DexRoam: Learning Mobile Bimanual Dexterous Manipulation from Egocentric Whole-Body Human Demonstrations

Rui Zhou[1,2,*], **Yibo Yuan**[4,2,*], **Junkai Zhao**[2,*,†], **Fangyuan Zhao**[3],
Xiaoguang Zhao[5], **Shanghang Zhang**[3,✉], **Sirui Han**[1,✉]
[1]The Hong Kong University of Science and Technology [2]Beijing Academy of Artificial Intelligence
[3]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University
[4]Beihang University [5]Institute of Automation, Chinese Academy of Sciences
[*]Equal contribution †Project Leader ✉Corresponding author

Project Page: https://dexroam.github.io/

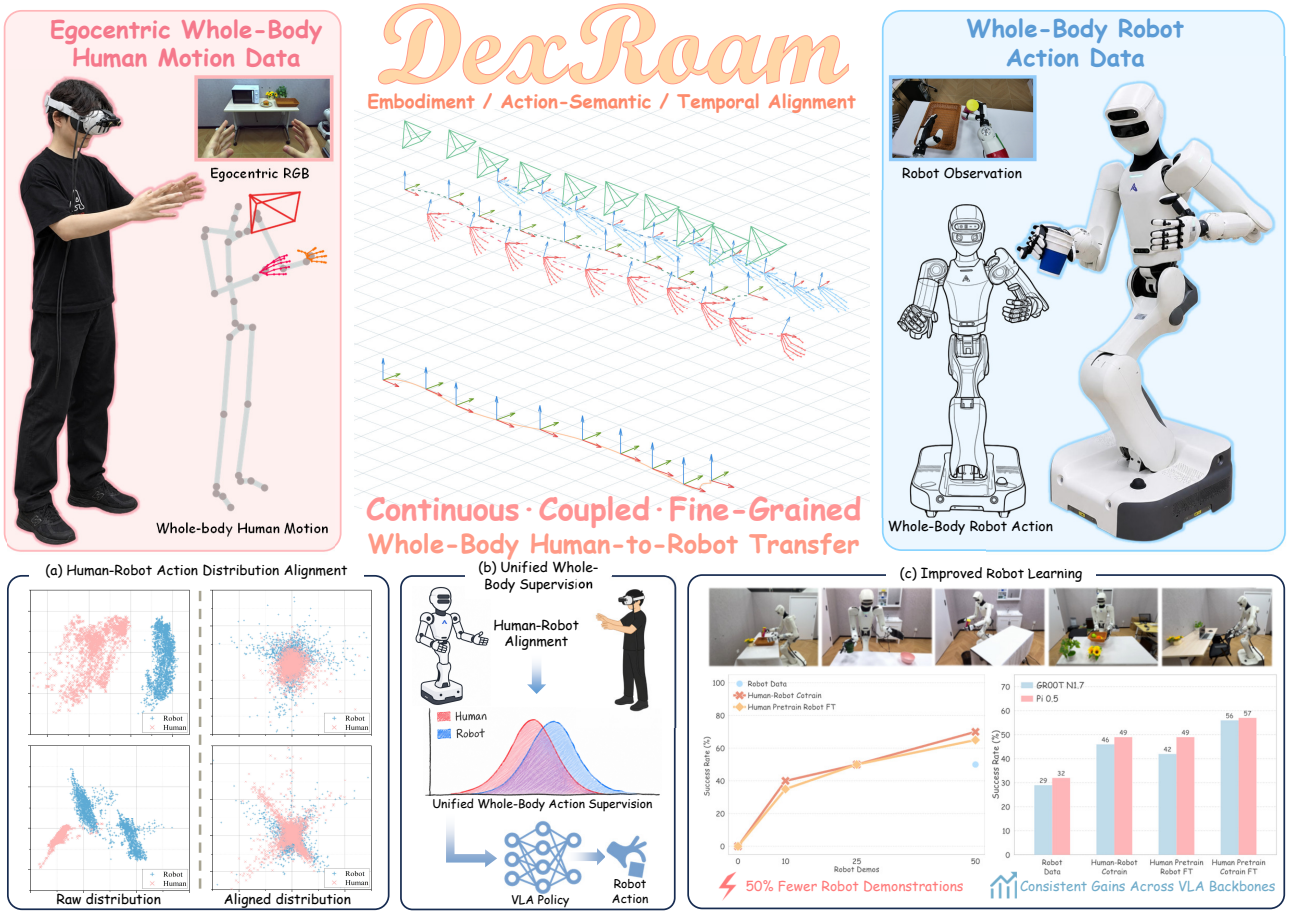

Figure 1. Overview of the DexRoam framework. DexRoam unlocks motion-level human data for mobile bimanual dexterous manipulation by transforming egocentric whole-body demonstrations into a continuous, coupled, robot-compatible action space. DexRoam provides a complete pipeline for tracker-free human motion capture, human-to-robot alignment, and VLA-based policy learning, enabling scalable learning from human demonstrations.

***Abstract*—Mobile bimanual dexterous manipulation requires continuous coordination of locomotion, whole-body motion, and finger-level dexterity within a single trajectory, creating a severe robot demonstration bottleneck. Egocentric human demonstrations offer a scalable alternative, but prior approaches ease the transfer by simplifying human motion, discarding exactly the fine-grained, coupled structure such tasks depend on. We present DexRoam, a complete system for learning mobile bimanual dexterous manipulation from human demonstrations, in which whole-body motion remains continuous and coupled throughout the human-to-robot transfer process. To enable scalable collection of whole-body human manipulation demonstrations, we develop a tracker-free capture system using only a consumer VR headset and a head-mounted stereo camera, without external cameras or motion trackers. We then perform three explicit alignment stages—embodiment, action-semantic, and temporal—to map captured motion into the robot action space, preserving fine-grained whole-body motion and allowing human and robot demonstrations to be jointly learned by standard VLA policies. Real-world experiments with different VLA backbones show that human demonstrations consistently improve policy learning across training paradigms, raising average success from 29% to 56% on GR00T N1.7 and from 32% to 57% on $\pi_{0.5}$, while matching robot-only training with half the robot demonstrations. Ablations confirm that each alignment stage is necessary. These results highlight the potential of human demonstrations for scalable whole-body mobile manipulation with preserved fine-grained motion structure.**

I. Introduction

Mobile bimanual dexterous manipulation requires robots to simultaneously perform locomotion, torso adjustment, bimanual coordination, active vision, and finger-level dexterous interaction within a single trajectory. Unlike isolated manipulation or locomotion, these tasks require continuous coordination among multiple body components throughout execution. Training such systems therefore demands large-scale, high-quality whole-body action data. However, collecting robot demonstrations at this scale remains challenging as robot embodiments become increasingly complex. Existing mobile manipulation platforms, such as Human Support Robot [1] and Stretch [2], provide important foundations for embodied intelligence, while recent whole-body teleoperation systems extend data collection toward bimanual and humanoid manipulation through bilateral interfaces, kinematic-matched platforms, and wearable devices [3]–[6]. Nevertheless, these approaches require increasingly specialized hardware interfaces, making large-scale data collection for mobile bimanual dexterous robots expensive and difficult to scale.

Egocentric human demonstrations provide a scalable alternative for alleviating the robot data bottleneck. Large-scale egocentric datasets have revealed rich human activity priors [7]–[9], while recent systems enable collecting robot-relevant human demonstrations through scalable interfaces [10]–[13]. Building on these advances, recent works demonstrate that human demonstrations can directly improve robot policy learning for dexterous manipulation [14]–[16], bimanual manipulation [17], and vision-language-action models trained from egocentric data [18]. However, directly transferring the full human motion space to robots remains challenging due to differences in morphology, control interfaces, and execution timescales between humans and robots.

To reduce this transfer difficulty, prior human-to-robot works often simplify or restructure human motion representations before policy learning. Instead of directly transferring the complete human motion space, these approaches introduce structured intermediate representations that reduce embodiment discrepancies and make robot learning more tractable. Recent mobile whole-body manipulation systems adopt different forms of such representation transformations, including simplified locomotion and hand representations [19], phase-based decomposition of navigation and manipulation [20], and controller-assisted whole-body execution through task-level targets [21]–[24]. Other approaches avoid directly transferring human actions and instead leverage human videos for representation learning or foundation model pretraining [25]–[28].

These representation transformations reduce the complexity of human–robot transfer, but they inherently introduce a trade-off between transferability and preserving human motion information. This trade-off is often acceptable for robots with limited action complexity and weak whole-body coupling, but becomes restrictive for mobile bimanual dexterous manipulation, where successful execution requires continuous coordination among locomotion, upper-body motion, and dexterous interaction. Specifically, existing approaches face several limitations. First, compressing continuous human motion into discrete or low-dimensional representations can remove fine-grained motion variations, preventing policies from fully exploiting the structure in human demonstrations. Second, decomposing long-horizon behaviors into independent phases can weaken dependencies between different stages of execution, making it difficult to jointly learn locomotion and upper-body manipulation within a unified policy. For example, separating navigation and manipulation limits the ability to perform tasks requiring tightly coupled arm–base coordination. Finally, in mobile bimanual manipulation, the motion of one body component often changes the feasible workspace and control constraints of others. Representing the base, torso, arms, and hands as independent control objectives ignores these interactions, limiting the ability to learn the whole-body coordination required for mobile dexterous manipulation.

This raises the central question of this work: **Can human demonstrations be transferred to robots while resolving embodiment mismatch and preserving the fine-grained motion structure required for complex whole-body behaviors?**

We present **DexRoam** (Fig. 1), a complete system that answers this question by learning mobile bimanual dexterous manipulation from human demonstrations end to end, spanning portable egocentric capture, explicit whole-body human-to-robot alignment, and real-robot policy learning. The system is built around a single design principle: the continuous, coupled structure of whole-body motion is carried from the demonstrator's body to the robot's action chunks without discretized locomotion primitives, phase decomposition, or independent per-component control. Each stage is designed so that the continuous, coupled structure of whole-body motion survives from the demonstrator's body to the robot's action.

The first stage is a **tracker-free egocentric whole-body capture system**. It recovers whole-body human motion together with egocentric video using only a consumer VR headset and a head-mounted stereo camera, requiring no external cameras, environment markers or body-worn trackers. The camera stream is displayed inside the headset in real time, so the demonstrator's view during collection is exactly the observation the policy later sees. This keeps collection portable and scalable, while providing raw whole-body motion in human space as the input to the rest of the system.

The second stage is **explicit human-to-robot alignment**, which brings this motion into the robot's action space without compressing it. Three mismatches are resolved in turn. *Embodiment mismatch*, from structural differences between humans and robots, is handled by root-centric whole-body retargeting. *Action-semantic mismatch*, between human motion representations and robot control interfaces, is handled by robot-centric relative action construction. *Temporal mismatch*, from the different execution timescales of humans and robots, is handled by task-progress temporal resampling. After alignment, human and robot demonstrations occupy the same observation–action space, so a standard VLA policy trains on both jointly, without human-specific action heads, and directly predicts complete whole-body action chunks.

We validate the full pipeline on a real mobile bimanual dexterous robot across five manipulation tasks, two VLA

backbones, and varying robot-data budgets. Transferred human trajectories replay on the robot with centimeter-level error and no IK failures, and alignment substantially reduces the human–robot action distribution gap. Training on the aligned data improves policy performance consistently across tasks and backbones, and matches robot-only training with half the robot demonstrations. Ablations show that removing action-semantic or temporal alignment substantially degrades the policy, confirming that each stage is necessary for the pipeline to work. Our contributions are:

- *DexRoam*, a complete working system for learning mobile bimanual dexterous manipulation from human demonstrations, integrating portable egocentric capture, explicit whole-body alignment, and real-robot policy learning under a single continuous, coupled action interface.
- *A tracker-free egocentric whole-body capture system* that records full-body motion, hand articulation, and egocentric stereo video with only a consumer VR headset and a head-mounted camera, with in-headset streaming that keeps the demonstrator's view consistent with the policy observation.
- *An explicit three-stage alignment* resolving embodiment, action-semantic, and temporal mismatches while preserving continuous whole-body motion structure, so that human and robot data are interchangeable for a standard VLA policy.
- *Extensive real-world evaluation* across five tasks, two VLA backbones, and different robot-data budgets, covering trajectory executability, distribution alignment, policy improvement, data efficiency, and ablations.

II. Problem Definition

We study the transfer of human demonstrations to mobile bimanual dexterous robots. We define the robot whole-body action space as $\mathcal{A}$, spanning the mobile base, torso, dual arms, active head, and dexterous hands, and require that the transferred demonstrations satisfy a *fine-grained whole-body action interface* with three properties: **(1) Continuous action space.** All action dimensions remain continuous rather than being discretized. **(2) Temporally continuous prediction.** The complete action sequence is continuously predicted by a single policy, without phase switching at deployment. **(3) Coupled whole-body prediction.** Actions of all body components are jointly modeled in a unified policy output space, rather than being treated as independent control objectives during deployment. Under this setting, we study how to effectively transfer human demonstrations to mobile bimanual dexterous robots while satisfying the above action interface constraints.

The training corpus consists of two sources: robot demonstrations $\mathcal{D}_{\text{robot}}$ collected on the target embodiment, whose actions already lie in $\mathcal{A}$, and human demonstrations $\mathcal{D}_{\text{human}}$, whose motions reside in human space. We therefore require a human-to-robot transformation $\mathcal{R}$ satisfying two conditions:

$$\mathcal{R}(\mathcal{D}_{\text{human}}) \subset \mathcal{S}, \qquad p_{\mathcal{R}(\mathcal{D}_{\text{human}})} \approx p_{\mathcal{D}_{\text{robot}}},$$

where $\mathcal{S}$ is the observation–action space of the target embodiment, whose action component $\mathcal{A}$ satisfies the three properties above. The first condition makes both sources consumable by a single policy without human-specific action heads; the second ensures the transferred data do not pull the policy away from the action distribution it encounters at deployment. Only under both conditions can $\mathcal{R}(\mathcal{D}_{\text{human}})$ and $\mathcal{D}_{\text{robot}}$ be trained on jointly under a single objective. We define the policy as $\pi_\theta(\mathbf{A}_t \mid \mathbf{I}_t, \mathbf{P}_t, \ell)$, where the policy receives an egocentric visual observation $\mathbf{I}_t$, a proprioceptive state history $\mathbf{P}_t$, and a language instruction ℓ, and outputs a continuous whole-body action chunk $\mathbf{A}_t \in \mathcal{A}$. After alignment, transformed human demonstrations join the training corpus $\mathcal{D}$ alongside robot demonstrations, enabling joint learning under the same policy and action interface.

III. DexRoam

In this section, we present DexRoam. Sec. III-A introduces the data collection system, Sec. III-B presents the three-stage alignment addressing embodiment, action-semantic, and temporal mismatches, and Sec. III-B4 describes policy learning under the unified observation–action interface.

A. Hardware Platform and Data Collection System

*1) **Hardware platform**:* The Astribot–XHand platform comprises an omnidirectional mobile base, an actuated torso, two 7-DoF arms, a 2-DoF active head, and two 12-DoF XHand dexterous hands (Fig. 2a, right), supporting coordinated whole-body mobile dexterous manipulation.

*2) **Robot demonstration collection**:* We build a whole-body teleoperation system (Fig. 2a) for collecting robot demonstrations on the target mobile bimanual dexterous platform. The Meta Quest 3S captures upper-body poses, VR controllers provide wrist pose commands, and MANUS Pro gloves capture hand motions, with these inputs jointly mapped to the robot base, end-effectors, and XHand joints, enabling coupled whole-body mobile dexterous teleoperation. The system adopts a single-operator control paradigm, where the operator simultaneously controls robot locomotion and whole-body manipulation, eliminating the need for an additional operator dedicated to base navigation. Since wearing gloves makes conventional keyboard or button-based interaction inconvenient, we further design a gesture-based interface for episode state management, allowing the operator to start, stop, save, and discard episode directly through hand gestures.

*3) **Tracker-free egocentric whole-body data collection**:* Human demonstrations are captured with only a Meta Quest 3 headset and a head-mounted ZED Mini stereo camera (Fig. 2b), requiring no external motion-capture cameras, no environment markers and no body-worn trackers. The Meta Quest 3 captures whole-body motion through WebXR [29], including 6-DoF headset pose, 33 body-joint poses, and 25 hand landmarks per hand, which together form the whole-body human state $\mathbf{h}_t$. Meanwhile, the ZED Mini records egocentric stereo RGB images with timestamps, providing first-person visual observations for each demonstration. Fig. 2b (right) shows a captured whole-body mobile manipulation trajectory, in which head pose, hand keypoints, and the full-body skeleton evolve continuously as the demonstrator moves through the

Fig. 2. **DexRoam data collection system.** (a) Follow-style teleoperation lets one operator drive base, torso, arms, head, and dexterous hands simultaneously via a Meta Quest 3S, VR controllers, and MANUS Pro gloves, with gesture-based episode management. (b) Tracker-free egocentric capture uses a Meta Quest 3 and a head-mounted ZED Mini to record 6-DoF head pose, 33 body joints, 25 joints per hand, and timestamped egocentric stereo RGB, with no external cameras, marker or trackers. Real-time in-headset video streaming keeps the demonstrator's view identical to the policy observation.

scene. A gesture-based interface is also designed for episode management, allowing the operator to start, stop, save, and discard recordings directly through gestures.

The key design choice of the capture system is *egocentric video streaming*: the ZED Mini stream is displayed inside the headset in real time (Fig. 2b), so what the operator sees during collection is exactly the egocentric observation used later for policy training. This keeps visual observation, visual attention, and motor intent consistent throughout an episode. Such consistency is important for mobile dexterous manipulation, where the viewpoint continuously changes as the robot moves its base, body, and head.

B. *Human-to-Robot Alignment*

DexRoam instantiates the human-to-robot transformation $\mathcal{R}$ through three explicit alignment steps(Fig. 3), which resolve embodiment, action-semantic, and temporal mismatches between human demonstrations and robot execution spaces.

*1) **Embodiment Alignment**:* Humans and robots have different kinematic trees, so neither joint angles nor global poses can be copied directly into an executable action. To reduce dependence on global coordinate frames and preserve body-relative motion structure, we introduce a time-varying pelvis-centric root frame $\mathcal{F}_{\mathrm{H}}$ to align human motion with the robot embodiment. The origin of $\mathcal{F}_{\mathrm{H}}$ is the orthogonal projection of the pelvis onto the ground plane (*Projection* in Fig. 3a); its forward axis is the horizontal projection of the pelvis's local forward direction, and its vertical axis points opposite to gravity. Let $\mathbf{z}_{\mathrm{H}} = -\widehat{{}^{\mathrm{W}}\mathbf{g}}$ denote the upward axis of the human root frame, where ${}^{\mathrm{W}}\mathbf{g}$ is the gravity vector expressed in the world frame. The ground-plane projection operator is defined as $\mathbf{\Pi}_{\mathrm{G}} = \mathbf{I} - \mathbf{z}_{\mathrm{H}}\mathbf{z}_{\mathrm{H}}^{\top}$. Assuming that the world origin lies on the ground plane, the pose of $\mathcal{F}_{\mathrm{H}}$ is defined as:

$$\mathbf{x}_t = \mathbf{\Pi}_{\mathrm{G}}\,\widehat{{}^{\mathrm{W}}\mathbf{f}_{\mathrm{pelvis},t}}, \qquad \mathbf{y}_t = \mathbf{z}_{\mathrm{H}} \times \mathbf{x}_t,$$

$${}^{\mathrm{W}}\mathbf{T}_{\mathrm{H},t} = \begin{bmatrix} \mathbf{x}_t & \mathbf{y}_t & \mathbf{z}_{\mathrm{H}} & \mathbf{\Pi}_{\mathrm{G}}\,{}^{\mathrm{W}}\mathbf{p}_{\mathrm{pelvis},t} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Aligned with the robot base motion plane, $\mathcal{F}_{\mathrm{H}}$ establishes a human-centric reference frame shared by locomotion and manipulation. It serves two purposes: transferring the human locomotion trajectory to the robot chassis motion and representing upper-body movements as body-relative trajectories, reducing dependence on global scene coordinates. Specifically, torso motion, head orientation, and wrist trajectories are expressed in $\mathcal{F}_{\mathrm{H}}$ through

$${}^{\mathrm{H}}\mathbf{T}_{J,t} = ({}^{\mathrm{W}}\mathbf{T}_{\mathrm{H},t})^{-1}\,{}^{\mathrm{W}}\mathbf{T}_{J,t},$$

where $J \in \{\mathrm{chest}, \mathrm{head}, \mathrm{l{-}wrist}, \mathrm{r{-}wrist}\}$. The resulting relative trajectories are then converted into robot commands: the human root trajectory is mapped to the chassis motion, the chest pose is used as the torso target, and the left/right wrist poses are transformed into arm end-effector targets. For the head, although its pose is represented in $\mathcal{F}_{\mathrm{H}}$, only the yaw and pitch components are extracted as the 2-DoF head command due to the robot's mechanical constraints. Fig. 3b illustrates the resulting per-component trajectories in robot space. For the dexterous hands, we adopt dex-retargeting [30] to retarget human hand poses into the 12-DoF XHand commands, bridging the morphology gap between human and robot hands.

*2) **Action-Semantic Alignment**:* Retargeting yields poses, not actions: the goal is not merely to reach the robot's pose space, but to construct targets carrying the same semantics as robot control. Absolute poses fail this test, as they still encode demonstrator-specific configuration such as height, shoulder width, arm length, and initial posture. We therefore express continuous motion as relative transformations in each body component's local frame, which specify how a component moves from where it currently is rather than where it should

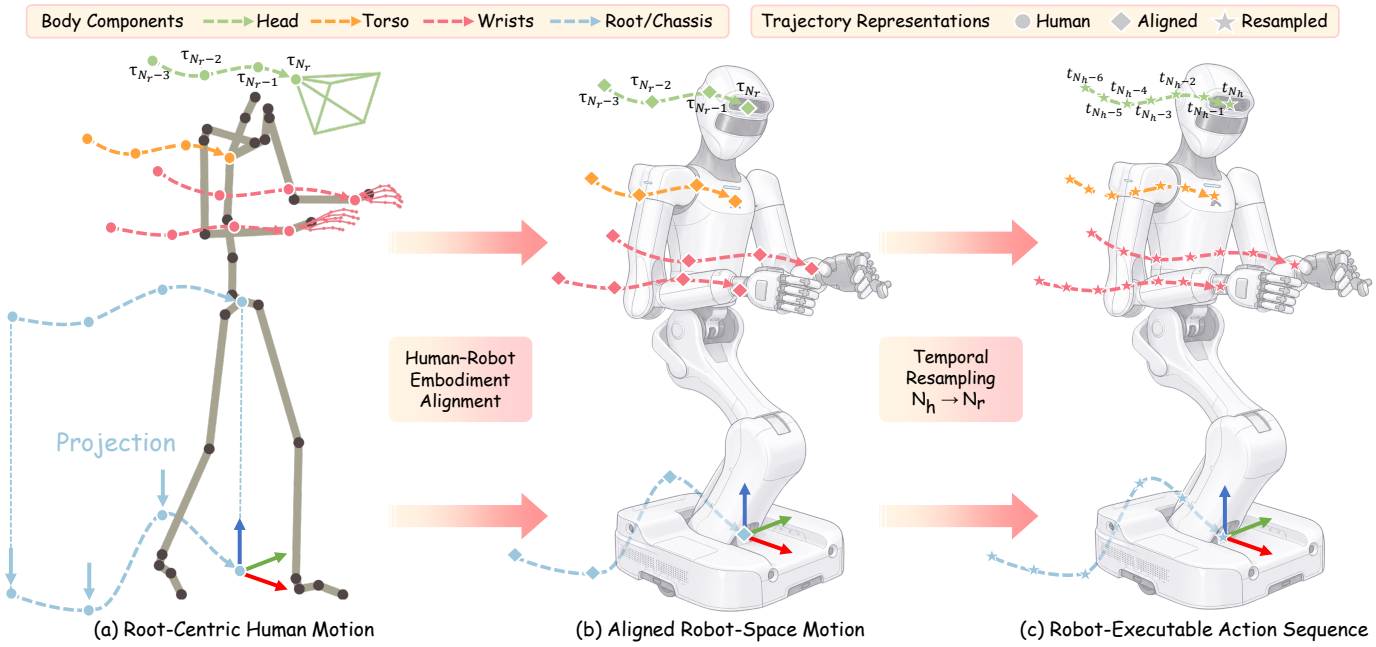

Fig. 3. **Explicit human-to-robot alignment in DexRoam.** (a) Human motion is expressed in a pelvis-centric root frame projected onto the ground plane, yielding root/chassis, torso, wrist, and head trajectories free of global scene coordinates. (b) Embodiment alignment maps each component to its robot counterpart, and action-semantic alignment re-expresses the results as robot-centric relative transformations. (c) Temporal alignment resamples N_h human steps into N_r waypoints evenly spaced in task progress, matching the execution timescale of robot demonstrations.

be in absolute coordinates, and are thus independent of the configuration in which the demonstration was recorded. For upper-body components, poses are already represented in the human root frame $\mathcal{F}_{\mathrm{H}}$, we compute the relative motion for $J \in \{\mathrm{chest}, \mathrm{l{-}wrist}, \mathrm{r{-}wrist}\}$ as

$$\Delta \mathbf{T}_t^J = ({}^{\mathrm{H}}\mathbf{T}_{J,t})^{-1\,{}^{\mathrm{H}}}\mathbf{T}_{J,t+1}$$

The resulting local transformations describe how each component evolves from its current state and preserve transferable motion patterns such as reaching, lifting, and retracting.

For locomotion, the root frame encodes the human movement. We compute the relative transformation between consecutive root frames and represent it as planar motion:

$$\Delta \mathbf{T}_t^{\mathrm{H}} = ({}^{\mathrm{W}}\mathbf{T}_{\mathrm{H},t})^{-1\,{}^{\mathrm{W}}}\mathbf{T}_{\mathrm{H},t+1}, \Delta \boldsymbol{\xi}_t^{\mathrm{H}} = [\Delta x_t, \Delta y_t, \Delta \psi_t]^\top$$

Since $\mathcal{F}_{\mathrm{H}}$ is constructed on the ground plane with a gravity-aligned vertical axis, its relative motion is inherently planar. Compared with absolute trajectories, this representation removes dependence on scene coordinates and initial positions while providing explicit locomotion semantics, enabling continuous coupling between base movement and whole-body manipulation. The 2-DoF head command and the 12-DoF hand command already lie in robot joint space and are used directly.

*3) **Temporal Alignment**:* We align demonstrations by task progress rather than wall-clock time. Humans complete the same task faster and in fewer steps than the robot, so the two advance task progress at different rates per policy step. We resample each retargeted human trajectory into $\bar{N}_r$ evenly spaced waypoints in normalized task progress $s \in [0, 1]$, where $\bar{N}_r$ is the mean robot episode length for that task:

$$\tilde{T}_k = \mathrm{Interp}(T_{1:N_h}, s_k),\ s_k = \frac{k-1}{\bar{N}_r - 1},\ 1 \le k \le \bar{N}_r$$

with N_h the original human trajectory length. After resampling, the two sources advance the same fraction of task progress per policy step, as illustrated in Fig. 3c. The resampled waypoints preserve the original spatial trajectory while matching the temporal distribution of robot executions. Positions are interpolated linearly and rotations by SLERP, while images and discrete signals use nearest-neighbor matching; base yaw is unwrapped before interpolation to avoid discontinuities around $\pm\pi$. Resampling only adjusts the temporal discretization of the trajectories and does not modify the robot control frequency or action chunk size.

After alignment, human and robot demonstrations become interchangeable under the same policy interface. The resulting action representation satisfies the three properties defined in Sec. II. First, all actions remain continuous without discretization. Second, the complete whole-body trajectory is predicted by a single policy without phase-based execution. Third, all body components are jointly represented in one unified action space. This unified interface enables human demonstrations to retain their fine-grained motion structure while becoming executable in the robot action space.

4) Policy Learning: Because aligned human data and robot data share the same observation–action space, they can be directly consumed by a standard VLA policy without human-specific action heads or input branches. We instantiate $\pi_\theta(A_t | I_t, P_t, \ell)$ with two VLA backbones, GR00T N1.7 [31] and $\pi_0.5$ [32]. Each policy takes egocentric RGB observations, proprioceptive state history, and language instructions as inputs, and predicts whole-body action chunks. We design three strategies for incorporating aligned human demonstrations into policy learning: **Human–Robot Cotrain**, jointly training on aligned human and robot demonstrations; **Human Pretrain + Robot FT**, pretraining on aligned human demon-

Fig. 4. **Real-world mobile bimanual dexterous manipulation tasks.** Rollouts of the five evaluation tasks, one per row: Pick Chips Can, Pour Water, Throw Trash, Deliver Fruit, and Push Chair & Close Laptop. They span fine-grained dexterity, base–manipulation coupling, and long-horizon whole-body interaction.

strations and finetuning on robot demonstrations only; and **Human Pretrain + Cotrain FT**, pretraining on aligned human demonstrations and finetuning on the human–robot mixture. All strategies use the same observation–action interface and downstream finetuning budget.

IV. Experiments

A. Experimental Setup

*1) **Tasks**:* We evaluate DexRoam on five tasks covering complementary mobile manipulation capabilities (Fig. 4). **Pick Chips Can** requires the robot to navigate to a table, grasp a chips can, and place it in a tray, emphasizing precise reaching and finger-level grasping. **Pour Water** requires grasping a cup and pouring into a target container, testing continuous wrist–hand coordination. **Throw Trash** requires grasping an object, moving to a bin, and releasing it accurately; the object must remain stable during locomotion, while successful disposal depends on precise coordination between base positioning and manipulation. **Deliver Fruit** requires carrying a basket bimanually to a target workstation, stressing bimanual transport and coordinated turning during locomotion. **Push Chair & Close Laptop** chains long-range navigation, contact-rich chair pushing, and laptop closing, testing long-horizon execution together with continuous base–upper-body coordination and fine manipulation.

*2) **Evaluation protocol**:* We evaluate DexRoam at three complementary levels: trajectory executability, human–robot action distribution alignment, and policy performance.

Trajectory replay. We directly replay transferred human trajectories on the real robot and report trajectory completion, IK failure rate, and positional and rotational replay errors for the torso, arms, and head.

Action distribution alignment. We further examine whether the alignment process brings human-derived actions closer to real robot actions. For each task, we compare the human and robot whole-body action distributions at the retargeted-absolute stage and after full alignment. We use Maximum Mean Discrepancy (MMD) and Sliced Wasserstein Distance (SWD) for quantitative comparison and PCA for qualitative visualization. Detailed normalization, metric computation, and visualization protocols are provided in the Appendix.

Policy evaluation. We evaluate all four training settings described in Sec. III-B4 on all five tasks, with 20 real-world trials per task. Success rate (SR) captures whether the final task goal is achieved, but provides limited information about partial progress, particularly for mobile manipulation tasks. We therefore additionally use task completion score (TCS), which decomposes each task into ordered stages and measures the fraction of stages completed. Detailed TCS criteria are provided in the Appendix. We report both SR and TCS

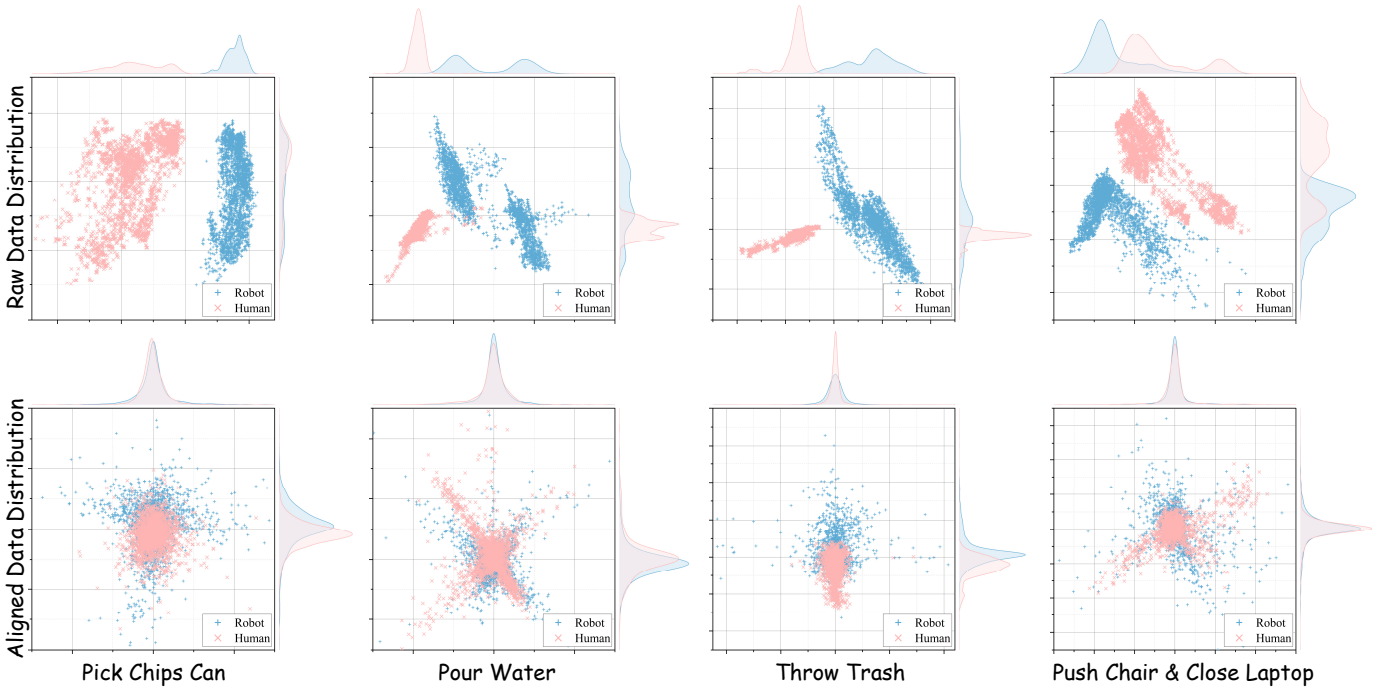

Fig. 5. **Human–robot action distribution before and after alignment.** PCA visualization of whole-body actions on representative tasks. Top: retargeted absolute human actions and robot actions before action-semantic and temporal alignment, which occupy clearly separated regions. Bottom: fully aligned robot-centric relative actions, where the two distributions largely overlap, with marginals shown along each axis.

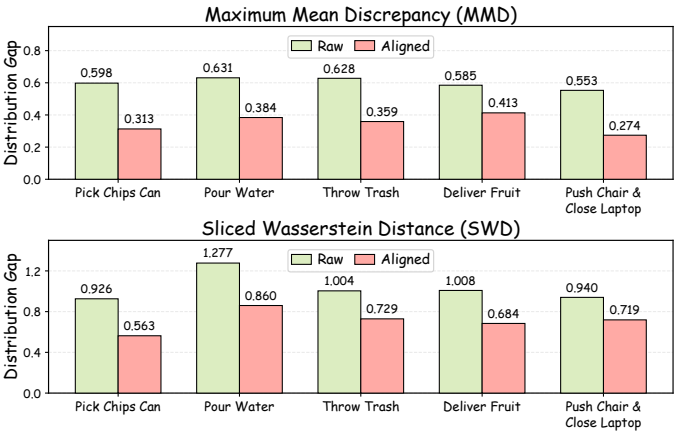

Fig. 6. **Quantitative human–robot action distribution alignment.** MMD and SWD between human-derived and robot whole-body actions on all five tasks, lower values indicate closer distributions. Alignment reduces both metrics on every task, by 41.7% in MMD and 31.1% in SWD on average.

for downstream policy evaluation. Based on these evaluation metrics, we investigate when, how, and why aligned human demonstrations benefit mobile bimanual dexterous manipulation. Specifically, we study four questions:

- **Q1: Transfer Quality.** How effectively does human-to-robot transfer convert human motion into executable and distribution-aligned robot training data?
- **Q2: Policy Learning.** Does aligned human supervision improve policy learning, and which training paradigm leverages it most effectively?
- **Q3: Data Efficiency.** How does aligned human data improve robot learning with limited robot demonstrations?
- **Q4: Alignment Analysis.** How do different alignment choices affect fine-grained human-to-robot transfer?

TABLE I: **Real-robot replay of transferred human trajectories.** Position and rotation errors of retargeted human motions across different body components, measured in millimeters (mm) and degrees (°).

Component	Metric	Pick Chips Can	Pour Water	Throw Trash	Average
Torso	Position	13.49	11.82	12.31	**12.54**
	Rotation	2.32	1.64	1.57	**1.84**
Left Arm	Position	3.75	4.27	4.02	**4.01**
	Rotation	0.51	0.58	0.57	**0.55**
Right Arm	Position	5.30	5.93	7.67	**6.30**
	Rotation	0.70	0.84	1.22	**0.92**
Head	Rotation	1.22	1.36	1.53	**1.37**
Completed		5/5	5/5	5/5	**15/15**

TABLE II: **Tracker-free vs. tracker-based capture.** Retargeting accuracy averaged over 15 trajectories across three tasks. Position and rotation errors are reported in millimeters and degrees, respectively.

Human Motion Capture	Torso		Left Arm		Right Arm		Head
	Pos.	Rot.	Pos.	Rot.	Pos.	Rot.	Rot.
Tracker-free (Ours)	**12.54**	**1.84**	**4.01**	**0.55**	6.30	0.92	1.37
PICO + Motion Trackers	13.16	2.06	4.12	0.57	**5.63**	**0.81**	**1.23**

B. Q1: Transfer Quality

*1) **Action distribution alignment**:* We examine whether the transferred human supervision is compatible with the action distribution of the real robot. Across all five tasks, alignment consistently reduces both distribution metrics, decreasing MMD by **41.7%** and SWD by **31.1%** on average (Fig. 6). The PCA visualization in Fig. 5 shows the same qualitative trend, with substantially greater overlap between human-derived and robot actions after alignment. Together with the replay results, these findings show that DexRoam produces human supervision that is not only executable on the target embodiment, but also better matched to real robot

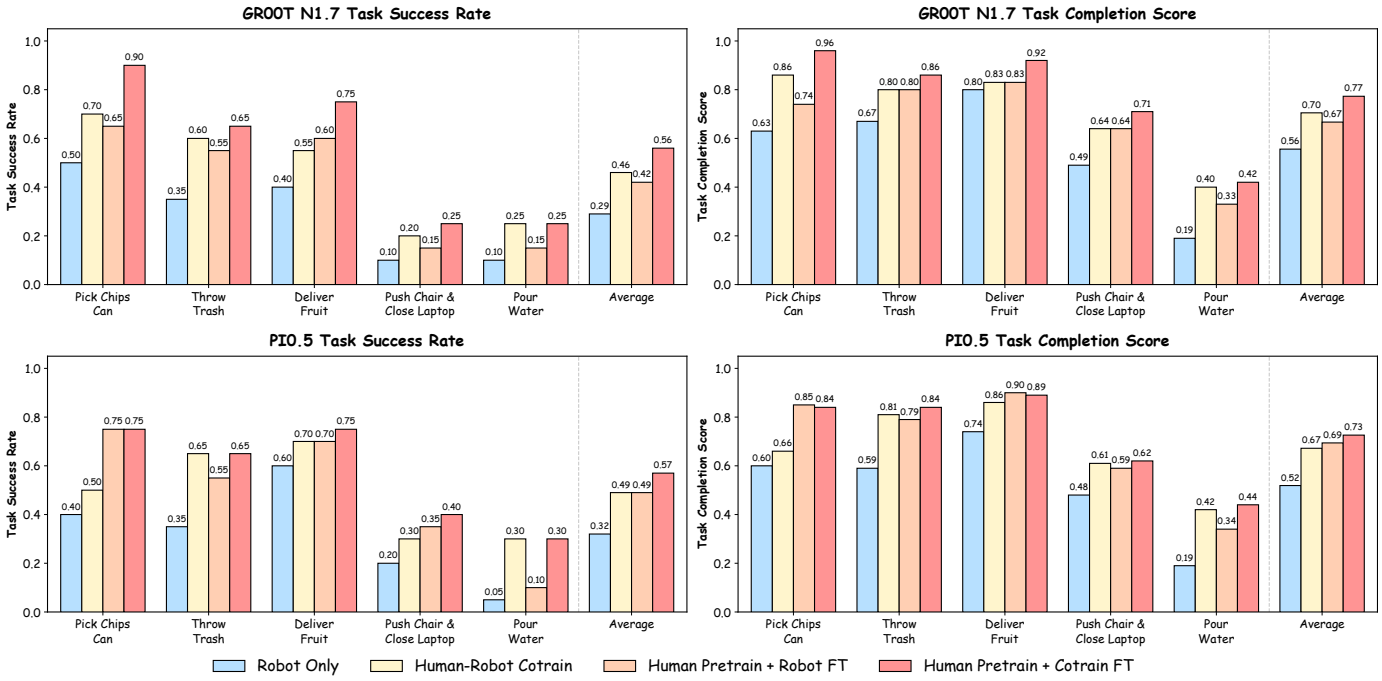

Fig. 7. **Main policy learning results.** Success rate (SR, left) and task completion score (TCS, right) of Robot-only and three human-assisted training strategies on five real-world tasks, for GR00T N1.7 (top) and $\pi_{0.5}$ (bottom). All three human-assisted strategies improve over Robot-only on every task and both backbones, with Human Pretrain + Cotrain FT best or tied-best on every task–backbone combination.

actions.

*2) **Real-robot executability**:* All 15 transferred human trajectories replay successfully on the real robot with no IK failures (Tab. I). Across the three tasks, positional replay errors remain at approximately the centimeter scale for the torso and millimeter scale for both arms, while rotational errors remain within a few degrees across all evaluated body components. Despite the different whole-body motion patterns involved, the transferred trajectories can therefore be consistently executed by the target robot, supporting their use as robot supervision.

*3) **Tracker-free capture fidelity**:* Since tracker-free human motion capture is a key component of our data collection system, we further examine whether this lightweight design compromises the quality of the transferred trajectories. For comparison, we construct a tracker-based counterpart using a PICO headset with five Motion Trackers and evaluate it under the same replay protocol. Both systems successfully replay all trajectories and achieve comparable replay errors across body components, with no systematic performance gap (Tab. II). This suggests that our tracker-free system substantially reduces sensing cost and wearable burden while preserving the quality required for subsequent human-to-robot transfer.

C. Q2: Policy Learning

Relative to Robot-only, all three human-assisted training paradigms improve both SR and TCS across both VLA backbones and all five tasks (Fig. 7). In terms of average SR, the three paradigms provide gains of **+17, +13, and +27 percentage points** on GR00T N1.7, and **+17, +17, and +25 percentage points** on $\pi_{0.5}$. The same positive trend holds across individual tasks, indicating that the benefit of aligned human supervision is not tied to a particular training paradigm, policy backbone, or manipulation setting. This remains true on Push Chair & Close Laptop, the longest-horizon and most tightly coupled task, where the best human-assisted setting improves success by **15 percentage points** on GR00T N1.7 and **20 percentage points** on $\pi_{0.5}$. Together, these results show that aligned human supervision provides a broadly useful learning signal for mobile bimanual dexterous manipulation.

The magnitude of the gain, however, depends on how the human supervision is used. Human Pretrain + Cotrain FT achieves the highest average performance on both backbones and is best or tied-best on every task–backbone combination. This suggests that using human data both to initialize task-relevant behavior and to retain human supervision during subsequent human–robot cotraining makes the most effective use of the transferred demonstrations.

Qualitatively, we also observe that human supervision can introduce useful motion variations absent from the robot demonstrations. On Pick Chips Can, the Robot-only policy predominantly approaches the can from the front and often knocks it over when its orientation is perturbed. With human supervision, the policy can instead adopt a side approach when appropriate, selecting different grasp motions according to the object orientation. Since this side-grasp pattern is absent from the robot demonstrations but present in the human data, the behavior provides a concrete example of how human supervision can expand the motion patterns available to the learned policy.

D. Q3: Data Efficiency

We fix 50 human demonstrations per task, vary the number of robot demonstrations over $\{0, 10, 25, 50\}$, and evaluate

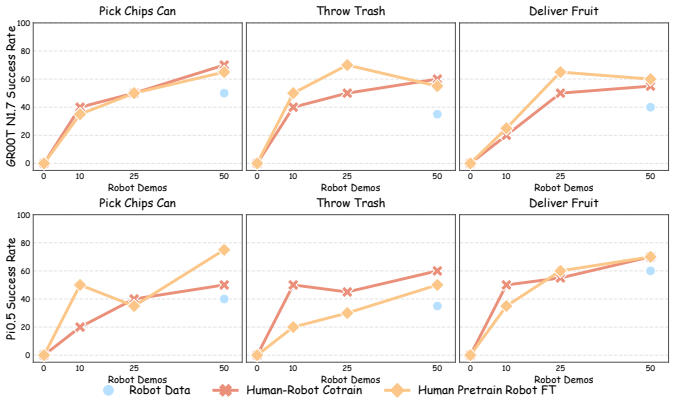

Fig. 8. **Robot-data efficiency with human supervision.** Success rate under robot demonstration budgets of $\{0, 10, 25, 50\}$ with 50 human demonstrations fixed per task. With aligned human data, success rises sharply once only a few robot demonstrations are available, and 25 already approach or match the 50-demo Robot-only baseline.

different human supervision training strategies, with Robot-only using 50 robot demonstrations as the reference (Fig. 8).

*1) **Human data reduces robot data requirements**:* With only 25 robot demonstrations, human-assisted training on GR00T N1.7 already matches or exceeds the 50-demo Robot-only baseline on all three tasks, effectively **reducing the required robot demonstrations by 50%**. On $\pi_{0.5}$, the same 25-demo setting remains within 5 percentage points of the 50-demo Robot-only baseline. These results show that aligned human supervision substantially improves robot-data efficiency, enabling comparable policy performance with fewer robot demonstrations.

*2) **Few-shot robot data activates human priors**:* With no robot demonstrations, neither human-assisted strategy achieves any task success. However, performance rises sharply once only 10 robot demonstrations are introduced. This suggests that aligned human data provides useful task and motion priors that can be activated with only a small amount of robot-specific supervision. At the same time, the failure of the human-only setting shows that human data amplifies rather than replaces robot data: robot demonstrations remain necessary to ground the policy in the target embodiment and execution domain.

The two training paradigms also exhibit different scaling behavior. Cotrain generally improves more steadily as robot data increases, whereas Human Pretrain + Robot FT is less stable. We hypothesize that finetuning exclusively on a limited robot dataset can partially overfit to the small robot distribution and weaken the diversity introduced during human pretraining, while cotraining continuously retains human supervision throughout optimization.

E. Q4: Alignment Analysis

We ablate the two alignment stages that can be removed in isolation, on Pick Chips Can and Throw Trash (Tab. III).

1) Absolute actions fail without semantic alignment: Replacing DexRoam's robot-centric relative action representation with absolute retargeted poses causes the policy to fail completely: success is 0% on both tasks and under all three training strategies (Tab. III). We attribute this failure to the fact that absolute poses map human and robot demonstrations into a shared coordinate representation while still retaining a substantial amount of human-specific information. During training, these factors pull the supervision signal away from the robot action distribution the policy will face at deployment, so the policy learns targets entangled with human-specific configurations rather than a stable robot control semantics.

TABLE III: **Ablation of alignment stages.** GR00T N1.7 success rate.

Task	Strategy	w/o Temporal Resampling	Absolute Action	**DexRoam (Ours)**
Pick Chips Can	Cotrain	0.65	0.00	**0.70**
	Pretrain + FT	0.30	0.00	**0.65**
	Pretrain + Cotrain FT	0.20	0.00	**0.90**
Throw Trash	Cotrain	0.40	0.00	**0.65**
	Pretrain + FT	0.50	0.00	**0.55**
	Pretrain + Cotrain FT	0.60	0.00	**0.65**

This distribution shift is further amplified over closed-loop execution. At the start of an episode, the robot's initial observation $\mathbf{o}_0 = (\mathbf{I}_0, \mathbf{P}_0)$ lies within the distribution of the robot data, so the first predicted chunk $\mathbf{a}_{0:H-1}$ still exhibits a plausible execution trend. However, because the training distribution has been pulled toward that of human-specific absolute actions, the model biases its predicted targets toward the demonstrator's absolute configurations and incurs a small error. Executing this error drives the next observation $\mathbf{o}_H = (\mathbf{I}_H, \mathbf{P}_H)$ off the training distribution, which in turn makes the following chunk $\mathbf{a}_{H:2H-1}$ more erroneous. Repeated over an episode, this forms a self-amplifying feedback loop that eventually drives the robot into a state from which it cannot recover, and the task fails.

Among the three strategies, Human Pretrain + Robot FT exhibits this behavior to a somewhat lesser degree under the absolute-action setting, since finetuning on robot data pulls the learned distribution partly back toward the robot's own. The finetuning budget, however, is not sufficient to recover it fully, and the resulting policy still completes no trials.

2) Temporal alignment improves policy learning: Removing task-progress temporal resampling reduces the average success rate from 68.3% to 44.2%, a drop of **24.1 percentage points** (Tab. III). This degradation is caused by the different execution timescales of human and robot demonstrations: without resampling, the same task progress can correspond to different future action chunks, introducing temporally inconsistent supervision. Task-progress temporal resampling aligns human and robot trajectories at the same execution phase, removing this temporal mismatch.

F. Generalization Under Distribution Shift

We additionally evaluate two distribution shifts on Pick Chips Can — a changed background scene and an unseen instruction phrasing. As shown in Fig. 9, we observe consistent improvements under both background and instruction shifts. For background shift, Robot-only drops from 50% to 20%, while Human Pretrain + Cotrain FT decreases from 90% to 55%, maintaining the highest robustness in the novel scene. For instruction shift, strategies that retain human data during

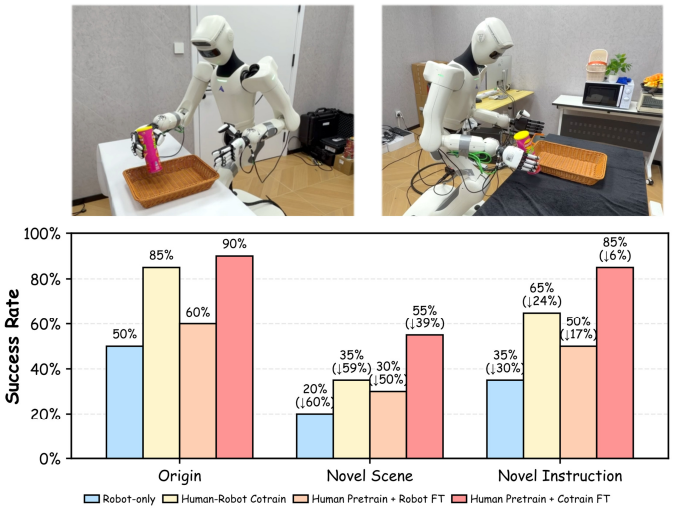

Fig. 9. **Generalization under distribution shift.** Evaluation on Pick Chips Can under the original setting, a novel scene, and an unseen instruction phrasing. The images above show the original and novel scenes. Human-assisted training degrades far less than Robot-only under both shifts, especially when human data is retained during finetuning.

finetuning show greater robustness, with Cotrain and Human Pretrain + Cotrain FT exhibiting 24% and 6% relative declines, respectively. These results suggest that aligned human supervision improves robustness beyond the training distribution, particularly when human data continues to participate during policy optimization.

V. Related Work

A. Mobile Manipulation

Embodied intelligence increasingly seeks generalist models that integrate multimodal perception, language understanding, world modeling, and action generation to accomplish diverse tasks in the physical world [31]–[44]. Mobile manipulation requires navigation, whole-body reachability, and object interaction to be coordinated over spatially extended tasks. Accessible platforms and data systems have progressed from Stretch and Mobile ALOHA to integrated or modular whole-body pipelines such as TidyBot++, BRS, TeleMoMa, MoMa-Teleop, HoMeR, and SuperSuit [2], [3], [45]–[50]. Humanoid systems including HumanPlus, HOMIE, TWIST/TWIST2, HumanoidExo, and Teleopit further extend teleoperation and policy learning to locomotion, posture, active vision, and dexterous hands [4], [5], [51]–[54]. Beyond collection interfaces, generalist and whole-body VLA or world-action models such as GR00T N1, PanoVLA, Ψ_0, OpenHLM, and ω-0 increasingly address language-conditioned bimanual, wheeled, and humanoid loco-manipulation [28], [31], [55]–[57]. HERMES also targets mobile bimanual dexterity but connects learned manipulation skills to a separate navigation module [58]. Despite this progress, many systems still require robot-in-the-loop collection, specialized interfaces, or substantial embodiment-specific robot data, motivating scalable robot-free demonstrations for continuous whole-body learning.

B. Human-Robot Alignment

Transferring human demonstrations to robots requires resolving differences in visual observations, morphology, action semantics, dynamics, and execution speed. Existing work learns cross-embodiment visual or task representations [25], [26], [59], expresses behavior through transferable affordances, flow, or point trajectories [60]–[63], or translates and retargets human data into robot-compatible observations and actions [64]–[71]. For mobile whole-body learning, recent systems often make transfer tractable through discrete locomotion commands, phase-aware policies, decoupled base and manipulation actions, or end-effector and keypoint targets executed by downstream controllers [19]–[24], [72]. These abstractions improve compatibility but may omit part of the original coupled motion. DexRoam instead explicitly aligns embodiment, action semantics, and task progress while retaining a continuous, temporally uninterrupted, jointly predicted base–torso–arm–head–hand action space.

C. Human Demonstrations for Robot Learning

Human demonstrations provide diverse behavioral priors at substantially lower collection cost than robot teleoperation. [73] Large egocentric datasets have progressed from first-person activities and hand–object interactions to paired viewpoints, metric geometry, full-body motion, and robot-oriented dexterous trajectories [7]–[9], [74]–[80]. Early work mainly extracted reusable representations or rewards, whereas portable interfaces and transfer pipelines now convert human experience into policy supervision for fixed-base, bimanual, and dexterous manipulation [10]–[18], [25], [26], [81]–[85]. Human-centric VLA and world-model pretraining further scale this idea across data sources and embodiments [86]–[90]. Complementary to these learning-based representations, recent works also investigate 3D-aware scene modeling for embodied perception [80], [84], [91]–[95]. More recently, EMMA, EgoHumanoid, HoMMI, HuMI, BifrostUMI, HALOMI, HERMES, WARP, and OpenHLM extend human supervision to mobile or whole-body policies [19]–[24], [56], [58], [69]. Preserving mobile, bimanual, and finger-level coordination in one robot-native policy interface nevertheless remains comparatively underexplored.

VI. Conclusion

We presented **DexRoam**, a complete system for learning mobile bimanual dexterous manipulation from egocentric whole-body human demonstrations. Rather than simplifying human motion to ease transfer, DexRoam preserves its continuous, coupled structure end to end. Whole-body motion and egocentric video are captured with only a consumer VR headset and a head-mounted stereo camera. Three explicit alignment stages—embodiment, action-semantic, and temporal—then map this motion into a unified whole-body robot action space, so that VLA policies can consume human and robot data jointly without human-specific action heads.

On a real mobile bimanual dexterous robot, transferred trajectories are directly executable, alignment substantially

closes the human–robot action distribution gap, and aligned human demonstrations consistently improve policy learning across five tasks and two VLA backbones while halving the required robot demonstrations. Ablations confirm that each alignment stage is necessary.

VII. Limitations

Our hands are position-controlled without force or tactile feedback, so commands cannot encode the closing force a firm grasp requires; combined with execution error, this yields grasps that touch but do not secure the object. Supervising hand commands from the Manus glove joint angles, rather than the robot hand state which saturates at contact, would partially help, while force or tactile feedback is a more fundamental remedy.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. T2600237, the Early Career Scheme (ECS) of the Research Grants Council (RGC) of the Hong Kong Special Administrative Region, China, under Project No. 26601126, the research funding under the HKUST-DXM AI for Finance Joint Laboratory (DXM25EG01), the National Natural Science Foundation of China under Grant No. 62476011, and the Beijing Natural Science Foundation under Grant No. L252060.

References

[1] T. Yamamoto, K. Terada, A. Ochiai, F. Saito, Y. Asahara, and K. Murase, "Development of human support robot as the research platform of a domestic mobile manipulator," *ROBOMECH Journal*, vol. 6, 04 2019.

[2] C. C. Kemp, A. Edsinger, H. M. Clever, and B. Matulevich, "The design of stretch: A compact, lightweight mobile manipulator for indoor human environments," 2022. [Online]. Available: https://arxiv.org/abs/2109.10892

[3] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," 2024. [Online]. Available: https://arxiv.org/abs/2401.02117

[4] Q. Ben, F. Jia, J. Zeng, J. Dong, D. Lin, and J. Pang, "Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit," 2025. [Online]. Available: https://arxiv.org/abs/2502.13013

[5] R. Zhong, Y. Sun, J. Wen, J. Li, C. Cheng, W. Dai, Z. Zeng, H. Lu, Y. Zhu, and Y. Xu, "Humanoidexo: Scalable whole-body humanoid manipulation via wearable exoskeleton," 2025. [Online]. Available: https://arxiv.org/abs/2510.03022

[6] R. Zhong, C. Cheng, J. Xu, Y. Wei, C. Guo, D. Zhang, W. Dai, and H. Lu, "Nuexo: A wearable exoskeleton covering all upper limb rom for outdoor data collection and teleoperation of humanoid robots," 2025. [Online]. Available: https://arxiv.org/abs/2503.10554

[7] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, A. Gebreselasie, C. Gonzalez, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolar, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanova, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Z. Zhao, Y. Zhu, P. Arbelaez, D. Crandall, D. Damen, G. M. Farinella, C. Fuegen, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. Newcombe, A. Oliva, H. S. Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, L. Torresani, M. Yan, and J. Malik, "Ego4d: Around the world in 3,000 hours of egocentric video," 2022. [Online]. Available: https://arxiv.org/abs/2110.07058

[8] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "The epic-kitchens dataset: Collection, challenges and baselines," 2020. [Online]. Available: https://arxiv.org/abs/2005.00343

[9] 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," 2024. [Online]. Available: https://arxiv.org/abs/2203.01577

[10] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu, "Dexcap: Scalable and portable mocap data collection system for dexterous manipulation," 2024. [Online]. Available: https://arxiv.org/abs/2403.07788

[11] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," 2024. [Online]. Available: https://arxiv.org/abs/2402.10329

[12] S. Chen, C. Wang, K. Nguyen, L. Fei-Fei, and C. K. Liu, "Arcap: Collecting high-quality human demonstrations for robot learning with augmented reality feedback," 2024. [Online]. Available: https://arxiv.org/abs/2410.08464

[13] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang, "Open-television: Teleoperation with immersive active visual feedback," 2024. [Online]. Available: https://arxiv.org/abs/2407.01512

[14] C. Yuan, R. Zhou, M. Liu, Y. Hu, S. Wang, L. Yi, C. Wen, S. Zhang, and Y. Gao, "Motiontrans: Human vr data enable motion-level learning for robotic manipulation policies," 2025. [Online]. Available: https://arxiv.org/abs/2509.17759

[15] R. Zheng, D. Niu, Y. Xie, J. Wang, M. Xu, Y. Jiang, F. Castañeda, F. Hu, Y. L. Tan, L. Fu, T. Darrell, F. Huang, Y. Zhu, D. Xu, and L. Fan, "Egoscale: Scaling dexterous manipulation with diverse egocentric human data," 2026. [Online]. Available: https://arxiv.org/abs/2602.16710

[16] T. Tao, M. K. Srirama, J. J. Liu, K. Shaw, and D. Pathak, "Dexwild: Dexterous human interactions for in-the-wild robot policies," 2026. [Online]. Available: https://arxiv.org/abs/2505.07813

[17] H. Bi, L. Wu, T. Lin, H. Tan, Z. Su, H. Su, and J. Zhu, "H-rdt: Human manipulation enhanced bimanual robotic manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2507.23523

[18] R. Yang, Q. Yu, Y. Wu, R. Yan, B. Li, A.-C. Cheng, X. Zou, Y. Fang, X. Cheng, R.-Z. Qiu, H. Yin, S. Liu, S. Han, Y. Lu, and X. Wang, "Egovla: Learning vision-language-action models from egocentric human videos," 2025. [Online]. Available: https://arxiv.org/abs/2507.12440

[19] M. Shi, S. Peng, J. Chen, H. Jiang, Y. Li, D. Huang, P. Luo, H. Li, and L. Chen, "Egohumanoid: Unlocking in-the-wild loco-manipulation with robot-free egocentric demonstration," 2026. [Online]. Available: https://arxiv.org/abs/2602.10106

[20] L. Y. Zhu, P. Kuppili, R. Punamiya, P. Aphiwetsa, D. Patel, S. Kareer, S. Ha, and D. Xu, "Emma: Scaling mobile manipulation via egocentric human data," 2025. [Online]. Available: https://arxiv.org/abs/2509.04443

[21] X. Xu, J. Park, H. Zhang, E. Cousineau, A. Bhat, J. Barreiros, D. Wang, J. Bohg, and S. Song, "Hommi: Learning whole-body mobile manipulation from human demonstrations," 2026. [Online]. Available: https://arxiv.org/abs/2603.03243

[22] Z. Zhao, Y. Zhao, G. Zhang, C. Liu, M. Zheng, and W. Lian, "Halomi: Learning humanoid loco-manipulation with active perception from human demonstrations," 2026. [Online]. Available: https://arxiv.org/abs/2606.18772

[23] H. Wang, C. Yu, Y. Hu, J. Zhang, Y. Li, and S. Luo, "Bifrostumi: Bridging robot-free demonstrations and humanoid whole-body manipulation," 2026. [Online]. Available: https://arxiv.org/abs/2605.03452

[24] R. Nai, B. Zheng, J. Zhao, H. Zhu, S. Dai, Z. Chen, Y. Hu, Y. Hu, T. Zhang, C. Wen, and Y. Gao, "Humanoid manipulation interface: Humanoid whole-body manipulation from robot-free demonstrations," 2026. [Online]. Available: https://arxiv.org/abs/2602.06643

[25] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," 2022. [Online]. Available: https://arxiv.org/abs/2203.12601

[26] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang, "Vip: Towards universal visual reward and representation via value-implicit pre-training," 2023. [Online]. Available: https://arxiv.org/abs/2210.00030

[27] S. Ye, J. Jang, B. Jeon, S. Joo, J. Yang, B. Peng, A. Mandlekar, R. Tan, Y.-W. Chao, B. Y. Lin, L. Liden, K. Lee, J. Gao, L. Zettlemoyer, D. Fox, and M. Seo, "Latent action pretraining from videos," 2025. [Online]. Available: https://arxiv.org/abs/2410.11758

[28] S. Wei, H. Jing, B. Li, Z. Zhao, J. Mao, Z. Ni, S. He, J. Liu, X. Liu, K. Kang, S. Zang, W. Yuan, M. Pavone, D. Huang, and Y. Wang, "ψ_0: An open foundation model towards universal humanoid loco-manipulation," 2026. [Online]. Available: https://arxiv.org/abs/2603.12263

[29] G. Yang, "VUER: An event-driven, declarative visualization toolkit for genai and robotics," 2025. [Online]. Available: https://github.com/vuer-ai/vuer

[30] 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 *Robotics: Science and Systems*, 2023.

[31] NVIDIA, J. Bjorck, N. C. Fernando Castañeda, X. Da, R. Ding, L. J. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, J. Jang, Z. Jiang, J. Kautz, K. Kundalia, L. Lao, Z. Li, Z. Lin, K. Lin, G. Liu, E. Llontop, L. Magne, A. Mandlekar, A. Narayan, S. Nasiriany, S. Reed, Y. L. Tan, G. Wang, Z. Wang, J. Wang, Q. Wang, J. Xiang, Y. Xie, Y. Xu, Z. Xu, S. Ye, Z. Yu, A. Zhang, H. Zhang, Y. Zhao, R. Zheng, and Y. Zhu, "GR00T N1: An open foundation model for generalist humanoid robots," in *ArXiv Preprint*, March 2025.

[32] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," 2025. [Online]. Available: https://arxiv.org/abs/2504.16054

[33] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky, "π_0: A vision-language-action flow model for general robot control," 2026. [Online]. Available: https://arxiv.org/abs/2410.24164

[34] P. Intelligence, A. Amin, R. Aniceto, A. Balakrishna, K. Black, K. Conley, G. Connors, J. Darpinian, K. Dhabalia, J. DiCarlo, D. Driess, M. Equi, A. Esmail, Y. Fang, C. Finn, C. Glossop, T. Godden, I. Goryachev, L. Groom, H. Hancock, K. Hausman, G. Hussein, B. Ichter, S. Jakubczak, R. Jen, T. Jones, B. Katz, L. Ke, C. Kuchi, M. Lamb, D. LeBlanc, S. Levine, A. Li-Bell, Y. Lu, V. Mano, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, C. Sharma, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, W. Stoeckle, A. Swerdlow, J. Tanner, M. Torne, Q. Vuong, A. Walling, H. Wang, B. Williams, S. Yoo, L. Yu, U. Zhilinsky, and Z. Zhou, "$\pi^*_{0.6}$: a vla that learns from experience," 2025. [Online]. Available: https://arxiv.org/abs/2511.14759

[35] J. Ji, D. Hong, B. Zhang, B. Chen, J. Dai, B. Zheng, T. A. Qiu, J. Zhou, K. Wang, B. Li, S. Han, Y. Guo, and Y. Yang, "PKU-SafeRLHF: Towards multi-level safety alignment for LLMs with human preference," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 31 983–32 016. [Online]. Available: https://aclanthology.org/2025.acl-long.1544/

[36] Y. Li, X. Wei, X. Chi, Y. Li, Z. Zhao, H. Wang, N. Ma, M. Lu, and S. Han, "Manipdreamer3d: synthesizing plausible robotic manipulation video with occupancy-aware 3d trajectory," in *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. [Online]. Available: https://doi.org/10.1609/aaai.v40i8.37595

[37] Y. Li, X. Wei, J. Cao, H. Wang, X. Chi, C. Bai, Q. Sun, J. Li, X. Zhang, P. Jia, J. Tang, S. Han, and S. Zhang, "Wam4d: Fast 4d world action model via spatial register tokens," 2026. [Online]. Available: https://arxiv.org/abs/2606.14048

[38] Q. Xu, J. Liu, R. Zhou, S. Shi, N. Han, Z. Liu, C. Gu, S. Gu, Y. Yue, G. Huang, W. Zheng, S. Han, P. Jia, and S. Zhang, "Twinrl: Digital twin-driven reinforcement learning for real-world robotic manipulation," 2026. [Online]. Available: https://arxiv.org/abs/2602.09023

[39] Y. Lou, Y. Ye, Y. Fu, J. Cen, X. Chi, Y. Lyu, P. Jia, S. Han, Z. Lu, and S. Zhang, "Dream-tac: A unified tactile world action model for contact-rich robot manipulation," *arXiv preprint arXiv:2606.08737*, 2026.

[40] X. Chi, P. Jia, C.-K. Fan, X. Ju, W. Mi, K. Zhang, Z. Qin, W. Tian, K. Ge, H. Li, Z. Qian, A. Chen, Q. Zhou, Y. Jia, J. Liu, Y. Dai, Q. Wuwu, C. Bai, Y.-K. Wang, Y. Li, L. Chen, Y. Bao, Z. Jiang, J. Zhu, K. Tang, R. An, Y. Luo, Q. Feng, S. Zhou, C. min Chan, C. Hou, W. Xue, S. Han, Y. Guo, S. Zhang, and J. Tang, "Wow: Towards a world omniscient world model through embodied interaction," 2025. [Online]. Available: https://arxiv.org/abs/2509.22642

[41] Y. Jia, J. Liu, S. Liu, R. Zhou, W. Yu, Y. Yan, X. Chi, Y. Guo, B. Shi, and S. Zhang, "Video2act: A dual-system video diffusion policy with robotic spatio-motional modeling," *arXiv preprint arXiv:2512.03044*, 2025.

[42] C.-K. Fan, X. Chi, X. Ju, H. Li, Y. Bao, Y.-K. Wang, L. Chen, Z. Jiang, K. Ge, Y. Li, W. Mi, Q. Wuwu, P. Jia, Y. Luo, K. Zhang, Z. Qin, Y. Dai, S. Han, Y. Guo, S. Zhang, and J. Tang, "Wow, wo, val! a comprehensive embodied world model evaluation turing test," 2026. [Online]. Available: https://arxiv.org/abs/2601.04137

[43] X. Chi, K. Ge, J. Liu, S. Zhou, P. Jia, Z. He, Y. Liu, T. Li, L. Han, S. Han, S. Zhang, and Y. Guo, "Mind: Learning a dual-system world model for real-time planning and implicit risk analysis," 2025. [Online]. Available: https://arxiv.org/abs/2506.18897

[44] K. Zhang, K. Ge, X. Chi, R. Zhang, S. Shi, Z. Dong, S. Han, and S. Zhang, "Can world models benefit vlms for world dynamics?" 2025. [Online]. Available: https://arxiv.org/abs/2510.00855

[45] J. Wu, W. Chong, R. Holmberg, A. Prasad, Y. Gao, O. Khatib, S. Song, S. Rusinkiewicz, and J. Bohg, "Tidybot++: An open-source holonomic mobile manipulator for robot learning," 2024. [Online]. Available: https://arxiv.org/abs/2412.10447

[46] Y. Jiang, R. Zhang, J. Wong, C. Wang, Y. Ze, H. Yin, C. Gokmen, S. Song, J. Wu, and L. Fei-Fei, "Behavior robot suite: Streamlining real-world whole-body manipulation for everyday household activities," 2025. [Online]. Available: https://arxiv.org/abs/2503.05652

[47] S. Dass, W. Ai, Y. Jiang, S. Singh, J. Hu, R. Zhang, P. Stone, B. Abbatematteo, and R. Martín-Martín, "Telemoma: A modular and versatile teleoperation system for mobile manipulation," 2024. [Online]. Available: https://arxiv.org/abs/2403.07869

[48] D. Honerkamp, H. Mahesheka, J. O. von Hartz, T. Welschehold, and A. Valada, "Whole-body teleoperation for mobile manipulation at zero added cost," 2025. [Online]. Available: https://arxiv.org/abs/2409.15095

[49] P. Sundaresan, R. Malhotra, P. Miao, J. Yang, J. Wu, H. Hu, R. Antonova, F. Engelmann, D. Sadigh, and J. Bohg, "Homer: Learning in-the-wild mobile manipulation via hybrid imitation and whole-body control," 2025. [Online]. Available: https://arxiv.org/abs/2506.01185

[50] T. Chen, H. Wu, J. Wang, X. Li, Z. Jin, and L. Fang, "Supersuit: An isomorphic bimodal interface for scalable mobile manipulation," 2026. [Online]. Available: https://arxiv.org/abs/2603.06280

[51] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "Humanplus: Humanoid shadowing and imitation from humans," 2024. [Online]. Available: https://arxiv.org/abs/2406.10454

[52] Y. Ze, Z. Chen, J. P. Araújo, Z. ang Cao, X. B. Peng, J. Wu, and C. K. Liu, "Twist: Teleoperated whole-body imitation system," 2025. [Online]. Available: https://arxiv.org/abs/2505.02833

[53] Y. Ze, S. Zhao, W. Wang, A. Kanazawa, R. Duan, P. Abbeel, G. Shi, J. Wu, and C. K. Liu, "TWIST2: Scalable, portable, and holistic humanoid data collection system," 2025. [Online]. Available: https://arxiv.org/abs/2511.02832

[54] B. Wu, Z. Xu, X. Fan, D. Li, and X. Huang, "Teleopit: A full-embodiment humanoid teleoperation system," 2026. [Online]. Available: https://arxiv.org/abs/2608.01834

[55] D. Yang, H. Chen, X. Chen, L. Liu, M. Li, C. Tu, K. Xu, X. Ma, and S. Liu, "Learning panorama-aware VLA for mobile manipulation with whole-body teleoperation," 2026. [Online]. Available: https://arxiv.org/abs/2608.02257

[56] Y. Hu, H. Zhu, B. Zheng, Y. Hu, T. Zhang, Z. Chen, J. Zhao, R. Nai, and Y. Gao, "Openhlm: An empirical recipe for whole-body humanoid loco-manipulation," 2026. [Online]. Available: https://arxiv.org/abs/2606.22174

[57] Z. Li, Z. Zhang, Y. Wei, W. Zhang, X. Yuan, P. Zhi, G. Li, X. Guo, F. Gao, J. Yang, and S. Zhang, "ω-0: A latent predictive world action model for concurrent humanoid loco-manipulation," 2026. [Online]. Available: https://arxiv.org/abs/2608.06375

[58] Z. Yuan, T. Wei, L. Gu, P. Hua, T. Liang, Y. Chen, and H. Xu, "HERMES: Human-to-robot embodied learning from multi-source motion data for mobile dexterous manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2508.20085

[59] K. Zakka, A. Zeng, P. Florence, J. Tompson, J. Bohg, and D. Dwibedi, "XIRL: Cross-embodiment inverse reinforcement learning," 2021. [Online]. Available: https://arxiv.org/abs/2106.03911

[60] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak, "Affordances from human videos as a versatile representation for robotics," 2023. [Online]. Available: https://arxiv.org/abs/2304.08488

[61] C. Yuan, C. Wen, T. Zhang, and Y. Gao, "General flow as foundation affordance for scalable robot learning," 2024. [Online]. Available: https://arxiv.org/abs/2401.11439

[62] J. Ren, P. Sundaresan, D. Sadigh, S. Choudhury, and J. Bohg, "Motion tracks: A unified representation for human-robot transfer in few-shot imitation learning," 2025. [Online]. Available: https://arxiv.org/abs/2501.06994

[63] Y. Han, L. Liu, Y. Gu, Y. Lu, H. Wang, S. Wai, I. Myrie, Q. Zhang, J. Yu, and G. Li, "A4A: Cross-embodiment transfer of action-oriented 4D affordances from human demonstrations," 2026. [Online]. Available: https://arxiv.org/abs/2609.05892

[64] 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," 2025. [Online]. Available: https://arxiv.org/abs/2502.16587

[65] H. Li, I. Zhang, R. Ouyang, X. Wang, Z. Zhu, Z. Yang, Z. Zhang, B. Wang, C. Ni, W. Qin *et al.*, "Mimicdreamer: Aligning human and robot demonstrations for scalable VLA training," 2025. [Online]. Available: https://arxiv.org/abs/2509.22199

[66] Y. Wang, P. Lin, X.-H. Chen, H. Yuan, Z. Liang, Y. Huang, A. Chen, Z. Lei, J. Zhang, T. Zhang *et al.*, "Ego2robot: Scalable robot data synthesis from egocentric human data," 2026. [Online]. Available: https://arxiv.org/abs/2608.02580

[67] J. Jeong, S. J. Joo, J. Kang, D. Kim, Y. Kim, H. Kim, and S. J. Kim, "Huro: Robotizing human videos for scalable VLA pretraining," 2026. [Online]. Available: https://arxiv.org/abs/2609.10706

[68] J. P. Araujo, Y. Ze, P. Xu, J. Wu, and C. K. Liu, "Retargeting matters: General motion retargeting for humanoid motion tracking," 2025. [Online]. Available: https://arxiv.org/abs/2510.02252

[69] Z. Chen, C. Kong, C. Zhang, Y. Yang, L. Y. Zhu, S. Kousik, and D. Xu, "Warp: Whole-body retargeting for learning from offline human demonstrations," 2026. [Online]. Available: https://arxiv.org/abs/2606.29940

[70] X. Lin, R. Yang, S. Lian, Z. Shen, B. Yu, C. Wu, H. Liu, Y. Zhang, H. Li, Q. Su *et al.*, "Human-as-humanoid: Enabling zero-shot humanoid learning from ego-exo human videos with human-aligned embodiments," 2026. [Online]. Available: https://arxiv.org/abs/2606.32009

[71] C. Xia and C. Ye, "Morphology-aware human motion retargeting for wheeled-humanoid loco-manipulation," 2026. [Online]. Available: https://arxiv.org/abs/2609.11357

[72] H. Huang, H. Dong, and H. Dong, "Mobile UMI: Cross-view diffusion policy with decoupled kinematics for mobile manipulation," 2026. [Online]. Available: https://arxiv.org/abs/2605.20894

[73] Y. Ye, Y. Fu, Y. Lv, B. Hou, J. Cen, L. Kong, D. Zheng, T. Chen, J. Liu, Z. Cao, Y. Lou, W. Chow, X. Sun, Y. Wang, K. Ge, X. Chi, X. Zhang, Z. Pang, Y. Zhong, S. Han, Z. Lu, W. Yuan, Q. Chen, M. Y. Wang, Y. Mu, Z. Liu, J. Yang, P. Luo, and S. Zhang, "Data pyramid for embodied manipulation: A survey," 2026. [Online]. Available: https://arxiv.org/abs/2607.24744

[74] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik *et al.*, "Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives," 2023. [Online]. Available: https://arxiv.org/abs/2311.18259

[75] L. Ma, Y. Ye, F. Hong, V. Guzov, Y. Jiang, R. Postyeni, L. Pesqueira, A. Gamino *et al.*, "Nymeria: A massive collection of multimodal egocentric daily motion in the wild," 2024. [Online]. Available: https://arxiv.org/abs/2406.09905

[76] P. Banerjee, S. Shkodrani, P. Moulon, S. Hampali, S. Han, F. Zhang, L. Zhang, J. Fountain *et al.*, "HOT3D: Hand and object tracking in 3D from egocentric multi-view videos," 2024. [Online]. Available: https://arxiv.org/abs/2411.19167

[77] R. Hoque, P. Huang, D. J. Yoon, M. Sivapurapu, and J. Zhang, "Egodex: Learning dexterous manipulation from large-scale egocentric video," 2026. [Online]. Available: https://arxiv.org/abs/2505.11709

[78] R. Punamiya, S. Kareer, Z. Liu, J. Citron, R.-Z. Qiu, X. Cai *et al.*, "Egoverse: An egocentric human dataset for robot learning from around the world," 2026. [Online]. Available: https://arxiv.org/abs/2604.07607

[79] Z. Li, B. Yang, C. Miao, K. Zhu, H. Chen, Q. Guan *et al.*, "Open-aoe: An open egocentric manipulation dataset and toolchain for embodied learning," 2026. [Online]. Available: https://arxiv.org/abs/2607.14183

[80] M. Lepert, J. Fang, and J. Bohg, "Phantom: Training robots without robots using only human videos," 2026. [Online]. Available: https://arxiv.org/abs/2503.00779

[81] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu, "Egomimic: Scaling imitation learning via egocentric video," 2024. [Online]. Available: https://arxiv.org/abs/2410.24221

[82] V. Liu, A. Adeniji, H. Zhan, S. Haldar, R. Bhirangi, P. Abbeel, and L. Pinto, "Egozero: Robot learning from smart glasses," 2025. [Online]. Available: https://arxiv.org/abs/2505.20290

[83] M. Xu, H. Zhang, Y. Hou, Z. Xu, L. Fan, M. Veloso, and S. Song, "Dexumi: Using human hand as the universal manipulation interface for dexterous manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2505.21864

[84] Y. Liu, S. Cheng, X. Yin, W. C. Shin, A. Cueva, Y. Yang, Z. Chen, C. Zhang, and D. Xu, "Egoengine: From egocentric human videos to high-fidelity dexterous robot demonstrations," 2026. [Online]. Available: https://arxiv.org/abs/2606.12604

[85] T. Yu, J. Wu, D. Li, B. Chen, H. Liu, X. Ma, and H. Liu, "SEED-UMI: Sharing the exoskeleton between human and robot for one-to-one dexterous demonstration," 2026. [Online]. Available: https://arxiv.org/abs/2609.11753

[86] H. Luo, Y. Feng, W. Zhang, S. Zheng, Y. Wang, H. Yuan, J. Liu, C. Xu, Q. Jin, and Z. Lu, "Being-h0: Vision-language-action pretraining from large-scale human videos," 2025. [Online]. Available: https://arxiv.org/abs/2507.15597

[87] H. Luo, Y. Wang, W. Zhang, S. Zheng, Z. Xi, C. Xu, H. Xu, H. Yuan, C. Zhang, Y. Wang, Y. Feng, and Z. Lu, "Being-h0.5: Scaling human-centric robot learning for cross-embodiment generalization," 2026. [Online]. Available: https://arxiv.org/abs/2601.12993

[88] H. Luo, W. Zhang, Y. Feng, S. Zheng, H. Xu, C. Xu, Z. Xi, Y. Fu, and Z. Lu, "Being-h0.7: A latent world-action model from egocentric videos," 2026. [Online]. Available: https://arxiv.org/abs/2605.00078

[89] X. Cai, R.-Z. Qiu, G. Chen, L. Wei, I. Liu, T. Huang, X. Cheng, and X. Wang, "In-n-on: Scaling egocentric manipulation with in-the-wild and on-task data," 2025. [Online]. Available: https://arxiv.org/abs/2511.15704

[90] B. Li, X. Yin, M. Lin, Y. Zhang, and D. Xu, "Egowam: World action models beyond pixels with in-the-wild egocentric human data," 2026. [Online]. Available: https://arxiv.org/abs/2607.08436

[91] M. Lepert, J. Fang, and J. Bohg, "Masquerade: Learning from in-the-wild human videos using data-editing," 2025. [Online]. Available: https://arxiv.org/abs/2508.09976

[92] Q. Sun, C. Shu, S. Zhou, R. Cheng, Y. Wei, Z. Yu, D. Yang, S. Han, and Y. Chun, "Gsrender: Deduplicated occupancy prediction via weakly supervised 3d gaussian splatting," *arXiv preprint arXiv:2412.14579*, 2024.

[93] Z. Qian, X. Chi, Y. Li, S. Wang, Z. Qin, X. Ju, S. Han, and S. Zhang, "Wristworld: Generating wrist-views via 4d world models for robotic manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2510.07313

[94] J. Xiu, F. Hong, Y. Li, M. Li, W. Wang, S. Han, L. Pan, and Z. Liu, "Egotwin: Dreaming body and view in first person," in *International Conference on Learning Representations*, vol. 2026, 2026, pp. 28 642–28 659.

[95] X. Qi, H. Zhang, Y. Wang, J. Pan, C. Liu, P. Li, X. Chi, M. Li, W. Xue, S. Zhang, W. Luo, Q. Liu, and Y. Guo, "Cocogesture: Toward coherent co-speech 3d gesture generation in the wild," 2025. [Online]. Available: https://arxiv.org/abs/2405.16874

APPENDIX

A. System and Implementation Details

This section provides additional implementation details of the DexRoam data acquisition pipeline. The robot platform, whole-body teleoperation interface, and tracker-free human demonstration system are described in the main paper. Here, we focus on the actual data rates, synchronization procedure, and visual preprocessing used in our implementation.

1) Robot Teleoperation and Data Acquisition: Robot demonstrations are collected using the Astribot–XHand whole-body teleoperation system. The system records egocentric stereo RGB observations, robot joint states, MANUS hand poses, and XHand joint feedback. These data streams run asynchronously at their native rates. In practice, the head stereo RGB stream operates at approximately 30 Hz, the robot joint states at approximately 250 Hz, and both the MANUS hand poses and XHand joint feedback at approximately 39 Hz. These values reflect the actual data rates observed during data collection rather than the target frequencies specified in the software configuration.

During offline preprocessing, we use the timestamps of the head visual stream as the reference timeline and align the remaining higher-frequency or asynchronous state streams to the corresponding visual frames. After synchronization, each training sample contains the visual observation, robot state, and action information associated with the same reference time, resulting in robot demonstrations organized at approximately 30 Hz.

The robot head camera stores stereo RGB observations as side-by-side JPEG images with a resolution of 1280×480. Each frame is split at the horizontal midpoint into left and right views, each with a resolution of 640×480, which are used as the stereo visual inputs to the policy.

2) Human Egocentric Capture System: The tracker-free human demonstration system consists of a Meta Quest 3 and a head-mounted ZED Mini camera. The ZED Mini records stereo RGB observations at approximately 30 Hz, while head and body poses are sampled within the same camera-driven acquisition loop and stored at the same rate. At each iteration, the latest head and body poses are read first, followed immediately by the corresponding ZED image capture, and the pose and stereo image are paired according to their capture order. The visual stream displayed inside the VR headset runs at approximately 20 Hz and is used only for real-time feedback to the demonstrator without affecting the recorded data frequency. Egocentric RGB observations are captured using the ZED Mini in `HD1080` mode at 30 fps and stored as side-by-side stereo images, with each view at approximately 1920×1080 and the full stereo frame at approximately 3840×1080. During preprocessing, each frame is decoded and converted to RGB, split into left and right views, center-cropped to 1280×720 per view, and resized to 640×480. The resulting left and right images are used as the stereo visual inputs for the human demonstrations.

TABLE IV: Detailed replay errors of the PICO + Motion Trackers capture system. Position and rotation errors are reported in millimeters (mm) and degrees (°), respectively.

Component	Metric	Pick Chips Can	Pour Water	Throw Trash	Average
Torso	Position	14.76	11.50	13.22	**13.16**
	Rotation	1.71	2.74	1.74	**2.06**
Left Arm	Position	3.85	4.11	4.41	**4.12**
	Rotation	0.54	0.58	0.60	**0.57**
Right Arm	Position	5.44	5.45	6.02	**5.63**
	Rotation	0.79	0.79	0.85	**0.81**
Head	Rotation	1.27	1.32	1.09	**1.23**

TABLE V: Motion smoothness of retargeted human trajectories after temporal resampling and teleoperated robot demonstrations. Higher LDLJ and SPARC indicate smoother motion.

Task	**Data**	**LDLJ ↑**	**SPARC ↑**
Pick Chips Can	Retargeted Human	-20.2802	-5.8049
	Teleoperated Robot	-20.4726	-7.0548
Pour Water	Retargeted Human	-20.5962	-5.9222
	Teleoperated Robot	-20.4183	-6.5718
Throw Trash	Retargeted Human	-20.8349	-6.6556
	Teleoperated Robot	-20.3022	-6.9229
Deliver Fruit	Retargeted Human	-20.3453	-5.4803
	Teleoperated Robot	-19.8355	-6.2193
Push Chair & Close Laptop	Retargeted Human	-20.5130	-7.2351
	Teleoperated Robot	-20.5575	-8.4923
Average	**Retargeted Human**	**-20.5139**	**-6.2196**
	Teleoperated Robot	**-20.3172**	**-7.0522**

B. Additional Experimental Results and Analysis

1) Detailed Tracker-Based Capture Results: In the main paper, we compare our tracker-free capture system with a tracker-based counterpart using a PICO headset and five body-worn motion trackers, and report the replay errors averaged over 15 trajectories. Here, we provide the detailed results underlying this comparison. For each of the three tasks, we evaluate five captured trajectories and report the positional and rotational replay errors of the torso, two arms, and active head.

2) Motion Smoothness Analysis: We further evaluate whether transferring human motion into the robot action space introduces additional motion jitter. We compare the final retargeted human trajectories with teleoperated robot demonstrations using two standard motion-smoothness metrics: log dimensionless jerk (LDLJ) and spectral arc length (SPARC).

LDLJ measures temporal irregularity through the third-order derivative (jerk) of a scalar trajectory $x(t)$, computed as

$$\mathrm{DLJ}(x) \propto \int_0^T \left(\frac{d^3 x(t)}{dt^3} \right)^2 dt, \tag{1}$$

$$\mathrm{LDLJ}(x) = -\ln \left| \mathrm{DLJ}(x) \right|. \tag{2}$$

where the jerk term is normalized by movement duration and scale to obtain the dimensionless quantity.

SPARC measures smoothness in the frequency domain from the velocity profile. Let $\hat{V}(\omega)$ denote the normalized Fourier

Fig. 10. Representative failure cases across the five tasks. Each row shows one task, with successive frames illustrating a failed rollout from left to right.

magnitude spectrum of the trajectory velocity. SPARC is defined as the negative arc length of the normalized spectrum,

$$\text{SPARC} = -\int_0^{\omega_c} \sqrt{\left(\frac{1}{\omega_c}\right)^2 + \left(\frac{d\hat{V}(\omega)}{d\omega}\right)^2}\, d\omega, \quad (3)$$

where ω_c is determined using a maximum cutoff frequency of 10 Hz and an amplitude threshold of 0.05. A smoother trajectory has a less complex velocity spectrum and therefore a higher SPARC value.

For each episode, LDLJ and SPARC are averaged across all scalar channels of the torso, left arm, right arm, and active head, and the reported values are then averaged over episodes. The retargeted human trajectories exhibit motion smoothness comparable to teleoperated robot demonstrations. Their mean LDLJ is close to that of robot trajectories (-20.51 vs. -20.32), while their mean SPARC is higher (-6.22 vs. -7.05). At the task level, the retargeted trajectories achieve higher SPARC on all five evaluated tasks, while LDLJ remains within a similar range to the robot demonstrations. These results indicate that the human-to-robot transfer pipeline preserves the smooth temporal structure of human motion and does not introduce evident additional jitter into the resulting robot-space trajectories.

C. Failure Mode Analysis

Across all five mobile dexterous tasks, the robot must coordinate the mobile base, dual-arm trajectories, head viewpoint, and dexterous-hand joints simultaneously. Although adding human demonstrations substantially improves overall success rates, several representative failure modes remain during real-robot deployment, as illustrated in Fig. 10. We analyze the main failure causes for each task below.

a) Pick Chips Can: The dominant failure is the dexterous hand failing to form a stable grasp. The robot can usually move close to the target and produce a visually plausible reach-and-close motion; however, in some failures the fingers stop tightening after only light contact with the can, leaving the grasp insufficiently secure. Although the end-effector pose and hand shape are close to those of successful demonstrations, inadequate finger closure prevents the robot from reliably lifting and transporting the can.

b) Pour Water: Two typical failures occur. First, the grasped cup may slip during transport or pouring because a small grasp-position error or insufficient finger closure leaves the grasp unstable. Second, premature finger contact during the approach can displace the lightweight cup from its expected position, causing the subsequent closure motion to miss the

TABLE VI: Composition of the 53-D proprioceptive state.

Component	Representation	Dim.
Left arm	3D Cartesian position + 6D rotation	9
Right arm	3D Cartesian position + 6D rotation	9
Torso	3D Cartesian position + 6D rotation	9
Left XHand	Joint feedback	12
Right XHand	Joint feedback	12
Active head	Joint state	2
Total		**53**

TABLE VII: Composition of the 56-D unified whole-body action.

Component	Representation	Dim.
Left XHand	Joint target	12
Right XHand	Joint target	12
Left arm	Relative translation + 6D rotation	9
Right arm	Relative translation + 6D rotation	9
Torso	Relative translation + 6D rotation	9
Active head	Joint target	2
Mobile base	$(\Delta x, \Delta y, \Delta\theta)$	3
Total		**56**

intended grasp configuration.

c) Throw Trash: Failures mainly arise from timing errors between base motion and object release. Because the bin opening is relatively small, the robot must release the object at an appropriate position after approaching the bin. In some failures, the base moves slightly beyond the desired release position before the action chunk completes, causing the object to miss the bin. This highlights the need for precise coordination among locomotion, arm motion, and release timing.

d) Deliver Fruit: The main failure arises from insufficient bimanual grasp stability during transport. In some rollouts, the two hands do not clamp the basket tightly enough, allowing the basket to oscillate while the robot moves. The resulting motion of the fruit changes the load distribution and shifts the effective center of mass of the basket, which further increases its tilt. These effects can accumulate during locomotion and eventually cause the basket to tip over.

e) Push Chair & Close Laptop: Failures in this long-horizon task often originate from the preceding chair-pushing stage. In some rollouts, the robot does not push the chair sufficiently close to the table, leaving the robot too far from the laptop when transitioning to the closing stage. As a result, the laptop falls outside the effective reaching range of the arm and the robot cannot complete the final closing motion. This illustrates how errors in an early stage of a long-horizon mobile manipulation task can propagate and make subsequent manipulation infeasible.

D. Policy Training Details

This section provides additional details of the unified policy interface and the training configurations used for the two VLA backbones. GR00T N1.7 and $\pi_{0.5}$ use the same stereo visual observations, 53-D proprioceptive state representation, and 56-D whole-body action space.

TABLE VIII: Training configuration for $\pi_{0.5}$.

Hyperparameter	Value
Base model	$\pi_{0.5}$
Number of GPUs	2
Maximum training steps	30,000
Global batch size	4
Per-GPU batch size	2
Gradient accumulation steps	1
Optimizer	AdamW
Learning rate	2.5×10^{-5}
Weight decay	1×10^{-10}
Warmup ratio	0.033
Warmup steps	1,000

TABLE IX: Training configuration for GR00T N1.7.

Hyperparameter	Value
Base model	NVIDIA Isaac-GR00T N1.7-3B
Number of GPUs	4
Maximum training steps	40,000
Global batch size	24
Per-GPU batch size	6
Gradient accumulation steps	1
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	1×10^{-5}
Warmup ratio	0.05
Warmup steps	2,000

1) Unified Policy Interface: The policy input at time step t is represented as

$$o_t = \left(I_t^L, I_t^R, p_t, \ell\right), \tag{4}$$

where I_t^L and I_t^R denote the left and right egocentric RGB observations, respectively, p_t denotes the robot proprioceptive state, and ℓ is the task language instruction. Both stereo views are represented at a resolution of 640×480. The current implementation uses a single-frame proprioceptive state without explicitly stacking multiple state histories.

a) Unified Policy State and Action Space: The robot proprioceptive state is represented as $p_t \in \mathbb{R}^{53}$, and the whole-body action is represented as $a_t \in \mathbb{R}^{56}$. Their compositions are summarized in Tables VI and VII.

The Cartesian states of the two arms and torso are represented by 3D positions and continuous 6D rotations. The policy state does not explicitly include the chassis pose. For actions, the two arms and torso use relative translation and 6D rotation, the mobile base uses local planar motion $(\Delta x, \Delta y, \Delta\theta)$, and the active head and two XHands use robot joint-space targets.

2) $\pi_{0.5}$ Implementation: For $\pi_{0.5}$, each training sample consists of the left and right RGB observations, the 53-D proprioceptive state, the task language instruction, and the corresponding 56-D whole-body action targets. We initialize the model from the pretrained $\pi_{0.5}$ base checkpoint. Since the

pretrained action interface does not match the 56-D DexRoam action space, the action input and output projections are randomly initialized. All model parameters are trainable during finetuning, with no frozen modules. The complete training configuration is summarized in Table VIII.

3) GR00T N1.7 Implementation: For GR00T, we use NVIDIA Isaac-GR00T N1.7-3B as the base model. The model takes stereo egocentric RGB observations, the 53-D proprioceptive state, and the task language instruction, and predicts actions in the same 56-D whole-body action space. During training, the pretrained LLM and vision backbones are frozen, while the projector and diffusion action decoder are optimized to adapt the pretrained model to the DexRoam action interface. The complete training configuration is summarized in Table IX.

E. Detailed Evaluation Protocol and Metrics

This section provides the detailed evaluation protocol for task progress and success, as well as the implementation details of the human–robot action distribution metrics used in the main paper.

1) Human–Robot Action Distribution Analysis: We analyze the distribution gap between human-derived and real-robot actions over the complete 56-D whole-body action space, including the two 12-D dexterous-hand actions, two 9-D arm actions, 9-D torso action, 2-D head action, and 3-D mobile-base action.

a) Action normalization: Before computing MMD, SWD, and PCA, each action dimension is independently normalized using pooled robust statistics. For the j-th action dimension, we use

$$\tilde{a}_j = \frac{a_j - \mathrm{median}(a_j)}{\max\left((Q_{75,j} - Q_{25,j})/1.349,\, 10^{-3}\right)}, \tag{5}$$

where $Q_{75,j}$ and $Q_{25,j}$ denote the 75th and 25th percentiles, respectively. For each episode, we randomly sample up to 256 frames to estimate the pooled statistics. Each data source is further limited to at most 4,000 frames, and all participating sources are downsampled to the same number of samples before pooling, such that each domain contributes equally. Human and robot actions are normalized in the same coordinate system.

b) Maximum Mean Discrepancy: We compute Maximum Mean Discrepancy (MMD) on the normalized 56-D action vectors using a single Gaussian RBF kernel. Given equally sized human and robot action sets $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, respectively, we use the biased V-statistic estimator

$$\mathrm{MMD} = \left[\max\left(\frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i, y_j) - \frac{2}{n^2}\sum_{i,j=1}^{n} k(x_i, y_j),\, 0\right)\right]^{1/2}. \tag{6}$$

where the Gaussian kernel is

$$k(x, y) = \exp\left(-\frac{\|x - y\|_2^2}{2b}\right). \tag{7}$$

The bandwidth parameter b is set to the median pairwise squared Euclidean distance over the pooled normalized samples, corresponding to $\sigma^2 = b$ under the standard Gaussian-kernel notation. The bandwidth is estimated separately for the absolute and relative action spaces, while the raw and aligned relative distributions share the same bandwidth. For each comparison, we sample the same number of human and robot actions without replacement, with at most 4,000 samples from each domain.

c) Sliced Wasserstein Distance: We use the first-order Sliced Wasserstein Distance (W_1) on the normalized 56-D action space. We sample $L = 128$ random projection directions from a standard multivariate Gaussian distribution and normalize each direction to unit length. For each direction, the projected human and robot samples are sorted, and the absolute differences between corresponding ordered samples are averaged. Specifically,

$$\mathrm{SWD} = \frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{n}\sum_{i=1}^{n}\left|x_{(i)}^{(\ell)} - y_{(i)}^{(\ell)}\right|, \tag{8}$$

where $x_{(i)}^{(\ell)}$ and $y_{(i)}^{(\ell)}$ denote the i-th ordered samples after projection onto direction ℓ. Human and robot samples are balanced and limited to at most 4,000 actions per domain. Random projection directions are independently sampled for different alignment stages.

d) PCA visualization: PCA is performed on the normalized 56-D action vectors using the same robust normalization described above. To avoid imbalance among data sources, samples are balanced across domains, with at most 6,000 samples used to fit each PCA model. The first two principal components are retained for visualization.

2) Task Completion Score: We use Task Completion Score (TCS) to characterize partial progress during long-horizon mobile manipulation. Each task is decomposed into a sequence of ordered milestones with task-specific weights. The TCS of a rollout is computed as the sum of the weights of all completed milestones, with the milestone weights for each task summing to 1.0. Detailed scoring criteria are provided in Table X.

TABLE X: Task Completion Score (TCS) criteria. Each task is decomposed into five ordered milestones with task-specific weights shown in parentheses.

Task	**Stage 1**	**Stage 2**	**Stage 3**	**Stage 4**	**Stage 5**
Pick Chips Can	Move forward to the table (0.2)	Move the hand into the grasping envelope without knocking over the can, with the fingertip center within 3 cm of the can axis (0.2)	Close the fingers and establish a stable grasp with at least two opposing fingers in contact with the can (0.2)	Lift the can above the table by at least the basket height and maintain the grasp for at least 3 s without slipping (0.2)	Successfully place the can into the basket (0.2)
Pour Water	Reach toward the container (0.1)	Establish a stable grasp on the container (0.2)	Lift the container at least 5 cm above the table (0.2)	Move toward the target while maintaining a stable grasp on the container (0.2)	Rotate the wrist and pour water into the target container (0.3)
Throw Trash	Move the hand into the grasping envelope of the chips can without knocking it over, with the fingertip center within 3 cm of the can axis (0.2)	Close the fingers and establish a stable grasp with at least two opposing fingers in contact with the object (0.2)	Move the mobile base until the trash bin is within the reachable workspace (0.2)	Release the object such that it falls into the trash bin (0.3)	Complete the task without displacing or knocking over the trash bin (0.1)
Deliver Fruit	Move both hands into the grasping envelope and establish contact with the fruit basket (0.2)	Lift the fruit basket at least 5 cm above the supporting surface using coordinated bimanual grasping (0.2)	Maintain the lifted basket for at least 3 s without dropping it (0.2)	Complete the required locomotion and turning while keeping the basket and its contents stable (0.2)	Place the fruit basket at the target location (0.2)
Push Chair & Close Laptop	Establish contact between the hand and the back of the chair (0.2)	Successfully push the chair without losing hand–chair contact (0.2)	Move the chair to the specified target position under the table (0.2)	Extend the right hand and successfully reach behind the laptop lid (0.2)	Successfully close the laptop (0.2)