenes. HRDexDB enables interaction-centric evaluation by providing calibrated observations, CAD models, and object 6D pose annotations under both human and robot grasping.

We evaluate three RGB-based object 6D localization methods, FoundPose [49], GigaPose [50], and PicoPose [51]. All methods use the same RGB+mask-conditioned protocol, sharing object identity, RGB image, SAM3-generated mask, camera intrinsics, CAD model, and BOP-style metrics. For refinement-based variants, the top five coarse hypotheses are refined using MegaPose [52].

Table 7 shows that all methods perform worse under robot grasping than under paired human grasping. This suggests that robotic hands introduce additional ambiguities for object localization, as rigid links and fingertips can overlap with object boundaries and produce object-like visual structures.

We also test whether HRDexDB can supervise adaptation to robot-object interaction by fine-tuning the MegaPose refiner with 100k GSO synthetic samples and 5.3k HRDexDB robot-grasp annotations. We evaluate on a held-out robot-grasp environment from the OmniRobotHome system [53], separate from HRDexDB. As shown in Table 8, the fine-tuned refiner improves mean ADD-S by 10.2% , suggesting that HRDexDB can help adapt pose refinement to interaction-centric settings.

Table 8: Effect of MegaPose refiner fine-tuning with HRDexDB robot-grasp data.

Refiner	ADD (cm) ↓	ADD-S (cm) ↓
Original	8.23	4.40
Fine-tuned	**7.90**	**3.95**
Rel. improv.	**3.99%**	**10.2%**

5 Conclusion

We present **HRDexDB**, the first dataset of paired human–robot dexterous manipulation across multiple embodiments, with 2.1K high-fidelity sequences over 100 objects featuring dense 3D hand and 6D object annotations and synchronized tactile sensing. By aligning real-world human and robot grasps on shared objects, HRDexDB provides a new resource for studying how dexterous grasp strategies transfer across embodiments. Through contact map transfer, latent-space grasp retrieval, 3D hand pose estimation, and object 6D pose estimation benchmarks, we show that HRDexDB supports both cross-embodiment grasp transfer and interaction-centric perception evaluation. Moving forward, we plan to expand HRDexDB toward 1,000 objects and more complex functional manipulation tasks.

6 Limitations

Despite the scale and multi-modality of **HRDexDB**, two limitations remain. **(1) Tactile Heterogeneity.** Tactile sensing is available only for robotic hands, and sensor specifications vary across

platforms, complicating unified tactile analysis. Future work should explore normalization strategies or shared latent tactile representations. **(2) Defining Trajectory Correspondence.** HRDexDB pairs human and robot grasps at the semantic level, but defining functionally equivalent motions across different hand morphologies remains an open problem for cross-embodiment imitation.

References

[1] G. Moon, S.-I. Yu, H. Wen, T. Shiratori, and K. M. Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *ECCV*, 2020.

[2] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *CVPR*, 2018.

[3] S. Brahmbhatt, C. Ham, C. C. Kemp, and J. Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *CVPR*, 2019.

[4] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, and T. Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *ICCV*, 2019.

[5] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *CVPR*, 2021.

[6] Y. Liu, Y. Liu, C. Jiang, K. Lyu, W. Wan, H. Shen, B. Liang, Z. Fu, H. Wang, and L. Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *CVPR*, 2022.

[7] Z. Fan, O. Taheri, D. Tzionas, M. Kocabas, M. Kaufmann, M. J. Black, and O. Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, 2023.

[8] X. Zhan, L. Yang, Y. Zhao, K. Mao, H. Xu, Z. Lin, K. Li, and C. Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *CVPR*, 2024.

[9] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020.

[10] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

[11] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE TPAMI*, 43(11):4125–4141, 2020.

[12] Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, S. Gao, X. He, X. Hu, X. Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

[13] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *ICRA*, 2024.

[14] S. Wu, X. Liu, S. Xie, P. Wang, X. Li, B. Yang, Z. Li, K. Zhu, H. Wu, Y. Liu, et al. Robocoin: An open-sourced bimanual robotic data collection for integrated manipulation. *arXiv preprint arXiv:2511.17441*, 2025.

[15] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. In *RSSW*, 2024.

[16] K. Wu, C. Hou, J. Liu, Z. Che, X. Ju, Z. Yang, M. Li, Y. Zhao, Z. Xu, G. Yang, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *arXiv preprint arXiv:2412.13877*, 2024.

[17] H.-S. Fang, H. Fang, Z. Tang, J. Liu, C. Wang, J. Wang, H. Zhu, and C. Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *ICRA*, 2024.

[18] T. Tao, M. K. Srirama, J. J. Liu, K. Shaw, and D. Pathak. Dexwild: Dexterous human interactions for in-the-wild robot policies. *RSS*, 2025.

[19] S. Xie, H. Cao, Z. Weng, Z. Xing, H. Chen, S. Shen, J. Leng, Z. Wu, and Y.-G. Jiang. Human2robot: Learning robot actions from paired human-robot videos. In *AAAI*, 2026.

[20] Y. Liu, H. Yang, X. Si, L. Liu, Z. Li, Y. Zhang, Y. Liu, and L. Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *CVPR*, 2024.

[21] J.-T. Song, J. Kim, J. Cao, Y. Lei, T. Yagi, and K. Kitani. Contact4d: A video dataset for whole-body human motion and finger contact in dexterous operations. In *3DV*, 2026.

[22] P. Banerjee, S. Shkodrani, P. Moulon, S. Hampali, S. Han, F. Zhang, L. Zhang, J. Fountain, E. Miller, S. Basol, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *CVPR*, 2025.

[23] R. Fu, D. Zhang, A. Jiang, W. Fu, A. Funk, D. Ritchie, and S. Sridhar. Gigahands: A massive annotated dataset of bimanual hand activities. In *CVPR*, 2025.

[24] Y. Liu, Y. Yang, Y. Wang, X. Wu, J. Wang, Y. Yao, S. Schwertfeger, S. Yang, W. Wang, J. Yu, et al. Realdex: Towards human-like grasping for robotic dexterous hand. *arXiv preprint arXiv:2402.13853*, 2024.

[25] R. Wang, J. Zhang, J. Chen, Y. Xu, P. Li, T. Liu, and H. Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. In *ICRA*, 2023.

[26] P. Li, T. Liu, Y. Li, Y. Geng, Y. Zhu, Y. Yang, and S. Huang. Gendexgrasp: Generalizable dexterous grasping. In *ICRA*, 2023.

[27] H. Zhang, S. Christen, Z. Fan, O. Hilliges, and J. Song. GraspXL: Generating grasping motions for diverse objects at scale. In *ECCV*, 2024.

[28] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM TOG (SIGGRAPH Asia)*, 36(6):245:1–245:17, Nov. 2017.

[29] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik. Reconstructing hands in 3d with transformers. In *CVPR*, 2024.

[30] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

[31] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield. Foundationstereo: Zero-shot stereo matching. In *CVPR*, 2025.

[32] B. Wen, W. Yang, J. Kautz, and S. Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *CVPR*, 2024.

[33] Z. Wu, R. A. Potamias, X. Zhang, Z. Zhang, J. Deng, and S. Luo. Cedex: Cross-embodiment dexterous grasp generation at scale from human-like contact representations. In *ICRA*, 2026.

[34] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017.

[35] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, and G. State. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

[36] J. Chen, Y. Ke, and H. Wang. Bodex: Scalable and efficient robotic dexterous grasp synthesis using bilevel optimization. In *ICRA*, 2025.

[37] B. Sundaralingam, A. Murali, and S. Birchfield. curobov2: Dynamics-aware motion generation with depth-fused distance fields for high-dof robots. *arxiv preprint arXiv:2603.05493*, 2026.

[38] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[39] Y. Qin, W. Yang, B. Huang, K. Van Wyk, H. Su, X. Wang, Y.-W. Chao, and D. Fox. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system. In *RSS*, 2023.

[40] R. A. Potamias, J. Zhang, J. Deng, and S. Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *CVPR*, 2025.

[41] H. Dong, A. Chharia, W. Gou, F. Vicente Carrasco, and F. D. De la Torre. Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba. In *NeurIPS*, 2024.

[42] K. Lin, L. Wang, and Z. Liu. Mesh graphormer. In *ICCV*, 2021.

[43] Y. Rong, T. Shiratori, and H. Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *ICCVW*, 2021.

[44] D. Xiang, H. Joo, and Y. Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, 2019.

[45] S. Jin, L. Xu, J. Xu, C. Wang, W. Liu, C. Qian, W. Ouyang, and P. Luo. Whole-body human pose estimation in the wild. In *ECCV*, 2020.

[46] C. Zimmermann and T. Brox. Learning to estimate 3d hand pose from single rgb images. In *ICCV*, 2017.

[47] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE TPAMI*, 45(6): 7157–7173, 2022.

[48] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.

[49] E. P. Örnek, Y. Labbé, B. Tekin, L. Ma, C. Keskin, C. Forster, and T. Hodaň. Foundpose: Unseen object pose estimation with foundation features. In *ECCV*, 2024.

[50] V. N. Nguyen, T. Groueix, M. Salzmann, and V. Lepetit. Gigapose: Fast and robust novel object pose estimation via one correspondence. In *CVPR*, 2024.

[51] L. Liu, J. Lin, Z. Liu, and K. Jia. Picopose: Progressive pixel-to-pixel correspondence learning for novel object pose estimation. In *CoRL*, 2025.

[52] Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, and J. Sivic. Megapose: 6d pose estimation of novel objects via render & compare. In *CoRL*, 2022.

[53] J. Lee, S. Han, J. Kim, I. Lee, M. Choi, J. Kim, W. Woo, and H. Joo. Omnirobothome: A multi-camera platform for real-time multiadic human-robot interaction. *arXiv preprint arXiv:2604.28197*, 2026.