Praxis: Scaling One-Shot Human Demonstration to Generalist Policy for Whole-Body Manipulation

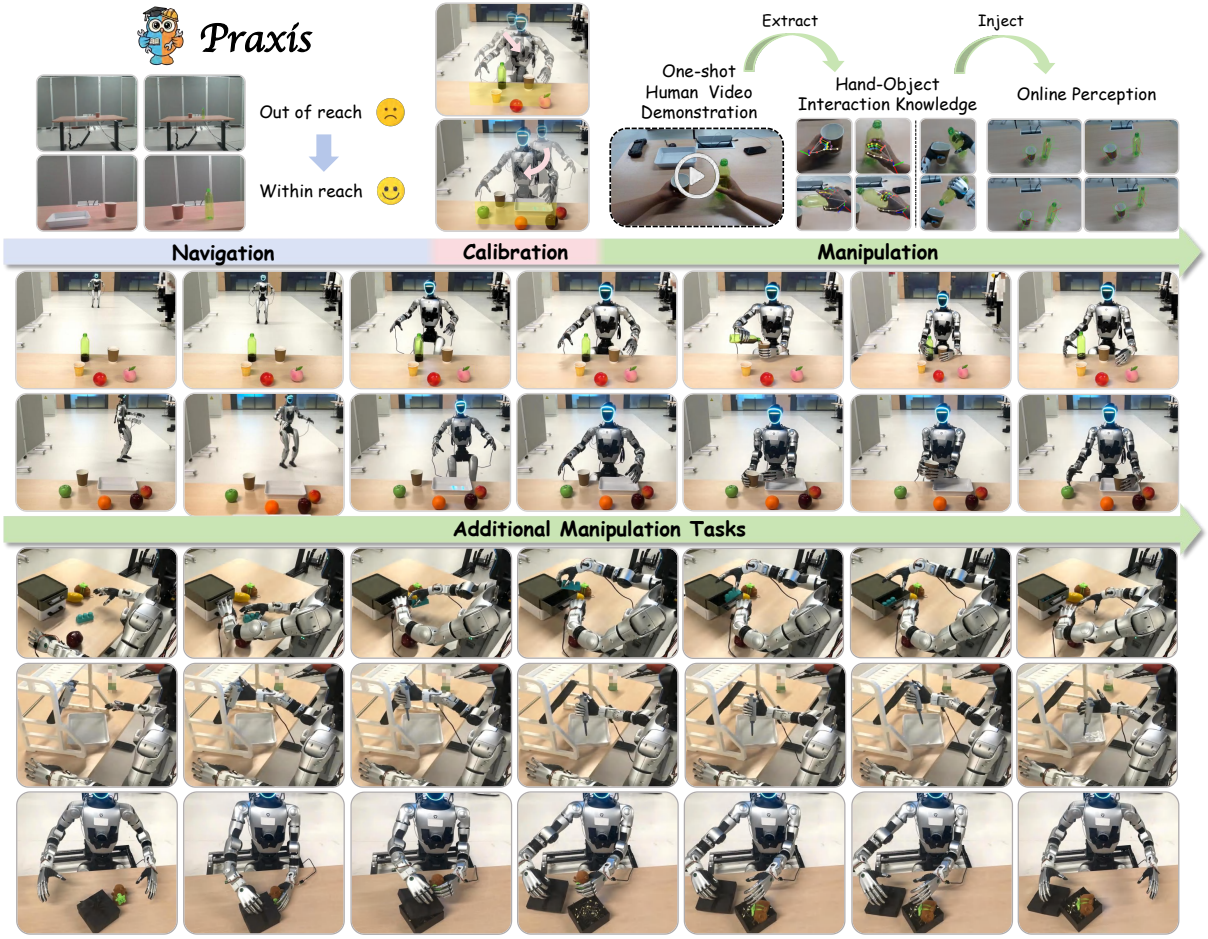

Fig. 1: Praxis is a three-stage whole-body manipulation framework that learns generalist policies from one-shot human videos. Top to bottom: `Pour Water`, `Hand Over`, `Pull Drawer`, `Manipulate Pipette`, `Open Lid`.

***Abstract*—Recent advancements in humanoid Whole-Body Control (WBC) have demonstrated promising performance in generating expressive behaviors like dancing and athletic movements. However, scaling WBC policies to achieve human-level dexterity in manipulating diverse objects within dynamic environments remains a fundamental challenge. To bridge this gap, we introduce Praxis, a generalist WBC framework for dexterous humanoid manipulation, driven by action-level semantic information from one-shot human video demonstrations. Specifically, Praxis disentangles WBC into three coordinated phases: 1) Navigation, where a vision-language-navigation model guides the robot from various initial positions to approach target objects; 2) Calibration, which optimizes whole-body posture to align the end-effectors within a functional workspace; and 3) Manipulation, where the robot physically interacts with objects to accomplish the specified tasks. Most critically, once trained, Praxis can be efficiently deployed to novel tasks in a learning-free manner, requiring only a single video demonstration from a human. We conduct extensive empirical studies to validate the spatial, visual, and cross-object generalizability of Praxis over diverse tasks and environments, and demonstrate its robustness in dynamically recovering from physical interferences at all stages of operation.**

I. Introduction

Developing general-purpose robotic manipulation policies is a foundational goal of embodied intelligence [11]. While recent studies have demonstrated impressive dexterous manipulation across a variety of objects, tasks, environments, and embodiments [28, 3, 34, 22, 2, 51], current success is largely confined to limited, static workspaces of fixed-based robots. Scaling underlying dexterity and generalization to larger, open-world, and dynamically changing environments in the wild remains a significant challenge [37].

To overcome limitations in workspace, recent studies have explored mobile manipulation [15, 58, 6, 35] that enables a

robot to locate and navigate to target objects before interacting with them. However, existing approaches often decouple mobility from manipulation, leading to a loss of the smooth coordination in human behaviors. For example, when humans retrieve an object, such as picking an item from the floor, they do not simply move to the location and then extend their arm. Instead, they seamlessly coordinate locomotion, torso motion, and reaching to maintain balance and optimize manipulability.

The development of Whole-Body Control (WBC) algorithms offers a pathway to such human-level coordination [19]. Following this trend, recent studies [21, 9, 70, 33, 56, 1] have introduced learning-based WBC models, typically by training control policies in simulated environments using reinforcement learning (RL) algorithms and transferring these policies to real-world systems for execution and demonstration. While learning-based WBC succeeds in generating expressive motions for dancing or athletic movements [9, 14, 59], extending these approaches to dexterous manipulation, requiring rich contact and precise interaction, remains an open problem.

To this end, our work aims to scale WBC toward generalizable, functional manipulation tasks. We argue that realizing true whole-body dexterity requires a system that can extend its reach through navigation to locate distant objects, refine to a manipulation-ready posture, and execute complex interactions, such as dexterous tool use and articulated manipulation. This process must be fully automatic without dependence on manual commands [1], teleoperation [31, 10], or explicit motion references [61]. Drawing inspiration from the emergence of human-to-robot transfer [26], we introduce **Praxis**, a hierarchical framework that extracts task semantics and motion intent from a single human video demonstration and transfers these skills to humanoid robots. Unlike language commands, video demonstrations provide the dense semantic information necessary for capturing the nuance of dexterous interaction [69]. By leveraging one-shot human demonstrations, Praxis bypasses the data scarcity bottlenecks inherent in training complex humanoid policies from scratch.

To implement Praxis, we decompose the complex problem of long-horizon whole-body manipulation into three coordinated stages (see Fig. 1): 1) *Navigation*, where a vision-language navigation model processes continuous visual observations to guide the robot from various locations to the target vicinity; 2) *Posture calibration*, which refines the humanoid's posture through closed-loop optimization, aligning the arm-hand workspace with target objects to maximize manipulability; and 3) *Manipulation*, where the robot executes dexterous, key-pose-grounded motions extracted from one-shot human video demonstrations, while simultaneously maintaining balance through coordinated lower-body control.

Through this design, Praxis achieves strong generalizability across variations in perception, dynamics, and task contexts. By decoupling invariant interaction semantics from environment-specific parameters and continuously regrounding control via visual and tactile feedback, our policy flexibly adapts to novel object poses, scene layouts, and physical properties without retraining or additional demonstrations. This integration of perception-conditioned WBC and one-shot semantic transfer allows Praxis to robustly transfer scalable human-demonstrated manipulation skills, diverse scenarios, enabling reliable whole-body manipulation under real-world dynamics.

We conduct extensive empirical studies across five long-horizon manipulation tasks. While all five are evaluated within a general manipulation study, we further stratify them into two categories: two tasks focus on long-distance whole body manipulation, and the remaining three represent challenging, contact-rich long-horizon tasks. Our framework achieves an average success rate of 76.97% despite drastic environmental variations, substantially outperforming four large-scale VLA baselines, the best of which attains only 10.30%. Notably, in long-distance scenarios, Praxis maintains robust performance, achieving success rates of 75% at 3 m and 50% at 5 m ranges. We further demonstrate system resilience under dynamic, external physical disturbances, recording success rates of 70%, 90%, and 90% over three operational stages, respectively. Details of the hardware system are described in Appendix A.

II. Related Works

Humanoid Whole-Body Manipulation. It couples locomotion, postural stability, and dexterous end-effector control to perform contact-rich object interactions. Prior work largely follows two paradigms: structured frameworks such as WoCoCo [64] and FALCON [68], which decompose whole-body control into sequential contact primitives or decoupled sub-agent architectures; and end-to-end systems [62, 14, 20, 31, 53, 13], which learn unified policies from imitation learning via teleoperation or from RL. While effective, these approaches are often data-intensive and require extensive retraining or manual engineering to adapt to novel task geometries. In contrast, Praxis extracts interaction knowledge from a single demonstration. It uses perception-driven regrounding to align trajectories with real-time feedback, enabling strong generalization and robust recovery in real-world settings.

Learn from Human Hand Videos. Human hand videos provide a scalable source of supervision for learning complex manipulation behaviors [17, 12, 63, 36, 18]. To mitigate the embodiment gap between humans and robots, prior work leverages intermediate representations, such as keypoints [41, 16, 55], affordances [30, 65, 39], and geometric correspondences [42, 29, 67], to enable motion retargeting [49, 32] and downstream action planning [27, 7, 65]. Closest to our work are YOTO [69] and OKAMI [32]. While YOTO demonstrates bimanual learning from binocular video, it is limited by implicit depth, a fixed base, and low-DoF grippers; in contrast, we leverage RGB-D sensing for mobile, high-DoF dexterous retargeting. OKAMI is restricted to stationary upper-body control via continuous trajectory warping; our framework integrates mobile navigation. It extracts discrete keyframes via geometric hand parsing to preserve interaction geometry. Furthermore, we enhance physical robustness through tactile-regulated grasping and online interference recovery, allowing adaptation to varying object stiffness and execution deviations.

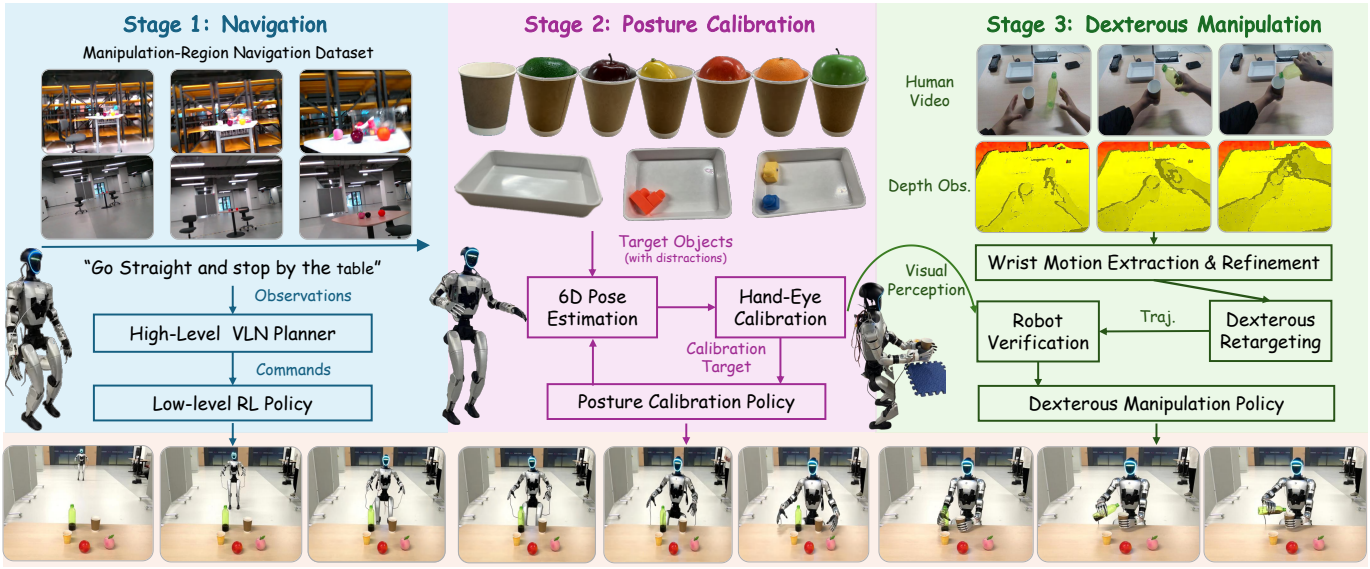

Fig. 2: The architecture of Praxis. Stage 1: Navigation to the manipulation region. Stage 2: Calibration to a manipulation-ready posture in a closed-loop manner. Stage 3: Learning hand-object interaction knowledge through human videos and applying online visual perception to adapt manipulation policies.

III. Method

Based upon the hardware system (Appendix A), we propose a hierarchical whole-body control (WBC) framework, Praxis, for humanoid manipulation in unstructured environments. The framework organizes whole-body manipulation into three sequential yet tightly coupled stages: navigation, posture calibration, and manipulation, all operating under a unified balance controller. In the following, we detail each stage in turn, beginning with navigation, which brings the robot from random initial configurations into the vicinity of the target objects, thereby establishing a feasible starting point for subsequent posture calibration and whole-body manipulation.

A. Humanoid WBC for Navigation

In real-world environments, target objects are often located beyond a humanoid robot's immediate manipulation workspace, making dexterous manipulation infeasible without prior repositioning. Accordingly, the navigation stage should reduce the spatial gap between the robot and the target objects by guiding the robot into their proximity, while leaving fine-grained posture optimization to the subsequent calibration stage. Inspired by recent advances in generalist navigation models [57], our navigation module follows a dual-system design, where a high-level planner guides a low-level whole-body controller for coordinated humanoid locomotion.

The *high-level* planner is a vision-language navigation (VLN) model, denoted as $\pi^{\text{up}}(\boldsymbol{c}_t | o_{t-H}, \ldots, o_t, l_t)$, which maps a sequence of historical visual observations $[o_{t-H}, \ldots, o_t]$ and language instructions l_t to discrete navigation commands $\boldsymbol{c}_t \in \Omega$, where Ω is a finite set of symbolic navigation actions. Prior VLN-based methods, such as NaVILA [8], operates on a restricted action space $\Omega_{\text{base}} = \{\text{move forward}, \text{turn left}, \text{turn right}, \text{stop}\}$. This action set proves inefficient in cluttered workspaces, where even simple lateral alignment necessitates multi-step 'turn-move-turn' sequences. This results in unnecessarily long and brittle trajectories. To address this limitation and exploit the whole-body mobility of our platform, we expand the action space to $\Omega = \Omega_{\text{base}} \cup \{\text{move left}, \text{move right}\}$, enabling direct lateral translation. This augmentation substantially simplifies navigation trajectories and improves positioning accuracy by eliminating redundant rotational maneuvers. To adapt the VLN model to the expanded action space, we fine-tune NaVILA on a heterogeneous dataset constructed from standardized navigation templates to achieve manipulation-region-oriented navigation. To promote robust generalization, we employ a multi-faceted randomization strategy during data collection. Further details are provided in Appendix C. By fine-tuning on these diverse template-based episodes, the high-level policy π^{up} learns to navigate and position the robot reliably in previously unseen environments, while maintaining low trajectory complexity and strong language grounding.

The *low-level* whole-body locomotion controller adopts HOMIE [69], which is trained using Proximal Policy Optimization [48] with an upper-body pose curriculum to improve balance robustness and adaptability. We represent this low-level controller as a control policy $\pi^{\text{low}}(a_t \mid q_{t-H}, \ldots, q_t, \boldsymbol{c}_t, a_{t-1})$. It maps the recent proprioceptive history and control commands to the corresponding motor actions. Here, q_t denotes the robot's proprioceptive state at timestep t, encompassing body angular velocity, joint positions, and joint velocities, while $\boldsymbol{c}_t$ represents the high-level navigation command. The same low-level controller is shared across posture calibration and manipulation to maintain whole-body stability.

B. Posture Calibration for Pre-Manipulation

Although the navigation policy brings the robot close to the target objects, it does not guarantee a posture suitable for dexterous, contact-rich tasks. We define a manipulation-ready

posture as one where the target object lies in the center of a 3D workspace region that yields low inverse kinematics (IK) error for the arm end-effector (EE).

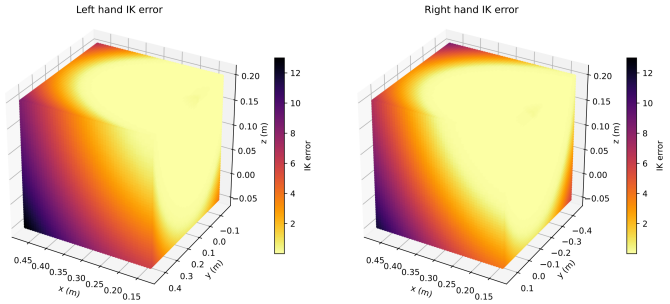

Fig. 3: Left and right arm EE IK error when the torso is fixed.

To address this, we introduce a lightweight closed-loop posture calibration module that reactively refines the robot base pose using object pose feedback. At each time step, the module reads object poses in the robot frame and extracts planar positions $\{(x_o, y_o, z_o)\}_{o\in O}$. If multiple relevant objects are present, the forward displacement is defined as, $x \triangleq \arg\max_{o\in O} |x_o|$, selecting the object with the largest absolute forward offset, while the lateral and vertical displacement is set to the midpoint, $(y, z) \triangleq \frac{1}{|O|}\sum_{o\in O}(y_o, z_o)$.

Our module first adjusts the robot's height to eliminate the vertical deviation $(z_{\text{ref}} - z)$, then performs closed-loop regulation by driving the object displacement (x, y) toward a reference $(x_{\text{ref}}, y_{\text{ref}})$, generating planar velocity commands (v_x, v_y) as bounded functions of the displacement errors and their temporal accumulation. z_{ref} is set to the average height of the target objects in the demonstrations. Fig. 3 visualizes the IK error distributions for both arms's EE; accordingly we set $(x_{\text{ref}}, y_{\text{ref}}) = (0.4, 0)$ where 0.4 (meter) is the farthest forward position with tolerant IK error and 0 lies at the center of low-error lateral region. To ensure control stability, we apply deadbands to suppress unnecessary corrective actions when the system is near the target, defined as $|x - x_{\text{ref}}| \le \varepsilon_x$, $|y - y_{\text{ref}}| \le \varepsilon_y$; and saturate command magnitudes to prevent over-actuation $v_i \leftarrow \text{clip}(v_i, -v_i^{\max}, v_i^{\max})$. The resulting commands are issued at a fixed rate and continuously updated until the robot posture establishes a low-IK-error region for the arm–hand workspace, yielding a manipulation-ready posture.

C. Wrist and Dexterous Hand Motion Extraction

Once the robot reaches a manipulation-ready whole-body posture, the problem reduces to generating precise arm and hand motions to accomplish the task. Human videos are easy to obtain and capture diverse real-world situations [26]. We therefore leverage one-shot human demonstrations, extracting wrist trajectories and dexterous finger motions to directly shape robot actions. By scaling human teaching data, either collected in-house or contributed by others, we can generate correspondingly scalable robot-executable action sequences.

We first collect human-taught demonstrations of long-horizon bimanual dexterous tasks within a workspace compatible with the G1 humanoid's dual arms and hands, with demonstrations recorded using the head-embedded D435i camera. When the robot torso obstructs free bimanual motion for a single demonstrator, the task can also be performed collaboratively by two operators, where one uses the left arm and hand and the other uses the right arm and hand to jointly complete a single demonstration. Moreover, demonstration videos recorded by other G1's D435i cameras can be directly reused, as they share identical eye-to-hand calibration. This allows demonstration motions to be transformed into a common robot-centric representation, which substantially *scales up* the amount of available demonstration data. We then apply vision-based techniques to extract rich manipulation features from the recorded videos. These features are post-processed into keyframe-based motion representations, including 6-DoF EE poses and hand motions which directly drive the coordinated control of the robot's dual arms and hands.

Human Wrist Motion Extraction. We begin with a single-view RGB-D video demonstration, represented as an image sequence $\{I_t\}_{t=1}^{T}$, of a human performing a bimanual manipulation task. For each frame I_t, we use a MANO-based hand reconstruction model [44] to estimate the 3D hand meshes $\mathcal{H}_j^{\diamond}$ for left ($\diamond = L$) and right ($\diamond = R$) hands. The wrist *rotation* is obtained from the MANO [47] global orientation in axis-angle (Rodrigues) form and converted to a rotation matrix $h_{r,t}^{\diamond} \in \text{SO}(3)$, while the wrist 3D *translation* $h_{p,t}^{\diamond}$ is recovered by back-projecting its 2D pixel (u, v) with depth d into the camera coordinate frame using the intrinsics. Treating the wrist as the robot arm EE, the resulting wrist pose in the camera frame is transformed to the robot frame via the eye-to-hand calibration matrix such that $a_{r,t}^{\diamond} = R_{\text{cam}}^{\text{robot}} h_{r,t}^{\diamond}$, $a_{p,t}^{\diamond} = T_{\text{cam}}^{\text{robot}} h_{p,t}^{\diamond}$, yielding the left/right arm EE trajectory $\{a_{r,t}^{\diamond}, a_{p,t}^{\diamond}\}_{t=1}^{T}$.

Keyframe-Based Motion Refinement. Continuous EE trajectories are often noisy and redundant, a problem that becomes more pronounced in long-horizon manipulation tasks. Following prior work [23, 50, 69], we abstract each continuous demonstration into a sequence of discrete keyframes (keyposes) that capture salient intermediate EE configurations. Specifically, we reduce each demonstration into a subset of discrete keyframes, $\{a_{r,k}^{\diamond}, a_{p,k}^{\diamond}\}_{k=1}^{K}, K \ll T$. Keyframes are selected at moments of significant motion change, typically when (i) a dexterous hand is interacting with the object, or (ii) the EE twist:

$$\begin{bmatrix} \mathbf{v}_t^{\diamond} \\ \boldsymbol{\omega}_t^{\diamond} \end{bmatrix} \approx \frac{1}{\Delta t}\left[\log\left(\begin{bmatrix} a_{r,t-1}^{\diamond} & a_{p,t-1}^{\diamond} \\ \mathbf{0}^{\top} & 1 \end{bmatrix}^{-1}\begin{bmatrix} a_{r,t}^{\diamond} & a_{p,t}^{\diamond} \\ \mathbf{0}^{\top} & 1 \end{bmatrix}\right)\right]^{\vee} \in \mathbb{R}^6$$

attains a local extremum, where $\log(\cdot) : \text{SE}(3) \to se(3)$ denotes the Lie group logarithm , and $(\cdot)^{\vee} : se(3) \to \mathbb{R}^6$ maps the resulting algebra element to a 6D twist vector of linear and angular velocities.

Dexterous Retargeting. While arm EE trajectories define the coarse motion of the manipulation, the corresponding hand poses, which encode fine-grained dexterous articulation, are extracted from the K keyframes via a dexterous retargeting module and mapped to robot hand motor commands for physically interactive manipulation. Since the MANO model relies on parametric pose fitting, which can introduce temporal

instability and fitting noise that adversely affect downstream control [52], we adopt MediaPipe [66] here as a monocular vision front-end. For each frame, MediaPipe extracts 21 hand landmarks per hand, based on which we geometrically infer per-joint kinematics, yielding an 11-DoF hand joint-angle vector (in degrees), $h^{\diamond}_{m,k}{}^{(11)} = \left[h^{\diamond}_{m,j,k}\right]_{j\in\mathcal{J}_{11}}$, which covers five fingers' proximal and distal joints as well as the thumb metacarpal joint. Following the Revo2 URDF, we map the 11-DoF hand pose to a 6-DoF controllable vector, $h^{\diamond}_{m,k}{}^{(6)} = \left[h^{\diamond}_{m,j,k}\right]_{j\in\mathcal{J}_6}$, with 5 distal joints recovered through predefined mimic relations on respective proximal joints. Finally, we convert the 6-DoF hand pose into low-level motor commands by normalizing each controlled joint j with a joint-specific maximum angle $h^{\max}_{m,j}$ and remapping to the integer joint motor range $\{0, 1, \cdots, 1000\}$: $\hat{h}^{\diamond}_{m,j,k} = \mathrm{clip}\Big(\frac{h^{\diamond}_{m,j,k}}{h^{\max}_{m,j}}, 0, 1\Big)$, $\tilde{h}^{\diamond}_{m,j,k} = \Big\lfloor 1000\,\hat{h}^{\diamond}_{m,j,k} \Big\rfloor$. Let $a^{\diamond}_{m,k} = [\tilde{h}^{\diamond}_{m,j,k}]_{j\in\mathcal{J}_6}$. This produces the joint motor command trajectory $\{a^{\diamond}_{m,k}\}_{k=1}^{K}$ for streaming dexterous control for each arm EE and hand. Appendix D provides illustrative examples.

Robot-Space Verification. Despite being expressed in the robot coordinate frame, the resulting trajectories are not guaranteed to be directly executable. In particular, some target poses may be kinematically unreachable, independently planned arm motions can induce self-collisions, and fingers within the same hand may collide during aggressive closures (e.g., clenching a fist). Rather than uncovering such failures through on-robot replay, we validate each keyframe offline using IK and hand-level control constraints to ensure reachability and collision-free execution. Valid keyframes are retained in the set of verified keyframe-based robot actions: $A^{\diamond} = \{(a^{\diamond}_{r,k}, a^{\diamond}_{p,k}, a^{\diamond}_{m,k})\}_{k=1}^{K}$, where K is the number of verified robot actions. Building on this, we factor the bimanual sequence into per-arm EE and hand trajectories:

$$A^{L} = \{(a^{L}_{r,k}, a^{L}_{p,k}, a^{L}_{m,k})\}_{k=1}^{K},\ A^{R} = \{(a^{R}_{r,k}, a^{R}_{p,k}, a^{R}_{m,k})\}_{k=1}^{K}.$$

D. One-shot Policy Generalization from Video

Although a one-shot demonstration contains only a single reference trajectory, it captures invariant, task-critical knowledge of robot–object interaction, including EE approach geometry and fine-grained, object-centric hand–object contact patterns. We keep this interaction logic fixed, while adapting the pre-interaction robot motion through online perceptual and tactile feedback to align with novel object poses, scene layouts, and material properties such as object stiffness, and to recover from execution failures. In this way, a single demonstration is lifted into a family of valid execution trajectories that differ in approach, alignment, and correction behavior but share the same interaction semantics, enabling robust generalization far beyond one-shot imitation.

Generalization across Object Poses. Spatial invariance is the ability to generalize task execution under varying object poses. Our method explicitly supports this by conditioning actions on online perception re-grounding rather than fixed demonstration trajectories. Specifically, we perform multi-object pose estimation using FoundationPose++ [60], which extends FoundationPose [54] with a 2D tracker and Kalman filtering for temporally consistent 6D pose estimates in dynamic scenes. To generate the required input masks, we employ LISA [45], which leverages a <SEG> token to integrate segmentation directly into the LLM's reasoning, enabling high-fidelity masking from implicit instructions where standard open-vocabulary methods fail (e.g., pipette). Let step k denote a keyframe where hand-object interaction occurs. Let step k denote a keyframe where hand-object interaction occurs; $T^{\diamond}_{o_i,k} \in \mathrm{SE}(3)$ and $T^{\diamond}_{a,k} = (a^{\diamond}_{r,k}, a^{\diamond}_{p,k}) \in \mathrm{SE}(3)$ be the pose of object o_i and the corresponding EE pose for $\diamond$ hand at step k in the demonstration, both expressed in the robot frame. We define the object-relative transformation, $\Delta T^{\diamond}_{k} = (T^{\diamond}_{o_i,k})^{-1} T^{\diamond}_{a,k}$, which expresses the EE pose in the object coordinate frame and encodes the demonstrated object-centric interaction geometry, invariant to the object's absolute pose. Given a newly estimated pose $\hat{T}^{\diamond}_{o_i,k}$ of the same object in a novel rollout with varied object pose, we recover the adapted EE pose via $\hat{T}^{\diamond}_{a,k} = \hat{T}^{\diamond}_{o_i,k}\,\Delta T^{\diamond}_{k} = (\hat{a}^{\diamond}_{r,k}, \hat{a}^{\diamond}_{p,k})$. Replacing $(a^{\diamond}_{r,k}, a^{\diamond}_{p,k})$ in the original EE trajectory $A^{\diamond}$ with $(\hat{a}^{\diamond}_{r,k}, \hat{a}^{\diamond}_{p,k})$ yields a perception-conditioned EE trajectory $\hat{A}^{\diamond}$ that preserves the demonstrated object-relative motion while adapting automatically to object pose variations.

Notably, the manipulated object can be replaced with another of similar shape and scale, enabling cross-object generalization. In addition, our method supports spatial edits for objects with multiple affordances, allowing the grasp pose to be adjusted accordingly (e.g., selecting different handles or compartments of a drawer by modifying z in $\hat{a}^{\diamond}_{p,k}$).

Generalization across Scene Variations. Beyond object-level variations, Praxis also generalizes across changes in scene configurations. Leveraging the left–right symmetry of the G1 humanoid, a demonstrated EE trajectory $A^{\diamond}$ can be reflected across the sagittal plane to generate a mirrored counterpart, enabling reuse of the same demonstration under symmetric task layouts. For instance, a demonstration of "grasp the bottle with the *right* hand" can be directly transformed into "grasp the bottle with the *left* hand", without additional data collection. Moreover, by leveraging the robustness of FoundationPose++ [60] to variations in lighting, scene layout, and surface appearance (e.g., tables with different shapes and textures), Praxis inherently generalizes across diverse scene configurations, enabling reliable perception-conditioned execution under substantial environmental variation.

Generalization across Object Stiffness. Vision alone is insufficient for physical interaction, as critical properties like stiffness often vary across visually similar materials. We therefore incorporate tactile sensing to ensure robust grasping, modeling the procedure as feedforward closure followed by tactile force regulation. Each finger i tracks contact force $\mathbf{f}_i = [f_{n,i}, \mathbf{f}_{t,i}]$. Contact is detected via $c_i = \mathbb{I}(f_{n,i} > \tau_n)$ and slip risk via $\rho_i = \|\mathbf{f}_{t,i}\|/f_{n,i}$. Upon stable multi-finger contact ($\sum c_i \geq M$), we estimate stiffness $k = \mathrm{mean}(\Delta f_{n,i}/\Delta x_{n,i})$. The target normal force is then computed as $f_n^{\star} = \mathrm{clip}(\alpha k + \beta, f_{\min}, f_{\max})$, subject to the no-slip

constraint $f^{\star}_{n,i} \geq \|\mathbf{f}_{t,i}\|/(\mu_i - \delta)$. We provide further details and visualizations in Appendix F.

Generalization to Failure Cases. While the above visual and tactile perception promotes stable manipulation under nominal conditions, real-world bimanual manipulation inevitably encounters execution failures. Failures in one arm, such as slippage or misalignment, can immediately perturb the coupled hand–object system, necessitating coordinated corrective actions from both arms to prevent task failure. To cope with such scenarios, we formalize self-recovery using a task-specific success indicator, which evaluates task completion based on the current perception and system state: $\psi_t = \mathbb{I}(\mathcal{C}(\Phi_t)) \in \{0, 1\}$, where Φ_t denotes object poses, EE poses, hand motions at time t; $\mathcal{C}(\cdot)$ specifies a task-dependent success condition. For example, in the "Pick cup and put it in tray" task, $\mathcal{C}(\Phi_t)$ encodes relative pose alignment between the cup and the tray. If the success condition remains unsatisfied ($\psi_t = 0$) after the manipulation attempt, updated perception is invoked, and a new arm EE and hand motion trajectory is replanned and executed to recover from failure cases.

IV. Experiments

In the experiments, we seek to address the following research questions. *Q1*: How effective is the proposed system in whole-body manipulation tasks? *Q2*: How well does the system generalize across variations in object poses, scene configurations, and object stiffness? *Q3*: How robust is the system recovery to external interference, including disturbances during navigation and posture calibration stage, and failures during manipulation stage?

A. Experiment Setups

1) Tasks: We evaluate Praxis across its three operational phases: navigation, posture calibration, and manipulation. To assess the system's robustness during navigation and calibration, we vary initialization distances of 3 m and 5 m relative to the workspace. We benchmark the performance of each distance under two distinct locomotion protocols: a direct linear approach versus a sequential 'orient-then-advance' strategy toward the workspace. For the manipulation phase, we conduct five real-world whole-body manipulation tasks: `Hand Over`, `Pour Water`, `Pull Drawer`, `Open Lid`, and `Manipulate Pipette`. The objects manipulated in these tasks span rigid, transparent, articulated, deformable, flexible linear, and non-prehensile categories. Solving them requires a diverse set of primitive skills, including pick-and-place, re-orient, pull-and-push, lift, and dexterous finger motions such as open, clench, flex, pinch, press, and release. Several skills demand coordinated bimanual control and multi-finger interaction. Moreover, all tasks are long-horizon, consisting of multiple sequential or parallel steps, which substantially increases their overall complexity. We next describe each task in detail, where L, R, L&R denote the step is performed by left, right, and both hands, respectively.

- `Hand Over`: involves a paper cup and a plastic tray; and consists of 5 steps: pick cup (R), close to left hand (R), hand over (L&R), move to tray (L), place cup (L).

TABLE I: Step-wise success rates and average completed step length across all three stages on two tasks. Navigation, Posture Calibration, and Manipulation are abbreviated as Nav., Post. Calib., and Manip., respectively. M. indicates reaching the manipulation region via translation only, while M.+T. requires turning during approach.

`Hand Over`								
Distance	Stage 1	Stage 2	Stage 3: Manip.					**Avg. Len.**
	Nav.	Post. Calib.	pick cup	close to right hand	hand over	move to tray	place cup	
3 m (M.)	10/10	9/10	9/10	9/10	9/10	9/10	9/10	6.4
5 m (M.)	7/10	7/10	7/10	7/10	6/10	6/10	6/10	4.6
3 m (M.+T.)	7/10	6/10	5/10	6/10	5/10	5/10	5/10	3.9
5 m (M.+T.)	4/10	4/10	4/10	3/10	3/10	3/10	3/10	2.4

`Pour Water`										
Distance	Stage 1	Stage 2	Stage 3: Manip.							**Avg. Len.**
	Nav.	Post. Calib.	pick cup	pick bottle	close to each other	pour water	place bottle	place cup		
3 m (M.)	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10		8.0
5 m (M.)	8/10	7/10	7/10	7/10	7/10	7/10	7/10	7/10		5.7
3 m (M.+T.)	6/10	6/10	6/10	6/10	6/10	6/10	6/10	6/10		4.8
5 m (M.+T.)	5/10	5/10	5/10	5/10	5/10	4/10	4/10	4/10		3.7

- `Pour Water`: involves a bottle with liquid and a paper cup; and consists of 6 steps: pick cup (L), pick bottle (R), close to each other (L&R), pour water (R), place bottle (R), place cup (L).
- `Pull Drawer`: involves a drawer with two handles and a building block; and consists of 4 steps: pull drawer (L), pick block (R), place block (R), push drawer (L).
- `Open Lid`: involves a box with a lid and a toy; and consists of 5 steps: lift lid (R), pick toy (L), move lid away (R), place toy in box (L), place lid (R).
- `Manipulate Pipette`: involves a pipette, an experiment shelf, and a plastic tray; and consists of 5 steps: pick pipette (R), press button (R, aspiration), release button (R), press button (R, dispensing), release button (R).

2) Baselines: We compare our method against four large-scale pretrained VLA models that generate upper-body actions, while sharing the same lower-body controller as Praxis. (1)-(3) π_0 [4]/π_0-FAST [43]/$\pi_{0.5}$ [22], which use hybrid multimodal supervision across heterogeneous multi-robot data to master long-horizon dexterous manipulation in the wild; and (4) GR00T N1.6 [2], which couples a vision-language backbone and diffusion-transformer head to facilitate generalized humanoid robot skills. We use teleoperation to construct fine-tuning datasets (85 trajectories per task) for VLA baselines, as detailed in Appendix E.

3) Metrics: We evaluate each method with 33 trials in every task. For a more granular comparison, we report the average completed step length (Avg. Len.; following CALVIN [38, 69]), and step-wise success rates (successful trials / total trials) where the last step indicates the overall task success rate. We exclude trials affected by hardware anomalies, such as G1 or Revo2 failures due to low battery or overheating. We refer to Fig. 3 to identify the low-IK-error workspace and the G1 head-camera FOV, whose intersection defines the feasible object placement region. To preserve sufficient manipulation

TABLE II: Quantitative performance comparisons on five long-horizon whole-body manipulation tasks. We report step-wise success rates (successful trials / total trials) and average completed task length (Avg. Lens). Bold entries denote the best-performing method. Colors (blue, orange, purple) indicate left arm, right arm, and both arms, respectively.

Methods	Hand Over						Pour Water						
	pick cup	close to left hand	hand over	move to tray	place cup	**Avg. Len.**	pick cup	pick bottle	close to each other	pour water	place bottle	place cup	**Avg. Len.**
π_0	22/33	18/33	9/33	8/33	4/33	1.64	23/33	26/33	21/33	12/33	8/33	9/33	2.79
π_0-FAST	21/33	19/33	11/33	10/33	6/33	1.55	21/33	25/33	19/33	10/33	7/33	7/33	2.67
$\pi_{0.5}$	27/33	25/33	16/33	12/33	10/33	1.97	25/33	25/33	22/33	17/33	15/33	15/33	3.42
GR00T N1.6	23/33	19/33	10/33	8/33	5/33	1.58	22/33	23/33	21/33	14/33	12/33	11/33	2.76
Praxis w/o Tact.	28/33	28/33	26/33	26/33	26/33	4.06	29/33	31/33	28/33	27/33	27/33	27/33	5.12
Praxis (Ours)	31/33	31/33	31/33	30/33	30/33	**4.64**	32/33	31/33	30/33	30/33	30/33	30/33	**5.55**

Methods	Pull Drawer					Open Lid						Manipulate Pipette					
	pull drawer	pick block	place block	push drawer	**Avg. Len.**	lift lid	pick toy	move lid away	place toy in box	place lid	**Avg. Len.**	pick pipette	press button	release button	press button	release button	**Avg. Len.**
π_0	11/33	22/33	7/33	5/33	0.76	5/33	26/33	4/33	2/33	2/33	1.03	15/33	1/33	1/33	0/33	0/33	0.33
π_0-FAST	12/33	20/33	6/33	5/33	0.79	6/33	25/33	5/33	4/33	3/33	1.03	13/33	2/33	2/33	0/33	0/33	0.21
$\pi_{0.5}$	14/33	25/33	11/33	8/33	0.88	8/33	26/33	6/33	5/33	5/33	1.21	17/33	2/33	2/33	0/33	0/33	0.42
GR00T N1.6	12/33	24/33	8/33	6/33	0.85	6/33	24/33	5/33	3/33	3/33	0.88	13/33	1/33	1/33	0/33	0/33	0.24
Praxis w/o Tact.	24/33	29/33	22/33	22/33	2.94	19/33	29/33	17/33	16/33	16/33	2.94	24/33	17/33	17/33	14/33	14/33	2.61
Praxis (Ours)	29/33	31/33	26/33	26/33	**3.39**	25/33	31/33	22/33	21/33	20/33	**3.61**	28/33	23/33	23/33	21/33	21/33	**3.52**

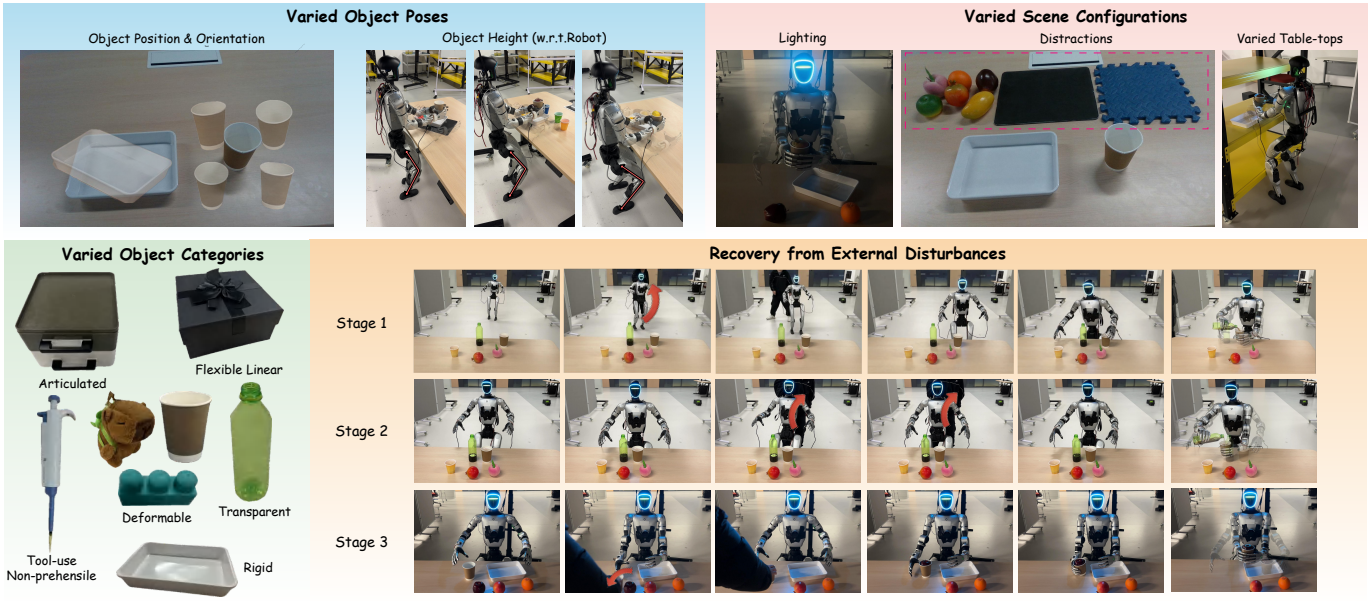

Fig. 4: Visualization of varied object poses, scene configurations, and object categories, and failure recovery across three stages.

clearance, we also avoid placing objects too close to the dexterous hand, especially near the back of the hand.

B. Experimental Results

In the following, we answer *Q1–Q3* in order, covering Praxis overall performance, generalization under diverse environmental variations, and robustness to external interference. For further visualizations, please refer to Appendix I and J.

***(Q1)* Effectiveness of Praxis in whole-body manipulation tasks.** Table I and V summarize the quantitative performance of Praxis on long-distance, long-horizon whole-body manipulation tasks. Results show that Praxis performs strongly in short-distance scenarios. At a distance of 3 m without turning in navigation, `Hand Over` achieves perfect navigation success and high reliability across manipulation steps, while `Pour Water` attains 100% success rate across all stages. At longer distances (5 m) with turning, which is a very challenging setting, Praxis achieves a non-trivial long-distance performance. The system achieves 30% to 40% success rates with average lengths of 2.4 and 3.7, respectively. We attribute this robustness to the combination of a manipulation-region–aware VLN policy and an object-centric posture calibration module, which together stabilize the transition from random initial configurations to manipulation-ready postures.

TABLE III: Step-wise success rates and average completed step length on recovery cases across all three stages.

Interference happens	Stage 1&2	Pour Water: Stage 3: Manip.						**Avg. Len.**
		pick cup	pick bottle	close to each other	pour water	place bottle	place cup	
Stage 1 & 2	8/10	7/10	8/10	7/10	7/10	7/10	7/10	5.1
Stage 2	9/10	9/10	9/10	9/10	9/10	9/10	9/10	6.3
Stage 3	10/10	10/10	10/10	10/10	9/10	9/10	9/10	6.7

TABLE IV: Comparison of the average success rate (Avg. SR) of methods over five manipulation tasks, derived from Tab. II.

Methods	π_0	π_0-FAST	$\pi_{0.5}$	GR00T N1.6	Praxis (Ours)
Avg. SR	5.45%	4.85%	10.30%	5.45%	**76.97%**

***(Q2)* Generalization of Praxis under diverse environmental variations.** To validate our method as a generalist policy, we introduce systematic variations in the initial object pose for each task. Specifically, we randomize the planar position, height relative to the robot, and orientation across trials, enforcing a minimum deviation of 3 cm in translation and 10° in rotation (along at least one axis) from the previous configuration. Furthermore, we vary lighting conditions and scene layouts, and introduce visual distractions, as illustrated in Fig. 4. As detailed in Tab. II and IV, Praxis substantially outperforms π series and GR00T N1.6 in both average completion length and average success rate. We attribute this performance gap to three critical design choices in our pipeline: (1) precise extraction of wrist and hand trajectories from human demonstrations; (2) object-centric execution, which preserves interaction geometry despite pose variations; and (3) tactile-informed grasping, which stabilizes contact against stiffness changes. Despite being finetuned on our specific task trajectories, generalist VLA baselines exhibit poor performance due to a fundamental embodiment and domain gap. These models are pretrained primarily on standard bimanual arms, making them ill-equipped for the high-DoF control required by a humanoid dexterous hand. Consequently, they remain effectively "out-of-distribution" regarding both embodiment and action interface, failing to generate the fine-grained EE and finger motions necessary for success. Praxis achieves near-perfect success on `Hand Over` and `Pour Water`, demonstrating stable perception and whole-body coordination over long horizons. The remaining tasks are significantly more challenging due to strict contact geometry constraints: `Pull Drawer` requires inserting fingers into a small handle, while `Open Lid` involves grasping a non-rigid ribbon. Both demand precise wrist rotation and force control to avoid slippage; `Manipulate Pipette` is particularly restrictive; the pipette is rack-mounted, constraining the grasp to a single viable pose that allows for button actuation while holding it still. Despite these geometric hurdles, Praxis maintains high success rates across varied environmental settings. Observed failures were largely localized to collision issues, such as approach collisions during the transition to the first keyframe, or the secondary button on the pipette obstructing finger placement, rather than fundamental policy errors. Appendix G provides a detailed per-task failure analysis.

***(Q3)* Robustness to external interference across all stages.** Beyond static environmental variations, we evaluate the dynamic recovery capability of Praxis under external disturbances across all stages, reporting step-wise success rates, average length, and average success rate in Tab. III and V. This task adopts a configuration where the robot navigates from a distance of 3 m to the operational area to perform manipulation tasks. When interference occurs in the navigation and calibration stages simultaneously, the system achieves a 70% success rate across manipulation steps. This indicates that Praxis effectively tolerates early perturbations and can re-approach the manipulation region. We attribute this to our fine-tuned VLN model, which allows the robot to re-localize and correct its trajectory even after significant displacement. With interference localized to Stage 2, success attains 90%. This reflects the tracking stability of our posture calibration policy, which reliably re-aligns the body and arms to a manipulation-ready state. When disturbances occur during manipulation, our system demonstrates exceptional robustness, achieving a 90% success rate, validating the efficacy of our recovery mechanism.

TABLE V: Performance of Praxis across different experimental scenarios, derived from Tab. I and III. Interf. is short for dynamic interference. S1,2,3 indicates Stage 1,2,3.

Scenarios	Base (3 m)	Long Dist. (5 m)	Interf. (3 m, S1&2)	Interf. (3 m, S2)	Self-Recovery (3 m, S3)
Avg. SR	75%	50%	70%	90%	90%

V. Conclusion and Limitation

Conclusion. In this paper, we propose Praxis, a novel hierarchical framework that orchestrates humanoid whole-body manipulation through three sequential stages: navigation, posture calibration, and manipulation, to ensure generalization and robustness across long-distance, long-horizon manipulation tasks. First, we adapt a foundation VLN model by fine-tuning it on a strategically curated dataset. This enables the robot to semantically interpret the surroundings and reliably perform navigation toward the manipulation region. Second, a posture calibration module perceives the target object area and reconfigures the humanoid's whole-body stance, aligning the robot pose into a state that maximizes manipulability. The manipulation stage distills physical interaction priors, specifically 6D wrist poses and fine-grained finger motions, directly from one-shot video keyframes. By injecting this knowledge into the online perception loop, the robot can dynamically retarget complex human interactions to new object poses and scene variations. This enables Praxis to handle diverse environmental variations in a fully learning-free manner, requiring only a single human teaching. We conduct extensive empirical studies across five challenging tasks and diverse environmental conditions. Our results demonstrate Praxis's superior generalization capabilities across spatial, visual, and cross-object domains. Furthermore, we highlight the system's strong robustness to dynamically recover from physical disturbances and resume execution at any stage of the pipeline.

Limitation. Despite the strong performance of Praxis in long-horizon whole-body manipulation and its generalization across diverse task environments, we identify three key avenues for future improvement: (1) While we successfully utilize demonstrations from the G1 platform, a substantial domain gap remains between these robot-centric data and the vast potential of in-the-wild internet videos. Bridging this gap is essential for further scaling object manipulation skills. (2) Praxis does not explicitly account for collision avoidance when unexpected interference occurs along planned trajectories. This may affect execution robustness in cluttered or dynamic settings. (3) Praxis would benefit from integrating higher-level autonomy modules, such as safe self-exploration and continual learning within the operational space, to better interpret and accomplish user-specified tasks.

References

[1] Qingwei Ben, Feiyu Jia, Jia Zeng, Junting Dong, Dahua Lin, and Jiangmiao Pang. Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit. *arXiv preprint arXiv:2502.13013*, 2025.

[2] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[3] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164*, 2024.

[4] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A Vision-Language-Action Flow Model for General Robot Control, 2026.

[5] Remi Cadene, Simon Alibert, Alexander Soare, Quentin Gallouedec, Adil Zouitine, Steven Palma, Pepijn Kooijmans, Michel Aractingi, Mustafa Shukor, Dana Aubakirova, Martino Russi, Francesco Capuano, Caroline Pascal, Jade Choghari, Jess Moss, and Thomas Wolf. LeRobot: State-of-the-art Machine Learning for Real-World Robotics in Pytorch. https://github.com/huggingface/lerobot, 2024.

[6] Sixiang Chen, Jiaming Liu, Siyuan Qian, Han Jiang, Lily Li, Renrui Zhang, Zhuoyang Liu, Chenyang Gu, Chengkai Hou, Pengwei Wang, et al. AC-DiT: Adaptive Coordination Diffusion Transformer for Mobile Manipulation. *arXiv preprint arXiv:2507.01961*, 2025.

[7] Yuanpei Chen, Chen Wang, Yaodong Yang, and Karen Liu. Object-Centric Dexterous Manipulation from Human Motion Data. In *8th Annual Conference on Robot Learning*, 2024.

[8] An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Xueyan Zou, Jan Kautz, Erdem Biyik, Hongxu Yin, Sifei Liu, and Xiaolong Wang. NaVILA: Legged Robot Vision-Language-Action Model for Navigation. In *Robotics: Science and Systems, RSS*, 2025.

[9] Xuxin Cheng, Yandong Ji, Junming Chen, Ruihan Yang, Ge Yang, and Xiaolong Wang. Expressive Whole-Body Control for Humanoid Robots. In *Robotics: Science and Systems, RSS*, 2024.

[10] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-TeleVision: Teleoperation with Immersive Active Visual Feedback. In *Conference on Robot Learning, CoRL*, volume 270, pages 2729–2749, 2024.

[11] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022.

[12] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023.

[13] Yuhui Fu, Feiyang Xie, Chaoyi Xu, Jing Xiong, Haoqi Yuan, and Zongqing Lu. DemoHLM: From One Demonstration to Generalizable Humanoid Loco-Manipulation. *arXiv preprint arXiv:2510.11258*, 2025.

[14] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. *arXiv preprint arXiv:2406.10454*, 2024.

[15] Zipeng Fu, Tony Z Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024.

[16] Jianfeng Gao, Xiaoshu Jin, Franziska Krebs, Noémie Jaquier, and Tamim Asfour. Bi-kvil: Keypoints-based visual imitation learning of bimanual manipulation tasks. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16850–16857. IEEE, 2024.

[17] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[18] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024.

[19] Zhaoyuan Gu, Junheng Li, Wenlan Shen, Wenhao Yu, Zhaoming Xie, Stephen McCrory, Xianyi Cheng, Abdulaziz Shamsah, Robert Griffin, C Karen Liu, et al. Humanoid locomotion and manipulation: Current progress and challenges in control, planning, and learning. *arXiv preprint arXiv:2501.02116*, 2025.

[20] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *arXiv preprint arXiv:2406.08858*, 2024.

[21] Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning Human-to-Humanoid Real-Time Whole-Body Teleoperation. In

IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, pages 8944–8951, 2024.

[22] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[23] Stephen James, Kentaro Wada, Tristan Laidlow, and Andrew J Davison. Coarse-to-fine q-attention: Efficient learning for visual robotic manipulation via discretisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13739–13748, 2022.

[24] Haoran Jiang, Jin Chen, Qingwen Bu, Li Chen, Modi Shi, Yanjie Zhang, Delong Li, Chuanzhe Suo, Chuang Wang, Zhihui Peng, and Hongyang Li. WholeBodyVLA: Towards Unified Latent VLA for Whole-Body Loco-Manipulation Control. *arXiv preprint arXiv:2512.11047*, 2025.

[25] Zhao Jin, Zhengping Che, Zhen Zhao, Kun Wu, Yuheng Zhang, Yinuo Zhao, Zehui Liu, Qiang Zhang, Xiaozhu Ju, Jing Tian, Yousong Xue, and Jian Tang. ArtVIP: Articulated Digital Assets of Visual Realism, Modular Interaction, and Physical Fidelity for Robot Learning. *arXiv preprint arXiv:2506.04941*, 2025.

[26] Simar Kareer, Karl Pertsch, James Darpinian, Judy Hoffman, Danfei Xu, Sergey Levine, Chelsea Finn, and Suraj Nair. Emergence of Human to Robot Transfer in Vision-Language-Action Models. *arXiv preprint arXiv:2512.22414*, 2025.

[27] Justin Kerr, Chung Min Kim, Mingxuan Wu, Brent Yi, Qianqian Wang, Ken Goldberg, and Angjoo Kanazawa. Robot See Robot Do: Imitating Articulated Object Manipulation with Monocular 4D Reconstruction. In *8th Annual Conference on Robot Learning*, 2024.

[28] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. OpenVLA: An Open-Source Vision-Language-Action Model. *arXiv preprint arXiv:2406.09246*, 2024.

[29] Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B Tenenbaum. Learning to Act from Actionless Videos through Dense Correspondences. In *The Twelfth International Conference on Learning Representations*, 2024.

[30] Gen Li, Nikolaos Tsagkas, Jifei Song, Ruaridh Mon-Williams, Sethu Vijayakumar, Kun Shao, and Laura Sevilla-Lara. Learning precise affordances from egocentric videos for robotic manipulation. *arXiv preprint arXiv:2408.10123*, 2024.

[31] Jialong Li, Xuxin Cheng, Tianshu Huang, Shiqi Yang, Ri-Zhao Qiu, and Xiaolong Wang. AMO: Adaptive Motion Optimization for Hyper-Dexterous Humanoid Whole-Body Control. *arXiv preprint arXiv:2505.03738*, 2025.

[32] Jinhan Li, Yifeng Zhu, Yuqi Xie, Zhenyu Jiang, Mingyo Seo, Georgios Pavlakos, and Yuke Zhu. Okami: Teaching humanoid robots manipulation skills through single video imitation. In *8th Annual Conference on Robot Learning*, 2024.

[33] Jiacheng Liu, Pengxiang Ding, Qihang Zhou, Yuxuan Wu, Da Huang, Zimian Peng, Wei Xiao, Weinan Zhang, Lixin Yang, Cewu Lu, et al. TrajBooster: Boosting Humanoid Whole-Body Manipulation via Trajectory-Centric Learning. *arXiv preprint arXiv:2509.11839*, 2025.

[34] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. RDT-1B: a Diffusion Foundation Model for Bimanual Manipulation. *arXiv preprint arXiv:2410.07864*, 2024.

[35] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *IEEE/ASME Transactions on Mechatronics*, 2025.

[36] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21740–21751, 2024.

[37] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093*, 2024.

[38] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022.

[39] Soroush Nasiriany, Sean Kirmani, Tianli Ding, Laura Smith, Yuke Zhu, Danny Driess, Dorsa Sadigh, and Ted Xiao. RT-Affordance: Affordances are Versatile Intermediate Representations for Robot Manipulation. *arXiv preprint arXiv:2411.02704*, 2024.

[40] NVIDIA. Isaac Sim. URL https://github.com/isaac-sim/IsaacSim.

[41] Georgios Papagiannis, Norman Di Palo, Pietro Vitiello, and Edward Johns. R+X: Retrieval and execution from everyday human videos. *arXiv preprint arXiv:2407.12957*, 2024.

[42] Weikun Peng, Jun Lv, Yuwei Zeng, Haonan Chen, Siheng Zhao, Jichen Sun, Cewu Lu, and Lin Shao. TieBot: Learning to Knot a Tie from Visual Demonstration through a Real-to-Sim-to-Real Approach. In *8th Annual Conference on Robot Learning*, 2024.

[43] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. FAST: Efficient Action Tokenization for Vision-Language-Action Models, 2025.

[44] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: End-to-end 3D Hand Localization and Reconstruction in-the-wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 12242–12254, 2025.

[45] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[46] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks, 2024. URL https://arxiv.org/abs/2401.14159.

[47] Javier Romero, Dimitris Tzionas, and Michael J Black. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics*, 36(6), 2017.

[48] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[49] Kenneth Shaw, Shikhar Bahl, Aravind Sivakumar, Aditya Kannan, and Deepak Pathak. Learning dexterity from human hand motion in internet videos. *The International Journal of Robotics Research*, 43(4):513–532, 2024.

[50] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023.

[51] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

[52] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[53] Lai Wei, Xuanbin Peng, Ri-Zhao Qiu, Tianshu Huang, Xuxin Cheng, and Xiaolong Wang. HMC: Learning Heterogeneous Meta-Control for Contact-Rich Loco-Manipulation. *arXiv preprint arXiv:2511.14756*, 2025.

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

[55] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. *arXiv preprint arXiv:2401.00025*, 2023.

[56] Haoyang Weng, Yitang Li, Nikhil Sobanbabu, Zihan Wang, Zhengyi Luo, Tairan He, Deva Ramanan, and Guanya Shi. HDMI: Learning Interactive Humanoid Whole-Body Control from Human Videos. *arXiv preprint arXiv:2509.16757*, 2025.

[57] Wansen Wu, Tao Chang, Xinmeng Li, Quanjun Yin, and Yue Hu. Vision-language navigation: a survey and taxonomy. *Neural Computing and Applications*, 36(7): 3291–3316, 2024.

[58] Zhenyu Wu, Yuheng Zhou, Xiuwei Xu, Ziwei Wang, and Haibin Yan. Momanipvla: Transferring vision-language-action models for general mobile manipulation. In *IEEE / CVF Computer Vision and Pattern Recognition Conference, CVPR*, pages 1714–1723, 2025.

[59] Weiji Xie, Jinrui Han, Jiakun Zheng, Huanyu Li, Xinzhe Liu, Jiyuan Shi, Weinan Zhang, Chenjia Bai, and Xuelong Li. KungfuBot: Physics-Based Humanoid Whole-Body Control for Learning Highly-Dynamic Skills. *arXiv preprint arXiv:2506.12851*, 2025.

[60] Wenhao Yan and Jie Chu. FoundationPose++, March 2025. URL https://github.com/teal024/FoundationPose-plus-plus.

[61] Lujie Yang, Xiaoyu Huang, Zhen Wu, Angjoo Kanazawa, Pieter Abbeel, Carmelo Sferrazza, C Karen Liu, Rocky Duan, and Guanya Shi. OmniRetarget: Interaction-Preserving Data Generation for Humanoid Whole-Body Loco-Manipulation and Scene Interaction. *arXiv preprint arXiv:2509.26633*, 2025.

[62] Haoqi Yuan, Yu Bai, Yuhui Fu, Bohan Zhou, Yicheng Feng, Xinrun Xu, Yi Zhan, Börje F Karlsson, and Zongqing Lu. Being-0: A Humanoid Robotic Agent with Vision-Language Models and Modular Skills. *arXiv preprint arXiv:2503.12533*, 2025.

[63] Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. OAKINK2: A Dataset of Bimanual Hands-Object Manipulation in Complex Task Completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 445–456, 2024.

[64] Chong Zhang, Wenli Xiao, Tairan He, and Guanya Shi. Wococo: Learning whole-body humanoid control with sequential contacts. *arXiv preprint arXiv:2406.06005*, 2024.

[65] Fan Zhang and Michael Gienger. Affordance-based Robot Manipulation with Flow Matching. *arXiv preprint arXiv:2409.01083*, 2024.

[66] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020.

[67] Xinyu Zhang and Abdeslam Boularias. One-Shot Imitation Learning with Invariance Matching for Robotic Manipulation. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[68] Yuanhang Zhang, Yifu Yuan, Prajwal Gurunath,

Ishita Gupta, Shayegan Omidshafiei, Ali-akbar Agha-mohammadi, Marcell Vazquez-Chanlatte, Liam Pedersen, Tairan He, and Guanya Shi. Falcon: Learning force-adaptive humanoid loco-manipulation. *arXiv preprint arXiv:2505.06776*, 2025.

[69] Huayi Zhou, Ruixiang Wang, Yunxin Tai, Yueci Deng, Guiliang Liu, and Kui Jia. You Only Teach Once: Learn One-Shot Bimanual Robotic Manipulation from Video Demonstrations. *Robotics: Science and Systems, RSS*, 2025.

[70] Ziwen Zhuang, Shenzhe Yao, and Hang Zhao. Humanoid Parkour Learning. In *Conference on Robot Learning, CoRL*, volume 270, pages 1975–1991. PMLR, 2024.

APPENDIX

A. Details on Hardware System

1) Robot Embodiment: Humanoid platforms extend the workspace of bimanual manipulation beyond fixed-base arms, and are suitable for human video-inspired whole-body manipulation. We adopt the Unitree G1 EDU humanoid robot[1], with 27 degrees of freedom (DoFs). We establish a wireless TCP interface between the robot and a host desktop equipped with an NVIDIA RTX 4090 GPU. This architecture supports real-time, asynchronous perception, planning, and control over a shared Wi-Fi network.

2) End-Effector Selection: Since the stock G1 humanoid's rubber hands do not support dexterous manipulation, we replace them with BrainCo Revo2 hands[2], a five-fingered tactile-sensing hand with 11 degrees of freedom (DoFs), with three in the thumb and two in each of the index, middle, ring, and pinky fingers. Each of the two Revo2 hands is connected to the G1 via an RS485-to-Type-C interface operating at a baud rate of 1M, enabling high-speed communication. The integrated tactile sensing substantially improves the hand's capability for precise and robust manipulation, particularly for articulated and deformable object interactions, as demonstrated in our real-world experiments.

3) Camera Observation: The G1 humanoid is originally equipped with an Intel RealSense D435i RGB-D camera embedded[3] integrated into the head with a fixed downward pitch of $48°$, which provides a fixed and limited field of view (FoV). To support long-range perception for navigation, we additionally mount a RealSense L515 solid-state time-of-flight LiDAR camera[4] on the forward-facing surface of the torso. The L515 is used for navigation perception, while the D435i supports pre-manipulation calibration and manipulation. The D435i operates at a resolution of 1280×720 at 25 Hz, with depth frames aligned to the RGB stream, while L515 operates at a resolution of 640×480 at 25 Hz.

Fig. 5 illustrates our hardware configuration, where the bracketed text denotes how each device is connected to the G1 onboard computer. To ensure responsive control, all I/O, including dual-camera image acquisition and proprioception, is handled asynchronously, avoiding stalls in the main control thread and enabling the whole-body controller to run reliably at 50Hz.

B. Comparison between Recent Whole-body Manipulation Systems and Praxis

As shown in Tab. VI, we compare four recent whole-body manipulation systems along eight dimensions—dexterity, one-shot, learning from video, vision-and-language input, long-distance navigation ($\geq$ 5m), long-horizon execution, recovery, and training-free deployment. Here, long-distance navigation indicates that the minimum distance the robot must travel

[1]https://www.unitree.com/g1
[2]https://www.brainco.cn/en-US/products/revo2
[3]https://www.realsenseai.com/products/depth-camera-d435i
[4]https://www.realsenseai.com/products/realsense-lidar-camera-l515

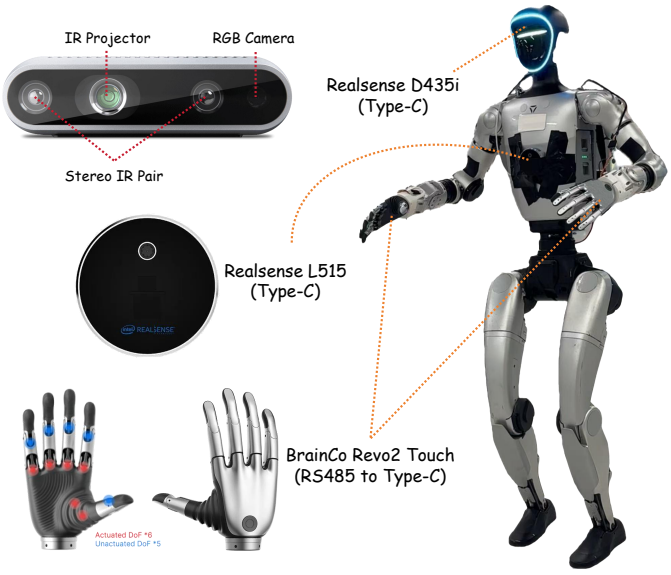

Fig. 5: Hardware system of Praxis.

to reach the manipulation workspace is at least 5 m, and long-horizon execution means the manipulation itself contains at least four distinct steps. WholeBodyVLA [24] leverages large-scale action-free videos and supports vision-language conditioning with broad spatial coverage, but it remains limited in fine dexterous manipulation and long-horizon task organization. TrajBooster [33] improves cross-embodiment transfer and whole-body coordination via trajectory lifting and subsequent adaptation, yet it still relies on training and a small amount of real data. DemoHLM [13] can scale from a single demonstration by synthesizing large amounts of simulation data and training policies, but it is not equivalent to real-video one-shot learning and still requires training. HMC [53] focuses on robust contact-rich control through hybrid/multi-mode control switching, addressing low-level execution rather than closed-loop capabilities such as language-guided navigation, long-distance mobility, and training-free task deployment. In contrast, Praxis unifies navigation and manipulation in a single pipeline: it uses vision-language navigation and calibration policy to reach targets over long distances, enables training-free deployment to new tasks via one-shot human video demonstrations, and incorporates a recovery mechanism (failure detection and replanning) to support long-horizon, multi-stage execution.

C. Details on Dataset Construction for Fine-tuning NaVILA

To adapt the general-purpose VLN model to our robotic embodiment and experimental environment, we construct a dedicated fine-tuning dataset. Beyond the strategy described in the main paper, we detail our model initialization, object-diverse scene augmentation, and evaluation protocol, and multi-faceted randomization strategy in data collection.

1) Model Initialization and Action Space: We initialize our policy using the official NaVILA checkpoint. Crucially, we select the version that has already undergone extensive fine-

TABLE VI: Comparison between recent whole-body manipulation systems and Praxis. Vis. & Lang. Input indicates vision–language input, and Long Dist. refers to long-distance navigation.

Method	Dexterity	One-shot	Learn from Video	Vis. & Lang. Input	Long Dist. ($\geq$ 5 m)	Long Horizon	Recovery
WholeBodyVLA [24]	✗	✗	✓	✓	✓	✗	✓
HMC [53]	✗	✗	✗	✗	✗	✗	✗
DemoHLM [13]	✗	✓	✗	✗	✗	✗	✗
TrajBooster [33]	✓	✗	✗	✓	✗	✗	✗
OKAMI [32]	✓	✓	✓	✗	✗	✗	✗
Praxis (Ours)	✓	✓	✓	✓	✓	✓	✓

tuning on large-scale navigation datasets. We leverage this established locomotor before adapting the model for coarse-level navigation in our specific environment. The objective of this fine-tuning is to enable the robot to recognize target objects and reach their immediate vicinity, after which a calibration module takes over to achieve manipulation-ready postures.

While the original NaVILA model primarily focuses on longitudinal and rotational movements, we expanded the action space by introducing two additional primitives: Move Left and Move Right. This augmentation allows the policy to utilize the robot's holonomic capabilities for a more flexible approach to trajectories during the navigation phase, ensuring efficient traversal without the constraints of non-holonomic kinematics.

Our fine-tuning dataset consists of 150 expert trajectories collected in the real world via teleoperation using a handheld remote controller. The distribution of these trajectories is detailed in Tab. VII.

TABLE VII: Distribution of trajectories in the fine-tuning dataset.

Action Type	**Count**
Move Forward	46
Turn Right	32
Turn Left	32
Move Right	20
Move Left	20
Total	**150**

2) Object Diversity and OOD Evaluation Protocol: To rigorously assess visual robustness, we adopt a strict out-of-distribution (OOD) protocol over scene objects. We curate seven target objects (one shelf and six tables) with high visual diversity in geometry (rectangular vs. circular), appearance (wood grain vs. metal), and size. We enforce a strict train–test split: trajectories for only four table variants are used for training, while the remaining two tables and the shelf are reserved exclusively for testing. This design ensures that our evaluation measures genuine generalization to unseen objects and environments rather than memorization.

3) Geometric Randomization: We vary the robot's initial pose, both radial distance and approach angle relative to the target, to encourage viewpoint invariance.

4) Instruction Augmentation Strategy: To enhance the robustness of instruction-following, we implement a "Semantic Consistency, Stylistic Diversity" annotation strategy. For each trajectory, we generate multiple instruction variants that strictly preserve the core semantic intent (e.g., target object and action type) while varying linguistic structure. For instance, a Move Forward action targeting a table is annotated with diverse phrasings such as:

- "Walk ahead and stop at the table."
- "Proceed forward and stop beside the table."

This strategy prevents the model from overfitting to specific lexical patterns and encourages deep semantic alignment.

5) Scalable Hybrid Data Acquisition: We combine robot-collected and human-collected trajectories to scale data efficiently. Human operators are instructed to match the robot's camera height and motion dynamics, preserving domain consistency while expanding coverage across diverse scenes (e.g., different room layouts and table geometries).

D. Qualitative Results of Wrist Detection and Dexterous Retargeting

Fig. 8–11 visualize the detected wrist poses (left: first row; right: third row) and the corresponding dexterous retargeting results on the Revo2 hand (left: second row; right: fourth row). Fig. 12 visualizes the detected wrist poses (right: first row) and the corresponding dexterous retargeting results on the Revo2 hand (right: second row). The predicted wrist trajectories and retargeted finger motions are consistent with the demonstrations, providing reliable robot-centric supervision for one-shot teaching.

E. Details on Data Acquisition and Preprocessing for Baseline VLA

The observation space is structured as follows: dimensions 0–6 correspond to the left arm, 7–13 to the right arm, 14–19 to the left hand, and 20–25 to the right hand. We collect 85 trajectories per task via teleoperation at the same resolution and frame rate, and include them in the finetuning dataset. Teleoperation is performed using the Meta Quest 3[5] as the XR (Extended Reality) device. For both the teleoperated and Praxis rollouts, the robot's upper initial pose is kept identical, and we ensure that the teleoperated trajectories do not differ substantially from the Praxis-generated rollouts. The training

[5]https://www.meta.com/ae/quest/quest-3/

batch size is set to 128, and we train for 5000 steps to ensure the VLA baseline can adequately learn the provided trajectories without overfitting.

During the preprocessing phase, we performed frame-level alignment for each episode, retaining only valid frames that contained both image and observation data. Each frame was assigned a timestamp calculated as $t = \text{frame\_index}/30$, along with a local frame index, an episode index, and a global index. The `next.done` flag was set to true for the final frame of each trajectory, and `next.reward` was initialized to zero.

Following preprocessing, we converted the dataset into the LeRobot [5] standard formats to support different training requirements: LeRobot v2.1 for GR00T N1.6 and v3.0 for π series.

1) v2.1 Conversion (Episode-Based): We organized the raw data into the LeRobot v2.1 format, utilizing an episode-based structure. Observations and RGB images were read frame-by-frame, temporally aligned, and written into individual Parquet files per episode, where both `observation.state` and `action` were stored as 26-dimensional vectors. Image sequences were encoded as H.264 MP4 videos at 30 FPS, with one video file generated for each episode.

Comprehensive metadata was generated, including:

- `meta/info.json`: specifies state and action dimensions, FPS, path rules, and indexing.
- `meta/episodes.jsonl`: records episode lengths.
- `meta/tasks.jsonl`: contains textual task descriptions.
- `meta/episodes_stats.jsonl` and `meta/stats.json`: stores statistical metrics (min, max, $1^{st}/99^{th}$ quantiles, etc.).
- `meta/modality.json`: defines the mapping between the visual input (head camera) and robot components (arms/hands).

2) v2.1 → v3.0 Conversion (File-Based): To facilitate efficient training for the π series, we consolidated the v2.1 data into the v3.0 file-based structure. This involved concatenating Parquet rows from multiple episodes into chunked files (e.g., `data/chunk-*/file-*.parquet`) and merging corresponding video segments into unified video files (e.g., `videos/{video_key}/chunk-*/file-*.mp4`).

Metadata was updated accordingly: `meta/episodes/*.parquet` was generated to index the range and video timestamps of each episode within the merged files, and `meta/tasks.parquet` was created for the task registry. Finally, `meta/info.json` was updated with the new path templates and version information. This transformation converts the data from episode-based storage to a highly efficient, streamable v3.0 format without altering the underlying data semantics.

F. Details and Qualitative Results of Tactile-informed Grasping

To illustrate tactile-informed grasping, we visualize in Fig. 13 the fingertip normal (blue) and tangential (orange) forces applied to the deformable toy in `Open Lid` under two-, three-, and five-finger grasps. While we compare these configurations for analysis, all experiments in the main paper use five-finger grasps to prioritize safety and robustness.

We formulate dexterous grasping as a two-stage process consisting of feedforward finger closure followed by tactile-informed force regulation, where each finger $i \in [5]$ estimates a local contact force $\mathbf{f}_i = [f_{n,i}, \mathbf{f}_{t,i}]$, with $f_{n,i} \geq 0$ denoting the normal force magnitude and $\mathbf{f}_{t,i} \in \mathbb{R}^2$ the tangential force vector at the fingertip. Contact is detected using the indicator $c_i = \mathbb{I}(f_{n,i} > \tau_n)$, where τ_n is a predefined normal-force threshold, and slip risk is evaluated via the friction ratio $\rho_i = \|\mathbf{f}_{t,i}\|/f_{n,i}$. Once stable multi-finger contact is achieved, defined by $\sum_i c_i \geq M$ with M denoting the minimum required number of contacting fingers, the contact stiffness $k = \text{mean}(\Delta f_{n,i}/\Delta x_{n,i})$ is estimated from the ratio between changes in normal force $\Delta f_{n,i}$ and normal indentation $\Delta x_{n,i}$. The desired normal grasp force is then set as $f_n^\star = \text{clip}(\alpha k + \beta, f_{\min}, f_{\max})$, where α and β are scalar gains and $[f_{\min}, f_{\max}]$ bounds the admissible force range, and further constrained by a no-slip condition $f_{n,i}^\star \geq \|\mathbf{f}_{t,i}\|/(\mu_i - \delta)$, where μ_i is the estimated local coefficient of friction and $\delta > 0$ is a safety margin.

G. Detailed Analysis of Failure Cases

Although Praxis demonstrates strong performance on long-horizon whole-body manipulation, we observe several recurring failure modes in real-world evaluation. We analyze execution failures across the five tasks that are reported in the main paper.

1) Hand Over & Pour Water: A common failure in these two tasks occurs at the very beginning of the manipulation stage: the target object is sometimes placed along the robot's transition path from its default posture to the first keyframe. As a result, the robot collides with the object before reaching the initial pose, causing early termination. This explains why the overall pick success rate is not 100%, despite the manipulation policy being effective once the first keyframe is reached.

2) Pull Drawer: During drawer pulling, failures are often caused by finger–handle disengagement. The handle provides a limited contact area and can induce slip or rolling contact under imperfect approach angles, leading to an unstable grasp and unsuccessful opening. In addition, when lifting an object near the drawer, the carried object may collide with the drawer front or surrounding structure, causing it to drop during transport and reducing the success rate of object placement.

3) Open Lid: The gift box lid is primarily grasped via a thin ribbon, which is a challenging target for finger closure. In several cases, the fingers fail to securely capture the ribbon due to its small thickness and deformability, resulting in an unsuccessful lift of the lid.

4) Manipulate Pipette: The pipette contains an additional nearby button used to eject the disposable tip. This tip-ejection button is located close to the aspiration button and can be accidentally contacted during execution. Such unintended presses change the hand's contact configuration, disturb the intended

finger placement, and reduce effective force transmission to the aspiration button, leading to failed button presses and contributing to the observed drop in success rate.

H. Object Masking via LISA for Robust Perception in FoundationPose++

To facilitate accurate 6D pose estimation through FoundationPose++, we utilize LISA (Large Language Instructed Segmentation Assistant) [45] to generate high-fidelity object masks as its input. Unlike contemporary open-vocabulary pipelines such as Grounded-SAM [46], which often struggle with implicit instructions or scenarios requiring complex semantic reasoning (e.g., "Find the object that is blocking the door" or "Find the screwdriver needed to fix this toy"), LISA employs an embedding-as-mask paradigm that integrates segmentation directly into a multi-modal LLM's reasoning process by expanding the vocabulary with a <SEG> token.

This integrated approach offers a dual advantage for the Praxis framework by combining superior linguistic sensitivity with extensive world knowledge to handle specialized tasks. In instances where standard models produce masking errors due to scene complexity or visual ambiguity, LISA allows users to provide more descriptive or context-aware prompts, leveraging its internal reasoning to actively correct and refine the segmentation. This capability is particularly vital for identifying and segmenting specialized or out-of-distribution objects, such as the pipette used in our dexterous manipulation experiments, which are frequently unrecognized by conventional segmentors but are well-represented within the LLM's pre-trained knowledge base.

I. Additional Implementation Results in Simulation

To augment our real-world findings, we evaluate the keyframe-based motion framework within high-fidelity simulation environments built in NVIDIA Isaac Sim [40], utilizing digital assets from the ArtVIP dataset [25], a comprehensive library of 206 high-fidelity articulated objects optimized for robotic manipulation. We leverage a Meta Quest 3 headset for teleoperation to facilitate scalable and efficient expert trajectory collection, during which the motion data undergoes a significant abstraction process: by identifying and processing local motion extrema, we isolate a compact set of approximately 10 to 20 keyframes for each trajectory, representing a substantial refinement from the original 500 to 1000 frames. Our empirical observations indicate that simple linear interpolation between these sparse keyframes robustly reconstructs the original demonstration trajectories while preserving interaction stability, as illustrated in Figure 6. These retargeted motion priors enable the robot to execute diverse manipulation primitives, confirming the kinematic feasibility of our keyframe-based representation before real-world deployment.

J. Qualitative Snapshots of Experimental Process across Five Tasks

This section provides a comprehensive visual walkthrough of the five real-world whole-body manipulation tasks conducted in our empirical studies. As visualized in Fig. 7, Praxis demonstrates the ability to coordinate bimanual control and dexterous finger motions to handle objects spanning transparent, rigid, articulated, deformable, and tool-use categories. These sequential snapshots (shadow means earlier state) highlight the system's proficiency in managing long-horizon tasks through a unified pipeline of perception-conditioned execution.

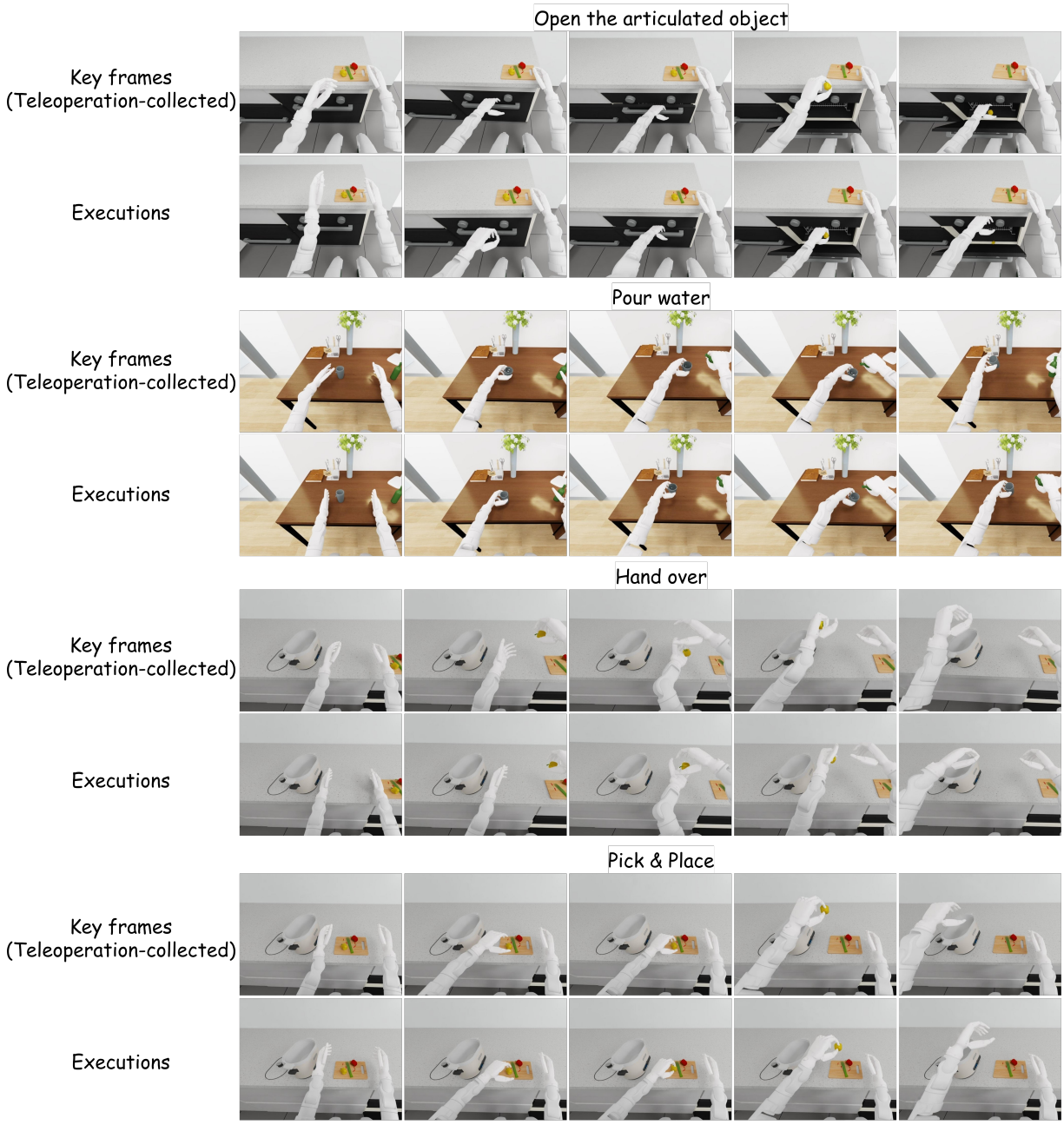

Fig. 6: Sequential snapshots of experiments in simulation. The top and bottom rows for each task display teleoperation-collected keyframes and their corresponding execution sequences, respectively. These results demonstrate the efficacy of our keyframe-based representation in maintaining precise interaction geometry across diverse manipulation tasks.

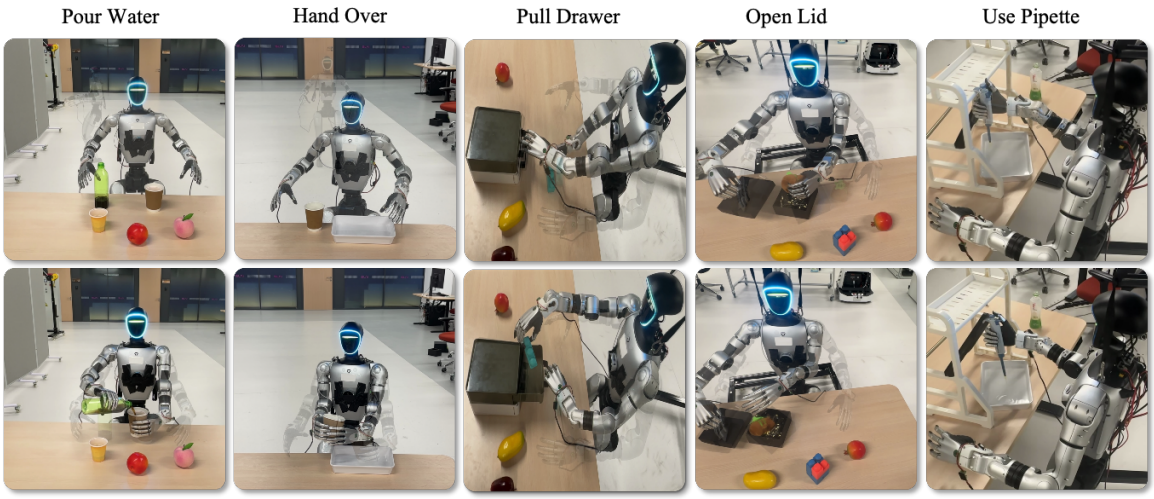

Fig. 7: Sequential snapshots showing the progression of the experiment.

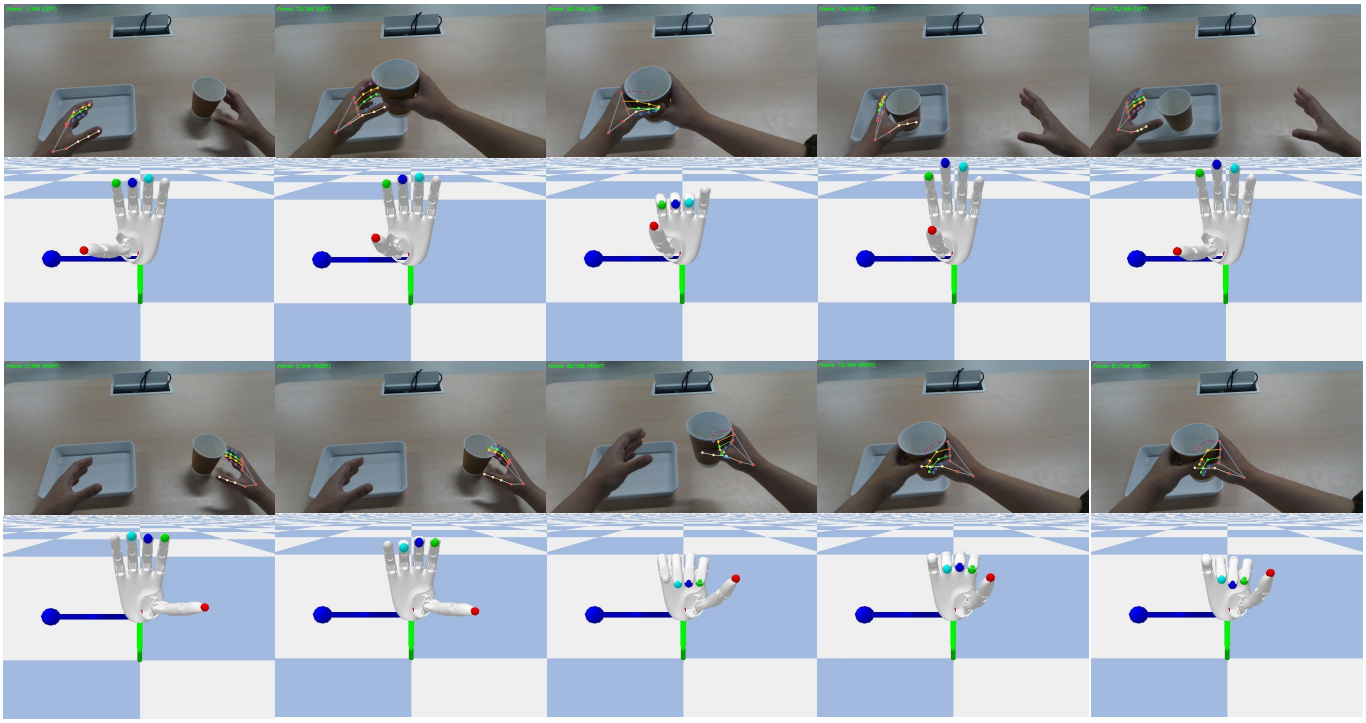

Fig. 8: Visualization of wrist pose detection and dexterous retargeting in task `Hand Over`.

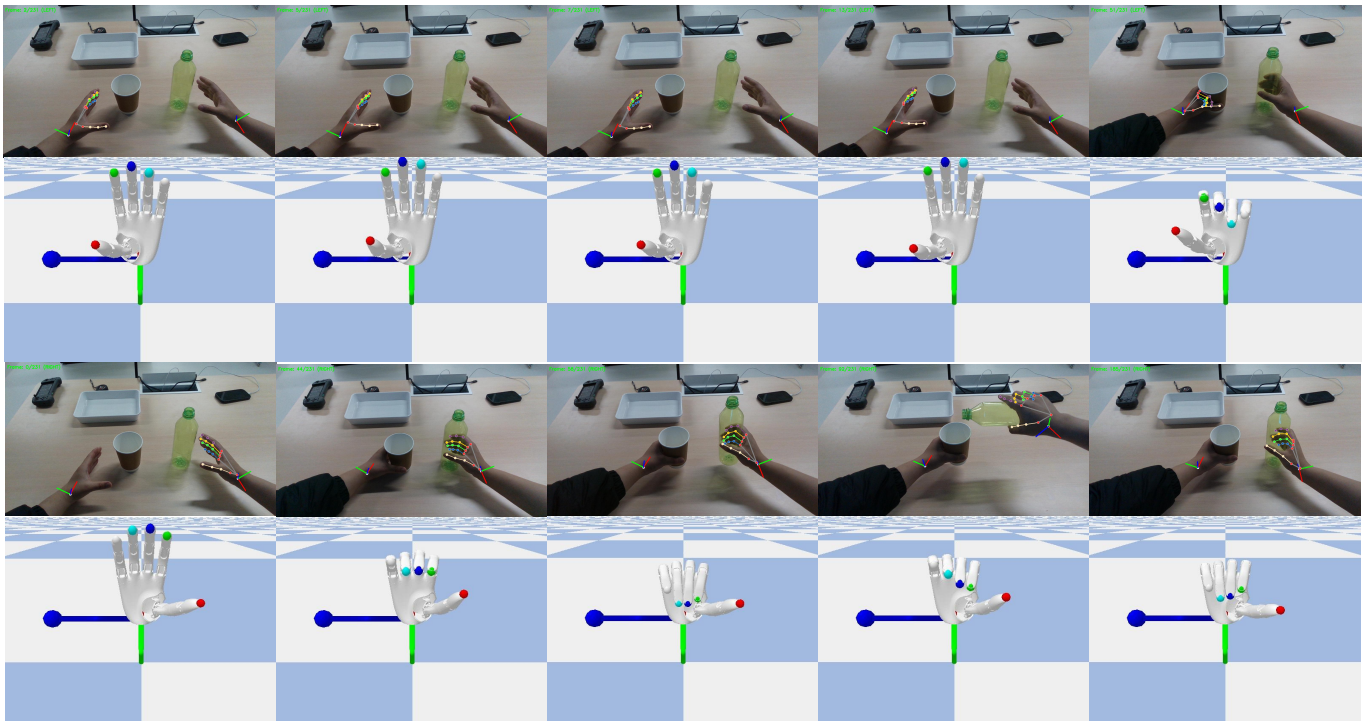

Fig. 9: Visualization of wrist pose detection and dexterous retargeting in task `Pour Water`.

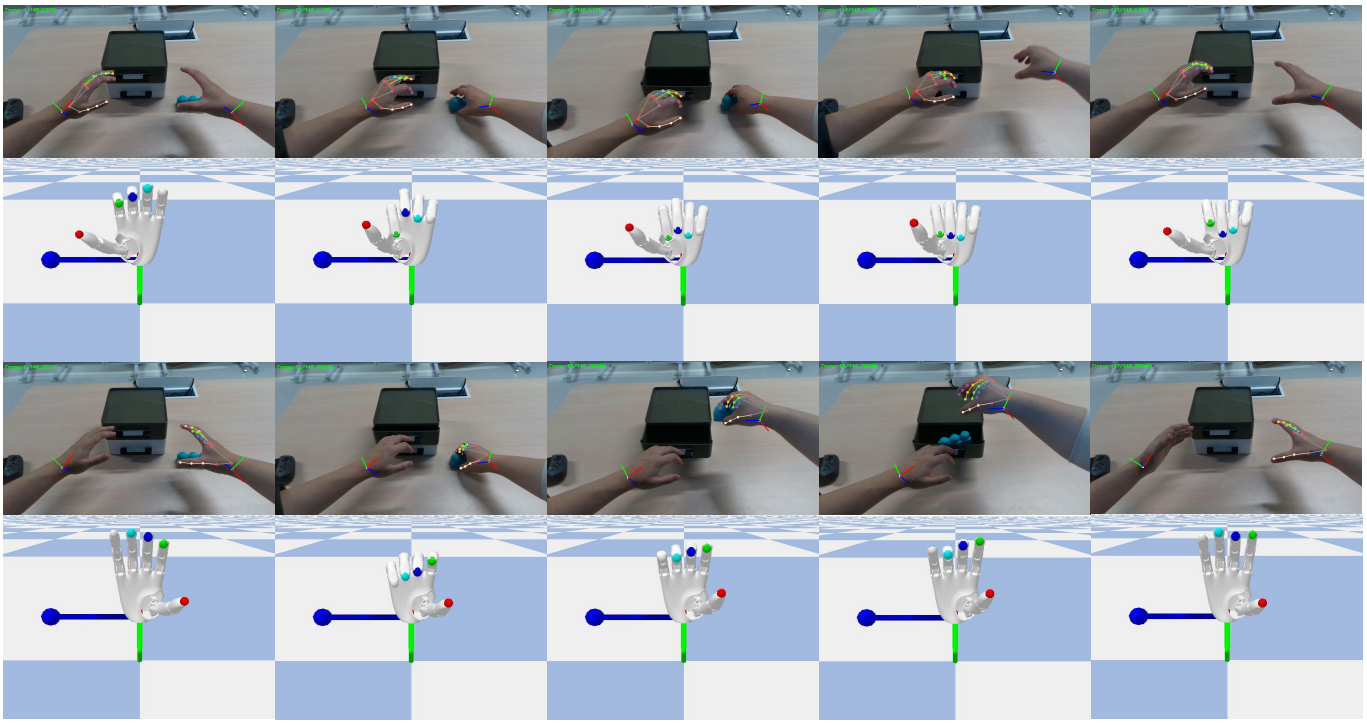

Fig. 10: Visualization of wrist pose detection and dexterous retargeting in task `Pull Drawer`.

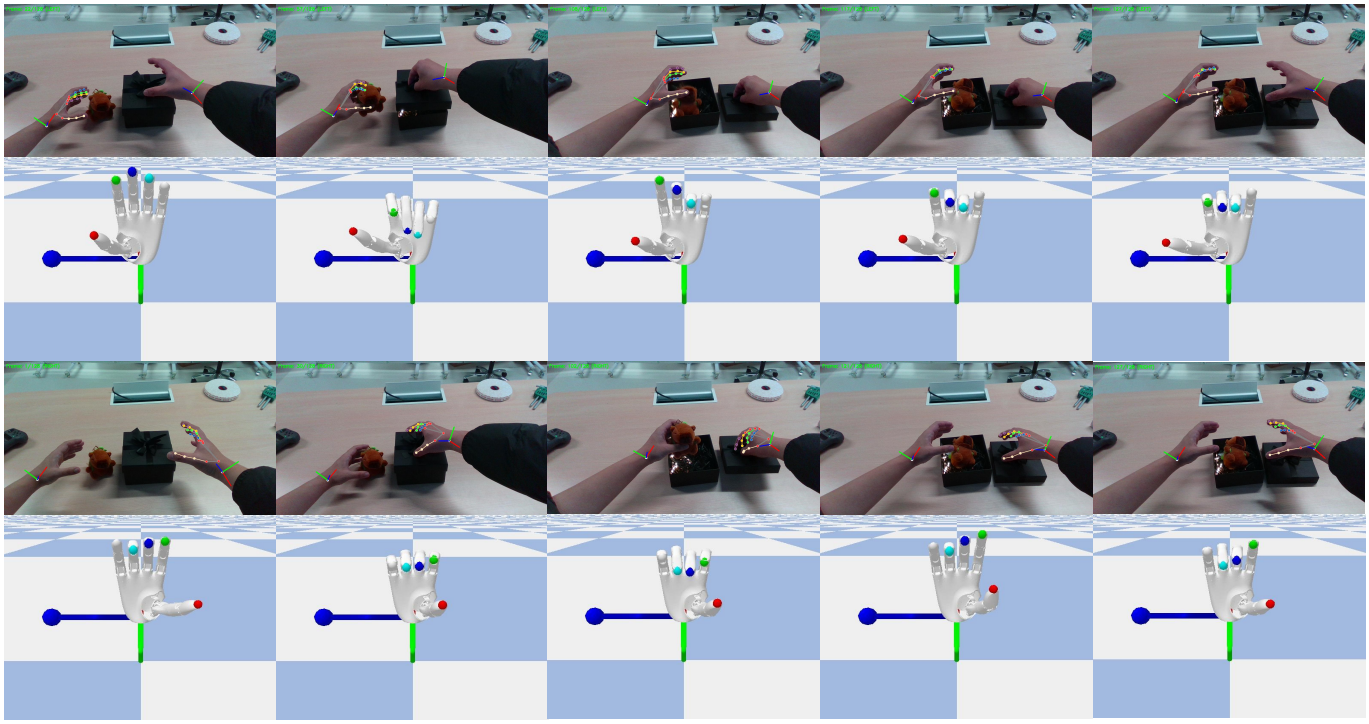

Fig. 11: Visualization of wrist pose detection and dexterous retargeting in task `Open Lid`.

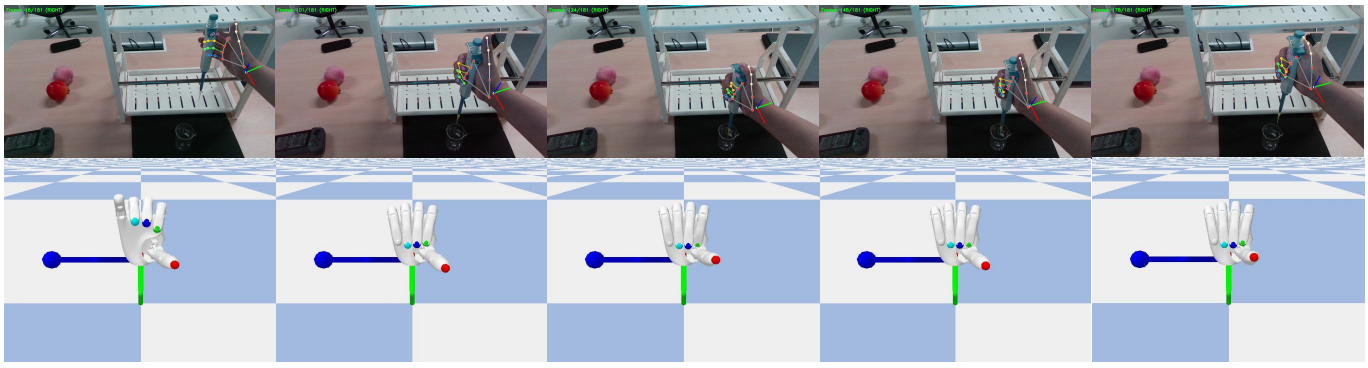

Fig. 12: Visualization of wrist pose detection and dexterous retargeting in task `Manipulate Pipette`.

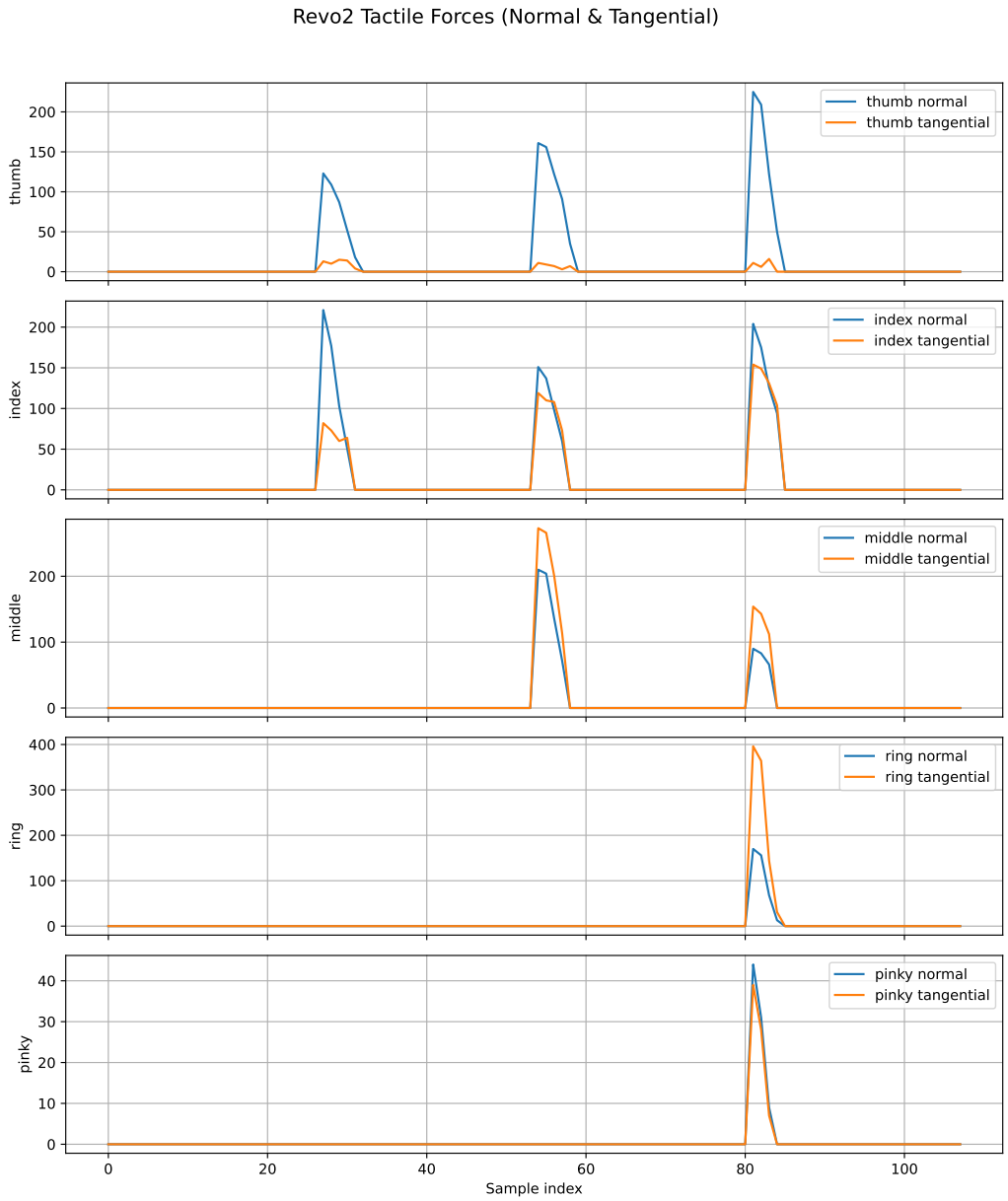

Fig. 13: Visualization of fingertip tactile forces when grasping the toy (step pick toy) in `Open Lid` using two-, three-, and five-finger configurations of the left Revo2 hand.