Human2Dex: Dexterous Robot Learning from Wrist-Centric Human Demonstrations

Anonymous Authors

Fig. 1. **Human2Dex overview.** A wrist-mounted camera records natural bare-hand demonstrations. The shared wrist-centric observation and 21-point hand interface are converted offline into separate executable training sets for the Linker O6 and Wuji hands. Visual adaptation and GPIR provide shared interface semantics, whereas retargeting, action labels, normalization, and diffusion policies remain hand–task specific.

***Abstract*— Collecting dexterous manipulation demonstrations for each robot hand requires repeated task-specific data acquisition. We ask whether one shared corpus of natural human demonstrations can instead serve as a reusable data interface for different dexterous hands, without collecting task demonstrations on each target robot. We present Human2Dex, a wrist-centric human demonstration interface that separates shared human observation and motion semantics from embodiment-specific execution. It aligns observations through local hand-appearance adaptation and a shared 21-point hand representation with consistent structural semantics across human and robot observations. It constructs executable supervision by retargeting fused human hand motion to each target hand, and introduces the Grasp-Pocket Interaction Representation (GPIR) to express tracked objects relative to a hand-centered Functional Grasp Pocket. The corpus is converted into separate training sets for Linker O6 and Wuji, while action labels, normalization, training, and diffusion-policy checkpoints remain hand-specific. Across six real-world tasks, Human2Dex achieves 109/120 (90.8%) successes on Linker O6 and 97/120 (80.8%) on Wuji. Across both hands, Human-Raw achieves 63/240 (26.3%) successes, whereas the complete Human2Dex configuration achieves 206/240 (85.8%). These results support the use of wrist-centric human demonstrations as a reusable interface across calibrated dexterous hands while retaining embodiment-specific policy learning.**

I. Introduction

Dexterous manipulation policies require demonstrations that jointly capture arm motion, high-DoF hand configuration, and contact-rich task progress. Acquiring such demonstrations separately for each target hand is costly and difficult to scale. Many existing systems make data collection more compatible with robot deployment through instrumented motion capture, robot-like interfaces, or wearable hardware [1]–[6]. These interfaces can reduce the correspondence burden during transfer, but they often establish robot compatibility partly through the acquisition setup, control interface, or embodiment-aware mapping. When the target embodiment changes, the corresponding conversion or embodiment-specific setup must still be established for the new hand. This motivates a higher-level question: how can a reusable human demonstration interface convert the same human manipulation episodes into executable training data for different dexterous hands, without collecting task demonstrations on each target robot?

We study this question from a wrist-centric perspective. Head-mounted egocentric video provides scalable human data and broad scene context [7]–[11], whereas a wrist-mounted camera provides local hand–object observations that more closely correspond to the eye-in-hand views used for close-range dexterous control. This similarity in observation arrangement makes wrist-centric acquisition a natural starting point for demonstration reuse. However, matching the observation arrangement does not by itself establish correspondence between human and robot data. The two settings can still differ in hand appearance, articulated structure, kinematics, command spaces, and grasp-related geometry.

Even under wrist-to-eye-in-hand alignment, three linked mismatches remain. First, human and robot hands differ in appearance and articulated structure, so the same visual content does not automatically provide the same structural semantics. Prior work has addressed this issue through appearance adaptation, robot-hand rendering, and structured hand representations [4], [12]–[14]. Second, human motion is not itself an executable command for a target hand: Linker O6 and Wuji have different kinematic structures, degrees of freedom, joint limits, and native action spaces. Human-to-robot retargeting and embodiment-aware motion mapping therefore remain necessary for constructing target-specific supervision [1], [13], [15], [16]. Third, even after observation semantics are aligned and executable supervision is constructed, a policy must infer how the task object is positioned relative to the hand's current grasp configuration. Image-centered object coordinates indicate where an object appears, but do not directly express its relation to the closing hand. Existing work has explored hand–object and functional grasp representations for grasp synthesis, contact modeling, and perception [9], [17]–[20]. Thus, structural correspondence supports observation interpretation, executable supervision supports target-hand policy learning, and an explicit object–hand relation can provide a more direct interaction cue, particularly when grasp alignment matters.

We present Human2Dex, a reusable wrist-centric human demonstration interface that shares human demonstrations and their observation and motion semantics across embodiments while retaining target-hand-specific executable supervision and policy learning. To align observations, Human2Dex combines mask-guided local hand-appearance adaptation with a shared 21-point hand representation, establishing consistent structural semantics for human and robot observations. To construct executable supervision, fused human hand motion is retargeted separately to each calibrated target hand, producing native action labels for Linker O6 and Wuji. To represent interaction geometry, Human2Dex introduces the Grasp-Pocket Interaction Representation (GPIR), which expresses tracked task objects relative to a hand-centered Functional Grasp Pocket defined by the thumb–index grasp center. The same human demonstration corpus is therefore converted into separate training sets for the two target hands, while action labels, normalization statistics, policy training, and diffusion-policy checkpoints remain specific to each hand–task pair. No task demonstrations are collected with the target robot hands.

We evaluate whether this interface supports human-data reuse across embodiments under real-world dexterous manipulation. A matched Linker O6–Croissant study examines wrist- versus head-centric observation, while the main evaluation covers Linker O6 and Wuji across six real-world tasks. Additional reference-frame and task-wise analyses examine selected design choices and the task dependence of interaction grounding.

Our contributions are:

- **A reusable human demonstration interface for multiple dexterous embodiments.** Human2Dex converts one wrist-centric human demonstration corpus into separate hand-specific training sets for Linker O6 and Wuji, allowing each hand to learn an independent policy without collecting task demonstrations on the target robots.
- **A grasp-centered interaction representation for policy conditioning.** The Grasp-Pocket Interaction Representation expresses tracked objects relative to a hand-centered Functional Grasp Pocket, complementing structural observation alignment and target-specific action supervision with an explicit object–hand relation.
- **A two-hand, six-task physical evaluation with targeted design analyses.** We evaluate the complete interface on two calibrated dexterous hands and six real-world tasks, complemented by controlled comparisons of observation viewpoint and object-reference coordinates and of interaction-grounding behavior.

II. Related Work

A. Learning Robot Policies from Human Demonstrations

Human-data interfaces differ in where compatibility between human demonstrations and robot execution is established. Portable robot-like tools, motion-capture systems, teleoperation platforms, and wearable interfaces introduce deployment-relevant visual, kinematic, or contact signals during collection [1]–[6], [14], [21]–[24]. Other approaches preserve more natural human behavior through egocentric video, human play, or human–robot co-training, while some convert human data through simulation or limited robot data [7]–[11], [25]–[29]. These studies establish useful routes for bringing human data into robot learning, but they differ in whether embodiment compatibility is built during acquisition or constructed afterward. Human2Dex focuses on the latter setting by reusing one wrist-centric human demonstration corpus and converting it into separate training sets for multiple dexterous hands.

B. Visual and Kinematic Cross-Embodiment Alignment

Cross-embodiment alignment involves both visual identity and articulated motion. Visual adaptation methods reduce differences between source and target hands through cross-painting, robot-hand rendering, or fine-grained hand–object segmentation [4], [12], [30]. Hand estimators and teleoperation systems provide structured landmarks or mappings for transferring human motion under different morphologies and kinematic constraints [13]–[16], [31]–[33]. Human2Dex uses a 21-point hand representation as an intermediate interface for shared observation semantics and motion fusion, while retaining embodiment-specific retargeting, action labels, and policy checkpoints for each target hand. Thus, the shared representation defines data semantics rather than a common robot action space.

C. Object-Centric and Functional Grasp Representations

Object-centric representations provide relational information beyond isolated hand or object states. FunctionalGrasp and D(R,O) Grasp use semantic hand–object structure or

robot–object descriptions for grasp generation, while ContactPose and HOT3D support contact modeling and egocentric hand–object tracking [17]–[20]. Human2Dex uses the Grasp-Pocket Interaction Representation (GPIR) for a narrower purpose: sequential wrist-view policy control. GPIR expresses tracked task objects relative to a hand-centered Functional Grasp Pocket and retains object scale and tracking reliability. It is therefore a low-dimensional grasp-relative interaction cue, rather than a general object reconstruction, universal grasp representation, or complete contact model.

III. Problem Formulation

We index target hands by r and tasks by τ. The experimental configuration c (Raw, Visual, Structured, or Human2Dex) is suppressed below for compactness; every hand–task–configuration tuple is trained independently. For each task, we consider the human demonstration subset

$$\mathcal{D}_{h,\tau} = \{I_t,\, T_t^w,\, P_t^p,\, P_t^v,\, U_t^v\}_{t=1}^{T_\tau}, \tag{1}$$

where $I_t \in \mathbb{R}^{H\times W\times 3}$ is a fisheye wrist image, $T_t^w \in SE(3)$ is the tracked human wrist pose, and $P_t^p, P_t^v \in \mathbb{R}^{21\times 3}$ are hand structures estimated by an external tracker and a wrist-RGB model, respectively, and $U_t^v \in \mathbb{R}^{21\times 2}$ contains direct image-space landmarks used for rendering and pocket construction. No task demonstration is collected with the target robot hand.

For each calibrated target hand r, a kinematic retargeter ρ_r converts the fused human structure into an executable hand state q_t^r. For task τ, the shared human source then produces a hand–task-specific training set (instantiated independently for each c)

$$\mathcal{D}_{r,\tau} = \Psi_{r,\tau}(\mathcal{D}_{h,\tau}) = \{\bar{I}_t, z_t, q_t^r, a_t^{r,\tau}\}_{t=1}^{T_\tau}, \tag{2}$$

where $\bar{I}_t$ is the structural observation, z_t is the optional functional interaction state, and $a_t^{r,\tau}$ contains arm and target-hand commands. We therefore train one task-specific policy for each hand–task pair within each configuration,

$$\pi_{\theta_{r,\tau}} : o_{t-h+1:t}^{r,\tau} \mapsto a_{t:t+H_a}^{r,\tau}. \tag{3}$$

The observation $o_t^{r,\tau}$ contains $\bar{I}_t$, the measured or retargeted hand state q_t^r, and z_t when functional grounding is enabled.

The objective is to reuse a single wrist-centric human-data interface across two calibrated dexterous hands rather than to share a policy checkpoint. Human RGB and motion are shared, whereas action labels, normalization statistics, and policy parameters $\theta_{r,\tau}$ remain hand–task specific within each configuration. Each policy must resolve three mismatches: hand appearance and structure, target action geometry, and the functional relation between the object and the closing hand.

IV. Method

A. Shared Human Source and Conversion Interface

Human2Dex converts one shared wrist-centric human demonstration source into hand–task-specific training data for multiple dexterous embodiments. The overall interface is illustrated in Fig. 1. For task τ, the human demonstration corpus is

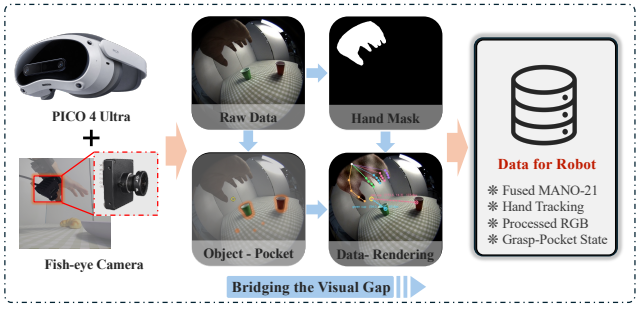

Fig. 2. Human2Dex conversion interface. Wrist RGB, PICO tracking, and a wrist-image hand model provide measurements for structural observations, target-specific action supervision, and grasp-centered interaction grounding. The resulting data are converted offline for each target hand.

$$\mathcal{D}_{h,\tau} = \{I_t,\, T_t^w,\, P_t^p,\, P_t^v,\, U_t^v\}_{t=1}^{T_\tau}, \tag{4}$$

where $I_t \in \mathbb{R}^{H\times W\times 3}$ is the wrist-camera RGB image, $T_t^w \in SE(3)$ is the human wrist pose, $P_t^p, P_t^v \in \mathbb{R}^{21\times 3}$ are 3-D hand structures from PICO tracking and a wrist-RGB hand model, and $U_t^v \in \mathbb{R}^{21\times 2}$ contains direct image-space landmarks. The wrist-image model uses a shared DINO encoder with a 3-D MANO branch and a direct 2-D keypoint branch. The 2-D landmarks support image-space rendering and grasp-reference construction, whereas the 3-D structures support motion fusion and target-hand retargeting.

For target hand r and task τ, Human2Dex constructs

$$\boxed{\mathcal{D}_{r,\tau} = \Psi_{r,\tau}\left(\mathcal{D}_{h,\tau}\right) = \left\{\bar{I}_t,\, q_t^r,\, z_t,\, a_t^r\right\}_{t=1}^{T_\tau}} \tag{5}$$

where $\bar{I}_t$ is the converted visual observation, q_t^r is the target-hand state, z_t is the interaction state, and a_t^r is the executable arm–hand action. The conversion contains three complementary branches: observation alignment, target-specific action supervision, and grasp-centered interaction grounding. The same human source is reused for Linker O6 and Wuji, while retargeting, action labels, normalization, training, and policy checkpoints are specific to each hand–task pair. No task demonstrations are collected with the target robot hands.

B. Hand-Structure-Aligned Observations

The observation branch adapts the human-hand appearance and adds a shared structural description. Given a hand mask M_t^h, local appearance randomization is defined as

$$I_t^{\text{aug}} = (1 - M_t^h) \odot I_t + M_t^h \odot \mathcal{A}(I_t), \tag{6}$$

where $\mathcal{A}$ is the appearance-randomization operator. The object and background remain unchanged.

Let Γ_t^h map the human image to the 224×224 policy image, and let $\mathcal{R}$ render the 21-point hand structure:

$$\bar{I}_t^h = \mathcal{R}\left(\Gamma_t^h(I_t^{\text{aug}}), \Gamma_t^h(U_t^v)\right). \tag{7}$$

A fixed topology and color assignment preserve landmark and finger identities across observations.

At deployment, robot landmarks are generated from forward kinematics and camera projection:

$$\bar{U}^r_{t,j} = \Gamma^r_t \left[\Pi \left(K_r, T^r_{C \leftarrow B}, F^r_j(q^r_t) \right) \right], \tag{8}$$

where F^r_j returns landmark j in the robot base frame, and $K_r, T^r_{C \leftarrow B}$ define the calibrated camera model. Human and robot observations use the same landmark topology, colors, and image-space coordinate convention.

RGB images, hand masks, object masks, and landmarks undergo the same crop, resize, and image-space perturbation. The observation branch outputs the registered structural observation $\bar{I}_t$.

C. Target-Specific Action Supervision

The action branch converts the shared human motion into executable commands for each target hand. The PICO and wrist-RGB 3-D structures are aligned in a wrist-centered MANO frame and fused using confidence and temporal consistency:

$$P_t = \mathrm{Fuse}\left(P^p_t, P^v_t\right). \tag{9}$$

Temporal filtering handles short tracking dropouts and preserves reference bone lengths.

For target hand r, the fused structure is retargeted as

$$q^r_t = \rho_r \left(P_t; \mathcal{K}_r\right), \tag{10}$$

where ρ_r is the embodiment-specific retargeter and $\mathcal{K}_r$ is the calibrated hand model. The human wrist trajectory is mapped through the calibrated arm transform to obtain relative end-effector translation. Native hand commands are constructed by

$$\begin{aligned} h^r_{t,k} &= \phi_r \left(q^r_t, q^r_{t+k}\right), \\ a^r_{t+k} &= \left[\Delta x^r_{t,k},\, r^{r,6D}_{t+k},\, h^r_{t,k}\right]^\top \in \mathbb{R}^{9+d_r}, \end{aligned} \tag{11}$$

where $\Delta x^r_{t,k}$ is relative end-effector translation, $r^{r,6D}_{t+k}$ is the rotation representation, and ϕ_r is the fixed command mapping used during data construction and deployment. The hand-command dimensions are $d_{\mathrm{O6}} = 6$ and $d_{\mathrm{Wuji}} = 20$. The 21-point structure serves as an intermediate human-motion interface, and native commands are generated separately for each target hand.

During training, q^r_t is obtained from retargeted human motion. During deployment, the corresponding state is measured from the target robot hand.

D. Grasp-Centered Interaction Representation

The interaction branch provides a compact description of the object relative to the current grasp location. For thumb–index grasps, fingertip tracks are smoothed with a causal three-frame median filter:

$$\begin{aligned} \widetilde{U}^v_{t,j} &= \mathrm{med}\left\{U^v_{t-2,j}, U^v_{t-1,j}, U^v_{t,j}\right\}, \\ p^h_{t,j} &= \Gamma^h_t \left(\widetilde{U}^v_{t,j}\right), \qquad p^r_{t,j} = \bar{U}^r_{t,j}. \end{aligned} \tag{12}$$

For either domain $d \in \{h, r\}$, the Functional Grasp Pocket is the image-space midpoint of the thumb and index fingertips:

$$g^d_t = \frac{p^d_{t,\mathrm{thumb}} + p^d_{t,\mathrm{index}}}{2}. \tag{13}$$

Robot fingertips are projected individually before their midpoint is computed.

For object i, SAM-based tracking [34] provides the policy-image centroid (u^i_t, v^i_t) and mask area A^i_t. Tracking confidence and memory validity are retained for intermittent updates. With $g_t = (u^g_t, v^g_t)$, the Grasp-Pocket Interaction Representation is

$$z^i_t = \left[\frac{u^i_t - u^g_t}{s_{\mathrm{ref}}}, \frac{v^i_t - v^g_t}{s_{\mathrm{ref}}}, \log \frac{A^i_t}{A^i_{\mathrm{ref}}}, c^i_t, m^i_t\right]^\top, \tag{14}$$

where s_{ref} is the image-coordinate normalization scale, A^i_{ref} is the reference object area, and c^i_t and m^i_t denote tracking confidence and memory validity. Multiple object states are concatenated. GPIR provides a compact whole-object, image-plane representation of grasp-relative alignment, scale, and tracking reliability for policy conditioning.

E. Policy Learning

The converted observation, target-hand state, and interaction state form the policy input:

$$o^{r,\tau}_t = \left\{\bar{I}_t,\, q^r_t,\, z_t\right\}, \tag{15}$$

where q^r_t is always included and z_t is included when interaction grounding is enabled. Each hand–task pair is trained independently.

The policy architecture is shown in Fig. 3. A DINOv2 ViT-B/14 Reg4 encoder [35] processes $\bar{I}_t$. Its features are combined with q^r_t and, when enabled, z_t to condition a 1-D U-Net. The network predicts a 16-step action sequence under the Diffusion Policy objective [36]:

$$\pi_{\theta_{r,\tau}} : o^{r,\tau}_{t-h+1:t} \mapsto a^r_{t:t+H_a}. \tag{16}$$

For configurations with auxiliary supervision, we regress the image-centered object position and log-area:

$$y^{\mathrm{aux}}_t = \left[\frac{u^i_t - W/2}{s_{\mathrm{ref}}}, \frac{v^i_t - H/2}{s_{\mathrm{ref}}}, \log \frac{A^i_t}{A^i_{\mathrm{ref}}}\right]_{i=1}^{N}. \tag{17}$$

The training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{DP}} + \lambda_{\mathrm{aux}} \mathcal{L}_{\mathrm{aux}}, \tag{18}$$

with $\lambda_{\mathrm{aux}} = 0$ when auxiliary supervision is disabled. The auxiliary head is omitted during deployment.

We train each policy with AdamW using a learning rate of 3×10^{-4}, a batch size of 32, one-frame observation history, and U-Net channel widths of 256, 512, and 1024. The diffusion schedule uses 50 training timesteps and 16 DDIM inference steps. Five percent of episodes are reserved for validation, with all appearance variants of an episode assigned to the same split.

V. Experiments

A. Research Questions

The experiments address three questions. **Q1:** Can one human demonstration corpus support independent policy learning for two dexterous hands? **Q2:** How do observation

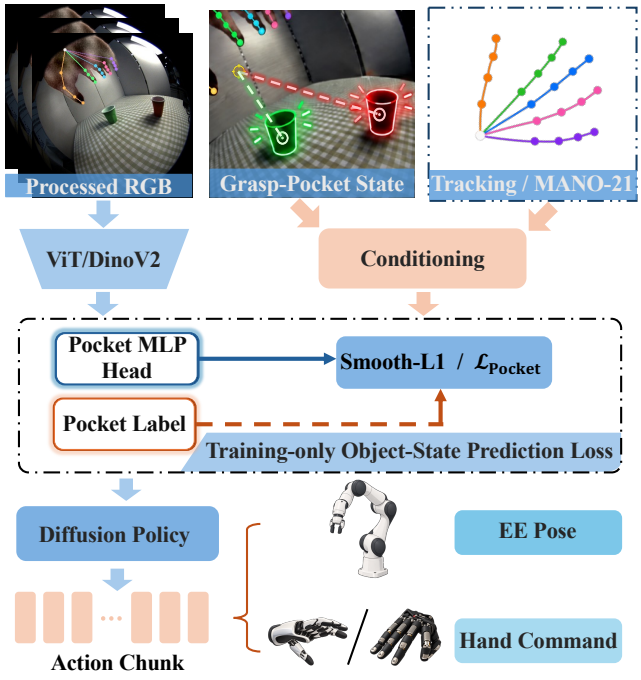

Fig. 3. Target-specific policy learning and inference. Converted human observations condition a diffusion policy that predicts native commands for one hand–task pair. The human-data interface is shared across embodiments, whereas action construction, normalization, policy training, and checkpoints remain hand-specific.

alignment and fused human-hand supervision affect performance, and how does wrist-centric observation compare with head-centric observation? **Q3:** When is grasp-centered interaction grounding useful, and how do reference coordinates and task geometry affect its performance?

B. Experimental Setup

Target Robot Hands. We evaluate Human2Dex across two anthropomorphic dexterous hands with different kinematic structures. Both hands are equipped with calibrated wrist-mounted fisheye cameras:

- **Linker O6:** eleven-DoF anthropomorphic dexterous hand with six active DoFs. The thumb has two active and one passive DoFs, while each remaining finger has one active and one passive DoF.
- **Wuji:** A fully-actuated hand with twenty active DoFs. Each finger has four active DoFs.

Tasks. The six real-world tasks cover compliant grasping, precise placement, object reorientation, coordinated finger contact, and articulated-object interaction:

- **Sponge [O6 & Wuji]:** Pick up a compliant sponge and place it on a target plate.
- **Croissant [O6 & Wuji]:** Pick up a croissant and place it on a target plate while maintaining a stable grasp.
- **Cup [O6 & Wuji]:** Place a green cup into a red cup, requiring accurate grasp alignment and insertion.
- **Bottle [O6 & Wuji]:** Lift a horizontal bottle and place it upright.
- **Egg Carton [O6 & Wuji]:** Open an egg carton by stabilizing the carton and lifting its lid with coordinated fingers.
- **Open Drawer [O6 & Wuji]:** Engage a drawer handle and pull the drawer open.

Shared Human Corpus. The six tasks contain 1,899 quality-controlled base episodes and 556,624 frames. Each base episode produces three appearance-rendered variants. All configurations use the same base episodes, episode-level split, sampled variants, and optimizer budget. O6 and Wuji share episode groups but use distinct actions, normalization, and checkpoints, with all variants of an episode assigned to the same split.

System Configurations. All configurations use the same embodiment-specific retargeter. Human-Raw and Human-Visual use PICO-derived hand structures, whereas Human-Structured and Human2Dex use the fused PICO/RGB structure. Human-Visual adds local visual alignment, and Human-Structured adds fused hand supervision. Human-Structured has no explicit object state or auxiliary object-regression head. Human2Dex additionally uses task-specific object tracking, Pocket-relative object input, and the auxiliary objective. Human-Absolute uses the same object processing and auxiliary supervision as Human2Dex but references object position to the image center.

Evaluation Protocol. Each method–task pair receives 20 physical trials with fixed reset ranges; a success must satisfy its terminal criterion for 2 s. Configurations share the base demonstrations, grouped split, backbone, optimization budget, and physical protocol. Wilson 95% intervals describe trial uncertainty for these fixed configurations. Cross-task pooled counts are benchmark summaries, not estimates for unseen-task populations.

C. Main System-Level Results

Table I addresses Q1. Both hands use the same human demonstration sources, while hand supervision and policies are independently constructed for each embodiment. The four rows are separately trained system configurations.

Human-Raw obtains $19/120$ on Wuji (15.8%) and $44/120$ on O6 (36.7%). Human2Dex reaches $97/120$ on Wuji (80.8%, Wilson 95% CI 72.9–86.9%) and $109/120$ on O6 (90.8%, 84.3–94.8%). Across both hands, the pooled success rates for Human-Raw, Human-Visual, Human-Structured, and Human2Dex are 26.3%, 60.0%, 67.9%, and 85.8%, respectively.

These results support the reuse of one human demonstration corpus for independent policy learning on two calibrated dexterous hands. The successive configuration comparisons provide system-level evidence for the integrated conversion framework; they do not isolate the causal effect of any individual component.

D. Observation and Human-Motion Supervision Analyses

The Raw-to-Visual and Visual-to-Structured comparisons address the roles of observation alignment and human-motion supervision. All configurations use the same

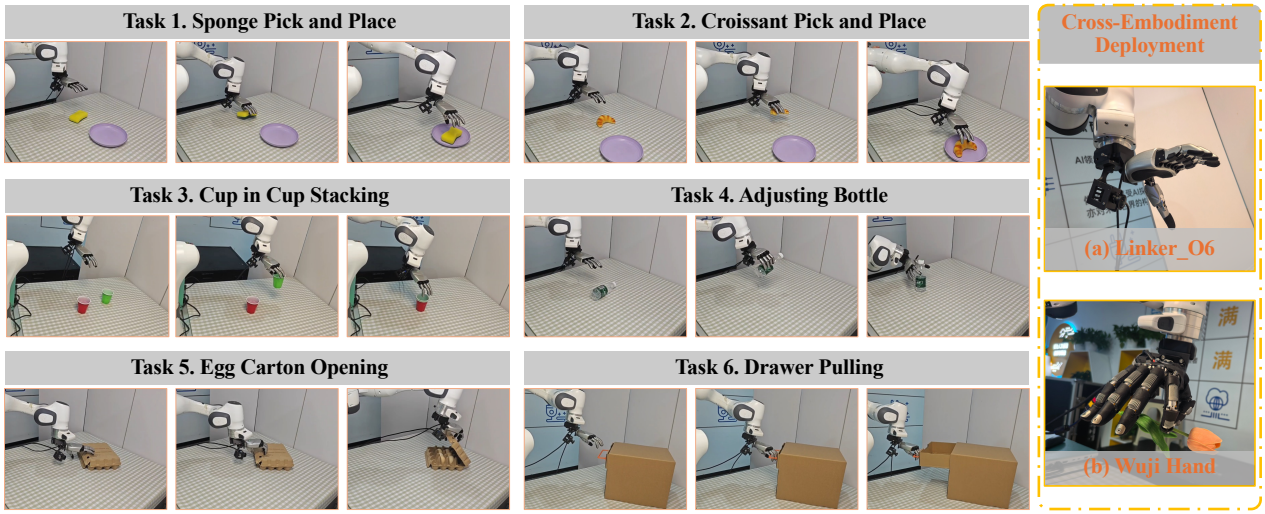

Fig. 4. The six real-world manipulation tasks evaluated on the Linker O6 and Wuji dexterous hands.

TABLE I

REAL-WORLD TASK SUCCESS (SUCCESSFUL TRIALS / 20). EACH HAND–TASK PAIR USES INDEPENDENTLY RETARGETED SUPERVISION AND A SEPARATE POLICY; THE FINAL ROW USES TASK-SPECIFIC OBJECT–POCKET OBSERVATIONS ON ALL SIX TASKS.

Method	Wuji						Linker O6					
	Sponge	Croissant	Cup	Bottle	Egg Carton	Open Drawer	Sponge	Croissant	Cup	Bottle	Egg Carton	Open Drawer
Human-Raw	5/20	3/20	2/20	0/20	2/20	7/20	10/20	6/20	5/20	6/20	12/20	5/20
Human-Visual	12/20	8/20	7/20	12/20	13/20	13/20	14/20	8/20	11/20	16/20	16/20	14/20
Human-Structured	12/20	10/20	7/20	15/20	**14/20**	**16/20**	15/20	9/20	11/20	17/20	**20/20**	**17/20**
Human2Dex	**19/20**	**17/20**	**17/20**	**16/20**	13/20	15/20	**20/20**	**18/20**	**16/20**	**19/20**	19/20	**17/20**

embodiment-specific retargeter; the Structured transition instead replaces PICO-only hand structures with fused PICO/RGB supervision. Visual alignment raises pooled success from 26.3% to 60.0%, while the transition to fused hand supervision reaches 67.9%. These are system-level configuration comparisons, with embodiment-specific retargeting fixed across conditions.

We evaluate a matched O6–Croissant comparison using synchronized head- and wrist-camera observations while fixing action supervision, architecture, optimization, and physical evaluation. First-Attempt Grasp denotes a stable first closure without reopening or regrasping. It is used here as a secondary viewpoint diagnostic, not as an independent measure of GPIR.

Wrist-view demonstrations achieve 6/20 final successes and 4/20 first-attempt grasps, whereas head-view demonstrations achieve 0/20 for both measures. This single-hand, single-task comparison supports wrist-centric observation as a useful design choice for the evaluated eye-in-hand setting.observations.

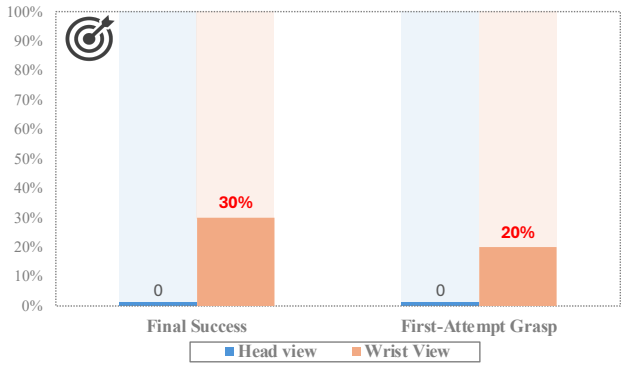

Fig. 5. Matched head- versus wrist-view comparison on O6–Croissant. 20 trials per condition, wrist-view demonstrations achieve 30% final success and 20% first-attempt grasp success, compared with 0% for both metrics using head-view demonstrations.

E. Interaction-Grounding Analysis

a) Integrated grounding.: The Structured-to-Human2Dex comparison evaluates the complete interaction-grounding configuration. This transition adds object tracking, Pocket-relative object input, and the auxiliary prediction objective together.

b) Reference-frame comparison.: We compare Human-Absolute and Human2Dex on Linker O6 for Croissant and Cup. The two configurations use the same base episodes, object processing, architecture, optimization, auxiliary target, and physical evaluation protocol. They differ only in whether object position is referenced to the image center or the

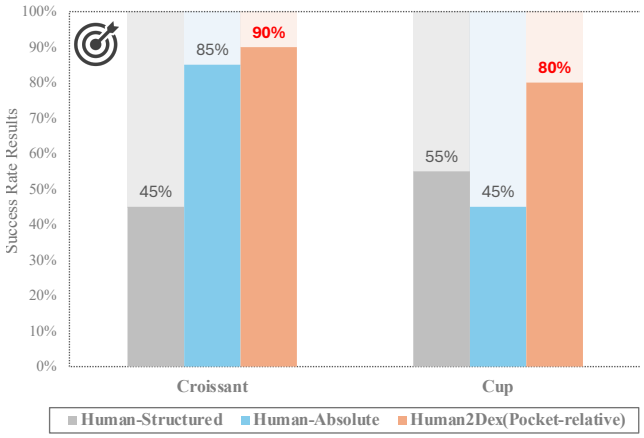

Fig. 6. Functional-reference comparison on Linker O6. Bars show task success rate over 20 physical trials. Human-Absolute and Human2Dex differ only in whether object position is referenced to the image center or the Functional Grasp Pocket.

Functional Grasp Pocket.

Human-Absolute reaches $17/20$ on Croissant, compared with $18/20$ for Human2Dex. On Cup, Human-Absolute reaches $9/20$, compared with $16/20$ for Human2Dex. Image-centered conditioning is competitive for Croissant, whereas Cup shows a larger improvement with the Pocket-relative reference in this evaluation.

c) Task-wise behavior.: Across the six tasks, the complete interface improves performance most on Sponge, Croissant, Cup, and Bottle, while the gains are smaller or absent on Egg Carton and Open Drawer. The former tasks primarily involve whole-object grasping, whereas the latter depend more on localized functional regions such as a lid edge or drawer handle. This pattern suggests that the current whole-object grounding is better suited to grasp-centric tasks.

d) Task-wise results.: Across the two hands, Human-Raw obtains $15/40$, $9/40$, $7/40$, $6/40$, $14/40$, and $12/40$ successes on Sponge, Croissant, Cup, Bottle, Egg Carton, and Open Drawer, respectively, whereas Human2Dex reaches $39/40$, $35/40$, $33/40$, $35/40$, $32/40$, and $32/40$.

Relative to Human-Structured, Human2Dex improves the combined success counts from $27/40$ to $39/40$ on Sponge, $19/40$ to $35/40$ on Croissant, $18/40$ to $33/40$ on Cup, and $32/40$ to $35/40$ on Bottle, while slightly decreasing performance on Egg Carton ($34/40$ to $32/40$) and Open Drawer ($33/40$ to $32/40$).

This task dependence is consistent with the scope of the current object-level grounding. Sponge, Croissant, Cup, and Bottle primarily require the hand to align with and grasp the object as a whole, for which an object-to-pocket relation provides a meaningful geometric cue. In contrast, Egg Carton and Open Drawer require interaction with localized functional regions—the lid edge and drawer handle—whose locations are not necessarily aligned with the centroid of the tracked object. These results suggest that the current GPIR is most effective for grasp-centric tasks, while tasks dominated by localized contact may benefit from part-level functional grounding.

VI. Discussion and Limitations

The main result is that one human demonstration source can support independent policy learning on two calibrated dexterous hands. This reuse is achieved by separating shared source semantics from target-hand execution: human observations and motion representations are shared, while retargeting, executable action labels, normalization, and policy training are constructed for each hand–task pair. The resulting interface avoids recollecting task demonstrations for each target hand while preserving the control structure required by each embodiment.

The reference-frame comparison shows that object localization and interaction referencing are distinct design choices. In the controlled comparison, image-centered conditioning is competitive on Croissant, whereas Pocket-relative conditioning performs substantially better on Cup, where grasp alignment and insertion are more precise. These results suggest that the value of a grasp-centered reference depends on how directly the object–hand relation determines task success.

The task-wise results further clarify the role of interaction grounding. The Structured-to-Human2Dex comparison adds object tracking, Pocket-relative object input, and the auxiliary prediction objective together, and therefore evaluates the integrated grounding configuration. Improvements on Sponge, Croissant, Cup, and Bottle are consistent with tasks that require alignment with and grasping of an object as a whole. Egg Carton and Open Drawer depend more on localized regions such as a lid edge or drawer handle, where a whole-object representation provides a less direct cue. These results motivate part-level or affordance-level interaction representations for localized contact tasks.

The present study covers two calibrated anthropomorphic hands and six real-world tasks, with a separate policy for every hand–task pair. The evaluation therefore establishes reuse across the tested embodiments, while broader transfer to arbitrary morphologies and shared action spaces remains open. GPIR currently uses a specified task object and reliable image-space tracking, and its whole-object representation does not explicitly identify localized contact regions. The viewpoint comparison is limited to one hand and one task. In addition, the absence of a robot-demonstration baseline leaves direct sample-efficiency comparisons between human and robot data for future work.

VII. Conclusion

Human2Dex establishes a reusable wrist-centric human demonstration interface for dexterous robot learning across multiple embodiments. It shares human demonstrations and their observation and motion semantics, then converts them into embodiment-specific visual observations and executable action supervision for each target hand, with GPIR providing a grasp-centered object–hand interaction cue. Using one human demonstration corpus, Human2Dex supports independent policy learning for Linker O6 and Wuji and achieves

109/120 and 97/120 successes across six real-world tasks, respectively. The reference-frame and task-wise analyses further suggest that whole-object grasp-relative grounding is most beneficial when task success depends strongly on grasp alignment, whereas localized interactions may require finer-grained representations.

REFERENCES

[1] C. Wang, H. Shi, W. Wang, *et al.*, "DexCap: Scalable and portable mocap data collection system for dexterous manipulation," *arXiv preprint arXiv:2403.07788*, 2024.

[2] C. Chi, Z. Xu, C. Pan, *et al.*, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," *arXiv preprint arXiv:2402.10329*, 2024.

[3] M. Seo, H. A. Park, S. Yuan, Y. Zhu, and L. Sentis, "LEGATO: Cross-embodiment imitation using a grasping tool," *IEEE Robotics and Automation Letters*, 2025.

[4] M. Xu, H. Zhang, Y. Hou, *et al.*, "DexUMI: Using human hand as the universal manipulation interface for dexterous manipulation," *arXiv preprint arXiv:2505.21864*, 2025.

[5] C. Xu, Y. Jiang, J. Huan, *et al.*, "RealDexUMI: A wearable universal manipulation interface for dexterous robot learning," *arXiv preprint arXiv:2606.06033*, 2026.

[6] H.-S. Fang, B. Romero, Y. Xie, *et al.*, "DEXOP: A wearable dexterous hand exoskeleton for general robot teleoperation," *arXiv preprint arXiv:2509.04441*, 2025.

[7] S. Kareer, D. Patel, R. Punamiya, *et al.*, "EgoMimic: Scaling imitation learning via egocentric video," *arXiv preprint arXiv:2410.24221*, 2024.

[8] R. Hoque, P. Huang, D. J. Yoon, M. Sivapurapu, and J. Zhang, "EgoDex: Learning dexterous manipulation from large-scale egocentric video," *arXiv preprint arXiv:2505.11709*, 2025.

[9] Z. Wang, B. He, K. Yu, *et al.*, "HumanEgo: Zero-shot robot learning from minutes of human egocentric videos," *arXiv preprint arXiv:2605.24934*, 2026.

[10] Z. Chen, S. Chen, E. Arlaud, I. Laptev, and C. Schmid, "ViViDex: Learning vision-based dexterous manipulation from human videos," in *IEEE International Conference on Robotics and Automation*, 2025.

[11] G. Zhang, Q. Xu, H. Zhang, *et al.*, "UniDex: A robot foundation suite for universal dexterous hand control from egocentric human videos," *arXiv preprint arXiv:2603.22264*, 2026.

[12] L. Y. Chen, K. Hari, K. Dharmarajan, *et al.*, "Mirage: Cross-embodiment zero-shot policy transfer with cross-painting," *arXiv preprint arXiv:2402.19249*, 2024.

[13] A. Handa, K. V. Wyk, W. Yang, *et al.*, "DexPilot: Vision based teleoperation of dexterous robotic hand-arm system," *arXiv preprint arXiv:1910.03135*, 2019.

[14] Y. Qin, W. Yang, B. Huang, *et al.*, "AnyTeleop: A general vision-based dexterous robot arm-hand teleoperation system," in *Robotics: Science and Systems*, 2023.

[15] S. Zhao, X. Zhu, Y. Chen, C. Li, X. Zhang, M. Ding, and M. Tomizuka, "DexH2R: Task-oriented dexterous manipulation from human to robots," *arXiv preprint arXiv:2411.04428*, 2024.

[16] S. Park, S. Lee, M. Choi, *et al.*, "Learning to transfer human hand skills for robot manipulations," *arXiv preprint arXiv:2501.04169*, 2025.

[17] Y. Zhang, J. Hang, T. Zhu, X. Lin, R. Wu, W. Peng, D. Tian, Y. Sun, *et al.*, "FunctionalGrasp: Learning functional grasp for robots via semantic hand–object representation," *IEEE Robotics and Automation Letters*, vol. 8, no. 5, pp. 3094–3101, 2023.

[18] Z. Wei, Z. Xu, J. Guo, Y. Hou, C. Gao, Z. Cai, J. Luo, and L. Shao, "$\mathcal{D}(\mathcal{R}, \mathcal{O})$ grasp: A unified representation of robot and object interaction for cross-embodiment dexterous grasping," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 4982–4988.

[19] S. Brahmbhatt, C. Tang, C. D. Twigg, C. C. Kemp, and J. Hays, "ContactPose: A dataset of grasps with object contact and hand pose," in *European Conference on Computer Vision*, 2020.

[20] P. Banerjee, S. Shkodrani, P. Moulon, *et al.*, "HOT3D: Hand and object tracking in 3d from egocentric multi-view videos," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[21] C. Wang, L. Fan, J. Sun, *et al.*, "MimicPlay: Long-horizon imitation learning by watching human play," *arXiv preprint arXiv:2302.12422*, 2023.

[22] S. P. Arunachalam, I. Guzey, S. Chintala, and L. Pinto, "Holo-Dex: Teaching dexterity with immersive mixed reality," *arXiv preprint arXiv:2210.06463*, 2022.

[23] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang, "Open-TeleVision: Teleoperation with immersive active visual feedback," *arXiv preprint arXiv:2407.01512*, 2024.

[24] X. Chen, Y. Pan, M. Li, and X. Ding, "DexViTac: Collecting human visuo-tactile-kinematic demonstrations for contact-rich dexterous manipulation," *arXiv preprint arXiv:2603.17851*, 2026.

[25] C. Yuan, R. Zhou, M. Liu, *et al.*, "MotionTrans: Human VR data enable motion-level learning for robotic manipulation policies," *arXiv preprint arXiv:2509.17759*, 2025.

[26] T. Tao, M. K. Srirama, J. J. Liu, K. Shaw, and D. Pathak, "DexWild: Dexterous human interactions for in-the-wild robot learning," in *Robotics: Science and Systems*, 2025.

[27] A. Sivakumar, K. Shaw, and D. Pathak, "Robotic telekinesis: Learning a dexterous policy from RGB videos," *arXiv preprint arXiv:2202.10448*, 2022.

[28] Z. Q. Chen, K. V. Wyk, Y.-W. Chao, *et al.*, "DexTransfer: Real world transfer for dexterous manipulation with reinforcement learning," *arXiv preprint arXiv:2209.14284*, 2022.

[29] Z. Jiang, Y. Xie, K. Lin, *et al.*, "DexMimicGen: Automated data generation for bimanual dexterous manipulation via imitation learning," in *IEEE International Conference on Robotics and Automation*, 2025.

[30] L. Zhang, S. Zhou, S. Stent, and J. Shi, "Fine-grained egocentric hand-object segmentation: A benchmark for automated meal preparation," in *European Conference on Computer Vision*, 2022.

[31] F. Zhang, V. Bazarevsky, A. Vakunov, *et al.*, "MediaPipe Hands: On-device real-time hand tracking," *arXiv preprint arXiv:2006.10214*, 2020.

[32] G. Pavlakos, D. Shan, I. Radosavovic, *et al.*, "Reconstructing hands in 3d with transformers," *arXiv preprint arXiv:2312.05251*, 2023.

[33] H. Yuan, B. Zhou, Y. Fu, and Z. Lu, "Cross-embodiment dexterous grasping with reinforcement learning," *arXiv preprint arXiv:2410.02479*, 2024.

[34] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, *et al.*, "SAM 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[35] M. Oquab, T. Darcet, T. Moutakanni, *et al.*, "DINOv2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[36] C. Chi, Z. Xu, S. Feng, *et al.*, "Diffusion policy: Visuomotor policy learning via action diffusion," *arXiv preprint arXiv:2303.04137*, 2023.