BEHAVIOR ROBOT SUITE: Streamlining Real-World Whole-Body Manipulation for Everyday Household Activities

Yunfan Jiang, Ruohan Zhang, Josiah Wong, Chen Wang, Yanjie Ze,
Hang Yin, Cem Gokmen, Shuran Song, Jiajun Wu, Li Fei-Fei
behavior-robot-suite.github.io

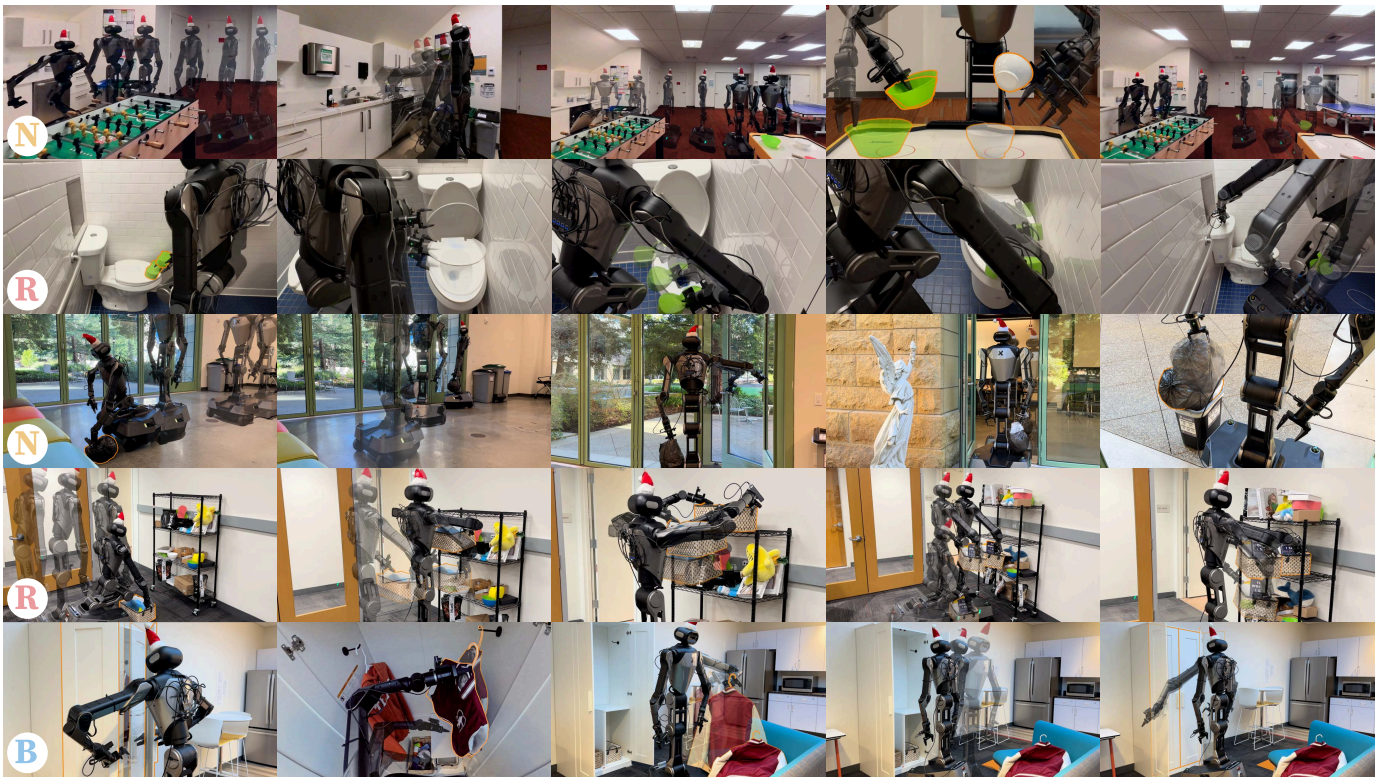


Fig. 1: **Everyday household activities enabled by BEHAVIOR ROBOT SUITE (BRS), showcasing its three core capabilities: bimanual coordination (B), stable and accurate navigation (N), and extensive end-effector reachability (R).** Each row illustrates the rollout trajectory of trained WB-VIMA policies, an imitation learning algorithm we developed, using data collected with JoyLo, our novel whole-body teleoperation interface. While every activity involves multiple capabilities, the most crucial capability for accomplishing each task is highlighted using B, N, and R. Activities from top to bottom are as follows. **1) Clean house after a wild party (N):** The robot navigates to a dishwasher and opens it, then moves to a gaming table to collect bowls. It returns to the dishwasher, places the bowls inside, and closes it. **2) Clean the toilet (R):** The robot picks up a sponge, opens the toilet cover, cleans the seat, then closes the cover and wipes it. Finally, it moves to press the flush button. **3) Take trash outside (N):** The robot navigates to a trash bag in the living room, picks it up, and carries it to a closed door. It opens the door, moves outside, and deposits the trash bag into a trash bin. **4) Put items onto shelves (R):** The robot lifts a box from the ground, moves to a shelf, and places the box on the appropriate level based on available space. **5) Lay clothes out (B):** The robot moves to a wardrobe, opens it, picks up a jacket on a hanger, lays the jacket on a sofa bed, then returns to the wardrobe and closes it.

***Abstract*—Real-world household tasks present significant challenges for mobile manipulation robots. An analysis of existing robotics benchmarks reveals that successful task performance hinges on three key whole-body control capabilities: bimanual coordination, stable and precise navigation, and extensive end-effector reachability. Achieving these capabilities requires careful hardware design, but the resulting system complexity further complicates visuomotor policy learning. To address these challenges, we introduce the BEHAVIOR ROBOT SUITE (BRS), a comprehensive framework for whole-body manipulation in diverse household tasks. Built on a bimanual, wheeled robot with a 4-DoF torso, BRS integrates a cost-effective whole-body teleoperation interface for data collection and a novel algorithm for learning whole-body visuomotor policies. We**

Stanford University

evaluate BRS on five challenging household tasks that not only emphasize the three core capabilities but also introduce additional complexities, such as long-range navigation, interaction with articulated and deformable objects, and manipulation in confined spaces. We believe that BRS's integrated robotic embodiment, data collection interface, and learning framework mark a significant step toward enabling real-world whole-body manipulation for everyday household tasks. BRS is open-sourced at `behavior-robot-suite.github.io`.

I. Introduction

Developing versatile and capable robots that can assist in everyday life remains a major challenge in human-centered robotics research [1–4], with increasing attention on daily household tasks [5–12]. *What key capabilities must a robot develop to achieve all these?* To investigate this question, we analyze activities from BEHAVIOR-1K [8], a human-centered robotics benchmark encompassing 1,000 everyday household tasks, selected and defined by the general public, and instantiated in ecological and virtual environments. Through this analysis, we identify three essential whole-body control capabilities for successfully performing these tasks: **bimanual** coordination, stable and accurate **navigation**, and extensive end-effector **reachability**.

Tasks such as lifting large, heavy objects require **bimanual manipulation** [13, 14], whereas retrieving objects throughout a house depends on stable and precise **navigation** [15–17]. Opening a door while carrying groceries demands the coordination of both capabilities [18–20]. In addition, everyday objects are distributed across diverse locations and heights, requiring robots to adapt their **reach** accordingly. To illustrate this, we analyze the spatial distribution of task-relevant household objects in everyday household tasks and scenes (Fig. 2). Notably, the multi-modal distribution of vertical distances highlights the necessity of extensive end-effector reachability, enabling a robot to interact with objects across a wide range of spatial configurations.

How, then, can a robot effectively achieve these capabilities? Carefully designed robotic hardware incorporating dual arms, a mobile base, and a flexible torso is essential to enable whole-body manipulation [21]. However, such designs introduce significant challenges for policy learning methods, particularly in scaling data collection [22–24] and accurately modeling coordinated whole-body actions. Current systems struggle to address these challenges comprehensively [21, 25–31], highlighting the need for more suitable hardware for household tasks, more efficient data collection tools, and improved models for whole-body control.

We introduce the BEHAVIOR Robot Suite (BRS), a comprehensive framework for learning whole-body manipulation to tackle diverse real-world household tasks (Table I). BRS addresses both hardware and learning challenges through two key innovations. The first is JoyLo, a low-cost, whole-body teleoperation interface designed for general applicability, with a concrete implementation on a wheeled dual-arm manipulator with a flexible torso. The second is the Whole-Body VIsuoMotor Attention (WB-VIMA) policy, a novel learning algorithm that effectively models coordinated whole-body actions.

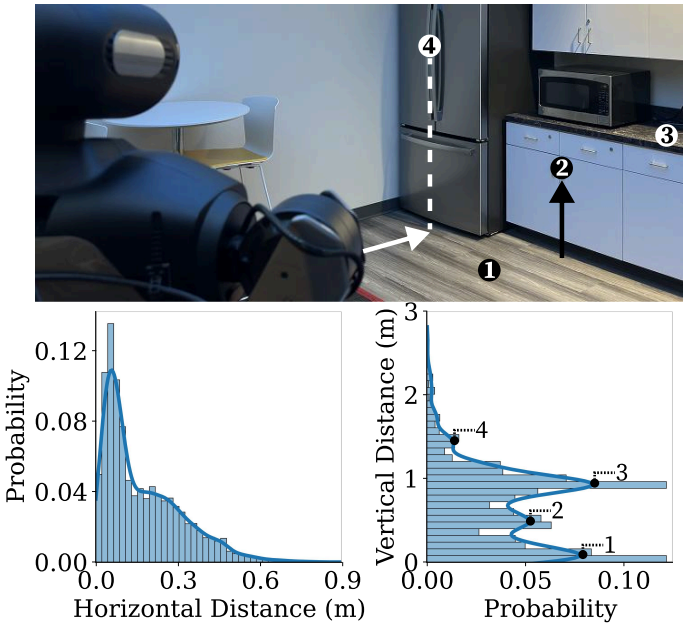


Fig. 2: **Ecological distributions of task-relevant objects in daily household activities.** Multiple distinct modes appear in the vertical distance distribution, located at $0.09\,\mathrm{m}$, $0.49\,\mathrm{m}$, $0.94\,\mathrm{m}$, and $1.43\,\mathrm{m}$, representing heights at which objects are typically found.

We evaluate BRS on five challenging real-world household tasks in unmodified human living environments. The learned WB-VIMA policies demonstrate strong performance, achieving an average success rate of 88% in short-horizon sub-tasks, and a peak success rate of 93% in long-horizon full tasks. We believe that BRS's integrated robotic embodiment, data collection interface, and learning framework mark a significant step toward real-world whole-body manipulation for everyday household tasks. BRS is open-sourced at `behavior-robot-suite.github.io`.

II. JoyLo: Joy-Con on Low-Cost Kinematic-Twin Arms

To enable seamless teleoperation of mobile manipulators with a high degree of freedoms (DoFs) and facilitate data collection for policy learning, we introduce **JoyLo**, a cost-effective whole-body teleoperation interface. As illustrated in Fig. 3, we implement JoyLo on the Galaxea R1 robot, a wheeled dual-arm manipulator with a 4-DoF torso (Appendix A), following design objectives detailed as follow. While we provide one specific instantiation of JoyLo, its design principles are general and can be adapted to similar mobile manipulators.

a) Efficient Whole-Body Control: Whole-body robot teleoperation methods vary widely in accuracy, efficiency, applicability, and user experience. At one extreme, kinesthetic teaching enables precise physical guidance [44–47], but is slow and not easily scalable. At the other extreme, motion retargeting

		BRS (Ours)	Mobile ALOHA [30]	TidyBot++ [32]	ACE [33]	BiDex [34]	HomeRobot [10]	TTT [21]	TeleMoMa [25]	RoboCopilot [35]	Open-TeleVision [36]	TRILL [37]	GR00T N1 [38]	ALOHA [39]	GELLO [40]	HATO [41]	FACTR [42]
		Mobile manipulation									Humanoid manipulation			Stationary manipulation			
Interface	Simultaneous control of arms, torso, and mobile base	✔	✘	✘	✘	✘	✘	✘	✔	✔	✘	✘	✘	✘	✘	✘	✘
	Bimanual control	✔	✔	✘	✔	✔	✘	✘	✔	✔	✔	✔	✔	✔	✔	✔	✔
	Torso control	✔	✘	✘	✘	✘	✘	✘	✔	✔	✘	✘	✘	✘	✘	✘	✘
	Single-operator mobile base control	✔	✔	✔	✔	✘	✘	✘	✔	✔	✘	✔	✘	✘	✘	✘	✘
	Untethered mobile base control	✔	✘	✔	✔	✔	✘	✘	✔	✔	✘	✔	✘	✘	✘	✘	✘
	Haptic feedback	✔	✘	✘	✘	✘	✘	✘	✘	✔	✘	✘	✘	✘	✘	✘	✔
	Cost[a]	$	$$$	$$	$$	$	N.A.	N.A.	$	$$$	$$$	$	$$$	$$$	$	$	$$
Capabilities[b]	Omnidirectional navigation	✔	✘	✔	✘	✔	✘	✔	✔	✔	✘	✘	✘	✘	✘	✘	✘
	Bimanual coordination	✔	✔	✘	✔	✔	✘	✔	✔	✔	✔	✔	✔	✔	✘	✔	✔
	Ground-level reach	✔	✘	✔	✔	✘	✘	✔	✔	✔	✘	✘	✘	✘	✘	✘	✘
	Comfortable overhead reach[c]	✔	✔	✘	✘	✘	✘	✔	✘	✘	✘	✘	✘	✘	✘	✘	✘
	Operation in confined spaces	✔	✘	✔	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘
	Coordinated whole-body manipulation involving hip, waist, and mobile base	✔	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘
Algorithm	Learning-based method	✔	✔	✔	✔	✔	✔	✘	✔	✔	✔	✔	✔	✔	N.A.	✔	✔
	Novel algorithm	✔	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✔	✔	N.A.	✘	✔
	Autoregressive whole-body action prediction	✔	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	N.A.	✘	✘
	Sensory observation modality	**Colored point cloud**	RGB	RGB	RGB	RGB	Depth + semantic seg.	RGB-D	RGB-D	RGB	RGB	RGB	RGB	RGB	N.A.	RGB-D + tactile	RGB
	Policy model backbone[d]	**XF**	XF	UNet	UNet	XF	RNN	N.A.	MLP/RNN	UNet	XF	RNN	XF	XF	N.A.	UNet	XF
	Open-source everything[e]	✔	✔	✔	Hardware + teleop.	Teleop.	✔	✘	Teleop.	✘	✔	✔	Weights + finetuning	✔	✔	✔	✔

[a] Interface hardware cost. **$**: $0 – $500 ; **$$**: $500 – $1000 ; **$$$**: $1000 +.
[b] We consider robot capabilities that are demonstrated by **learned autonomous policies**.
[c] Following Panero and Zelnik [43], we use 182.9 cm (72 in) as the maximum height for comfortable overhead reach.
[d] Neural network architecture of the policy backbone. "XF" stands for Transformer.
[e] Everything includes the interface hardware, teleoperation code, algorithm code, and documentation.

TABLE I: **Comparison of recent real-robot frameworks.** BRS is comprehensive, integrating a unique whole-body control interface JoyLo and a novel algorithm WB-VIMA for learning whole-body visuomotor policies, demonstrating several unprecedented robotic capabilities.

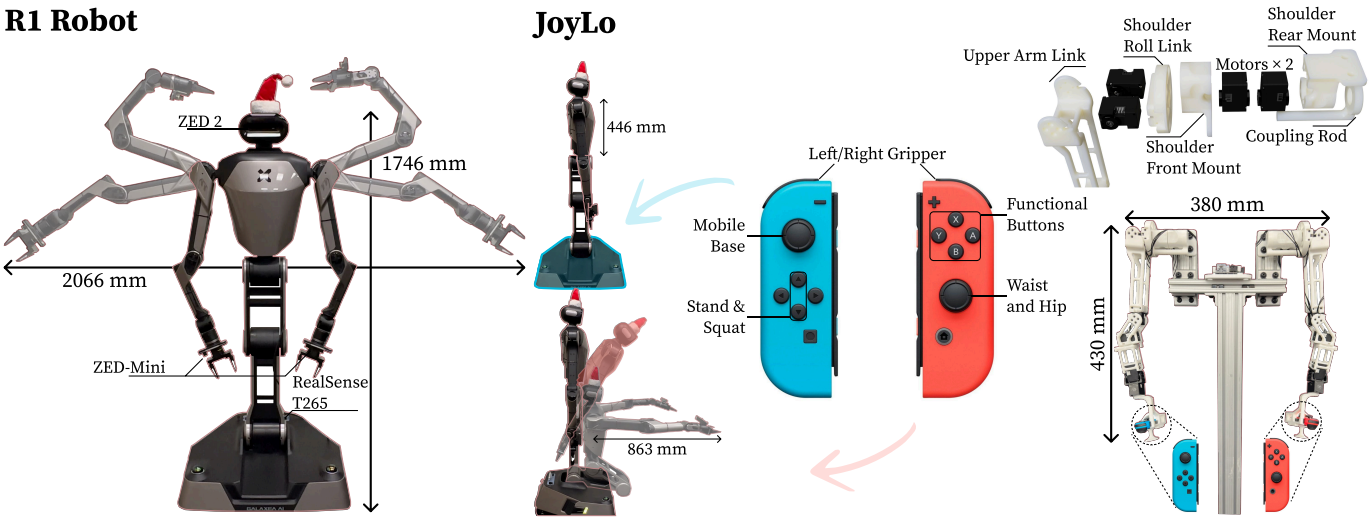


Fig. 3: **BRS hardware system. Left:** The R1 robot with two 6-DoF arms and a 4-DoF torso mounted on an omnidirectional mobile base. **Right:** The JoyLo system, consisting of compact, off-the-shelf Nintendo Joy-Con controllers mounted at the ends of two kinematic-twin arms. Joy-Con serves as the interface for controlling the grippers, torso, and mobile base.

techniques [25, 48–57] remove physical interaction but face embodiment mismatches and limited platform applicability. To balance intuitiveness, ease of use, and precision for manipulation tasks, we propose a puppeteering-based approach using kinematic-twin arms equipped with thumbsticks for torso and mobile base control. Specifically, we utilize off-the-shelf Nintendo Joy-Con controllers due to their compact size, integrated thumbsticks, and multiple functional buttons, which enable rich, customizable functionality. As illustrated in Fig. 3, the left thumbstick controls mobile base velocity; the right thumbstick adjusts waist and hips; arrow keys change torso height; triggers operate the grippers. With JoyLo, users can simultaneously control arm movements, gripper operations, upper-body motions, and mobile base navigation, enabling efficient whole-body control that is accurate, user-friendly, and scalable. Additionally, the kinematic constraints imposed by the leader arms prevent the operator from generating infeasible or undeployable actions, ensuring smooth and reliable demonstrations.

b) Rich User Feedback: JoyLo enhances teleoperation by providing haptic feedback through bilateral teleoperation [58, 59] without extra force sensors [60, 61]. The JoyLo arms, kinematically coupled with the robot arms,

act as leaders issuing commands while being regularized by the robot's joint positions. Let $\mathbf{q}_{\text{JoyLo}}$ and $\mathbf{q}_{\text{robot}}$ be their respective joint positions; the torques τ applied to the JoyLo arms are $\tau = \mathbf{K_p}\left(\mathbf{q}_{\text{robot}} - \mathbf{q}_{\text{JoyLo}}\right) + \mathbf{K_d}\left(\dot{\mathbf{q}}_{\text{robot}} - \dot{\mathbf{q}}_{\text{JoyLo}}\right) - \mathbf{K}$, where $\dot{\mathbf{q}}$ denotes joint velocities, and $\mathbf{K_p}$, $\mathbf{K_d}$, and $\mathbf{K}$ are proportional, derivative, and damping gains. This feedback discourages abrupt user motions and provides proportional resistance when the robot experiences contact.

c) Low Cost and Easy Accessibility: JoyLo is built from 3D-printed links, low-cost Dynamixel motors, and Joy-Con controllers, totaling under $500. Additionally, its modular design ensures that all components are replaceable, minimizing downtime and eliminating unnecessary repair costs. BRS provides an intuitive, real-time controller with Python interfaces for efficient operation.

III. WB-VIMA: Whole-Body VIsuoMotor Attention Policy

This section introduces **WB-VIMA**, a transformer-based model [62, 63] designed to learn coordinated whole-body actions for mobile manipulation tasks. Trained on data collected through JoyLo, it autoregressively decodes whole-body actions across the embodiment space and dynamically aggregates multi-modal observations using self-attention (Fig. 4).

a) Autoregressive Whole-Body Action Decoding: In mobile manipulators with multiple articulated components, small mobile base or torso errors can cause large end-effector deviations. For example, a $0.17\,\text{rad}$ ($10°$) knee movement in the R1 robot's neutral pose (Fig. 3) can shift the end-effector by up to $0.14\,\text{m}$ due to error amplification along the kinematic chain, highlighting the need for precise coordination in whole-body mobile manipulation. To address this issue, we leverage the inherent hierarchy in the robot's embodiment. Specifically, conditioning upper-body action predictions on the predicted lower-body actions enables the policy to better model coordinated whole-body movements. This approach ensures that downstream joints account for upstream motion, reducing error propagation. The whole-body action decoding follows an autoregressive structure: At timestep t, the mobile base trajectory $\mathbf{a}_{\text{base}} \in \mathbb{R}^{T_a \times 3}$ is first predicted using the action readout token $\mathbf{E}^a$ (encoded from observations, detailed later). $\mathbf{a}_{\text{base}}$ and $\mathbf{E}^a$ are then used to predict the torso trajectory $\mathbf{a}_{\text{torso}} \in \mathbb{R}^{T_a \times 4}$. Finally, $\mathbf{a}_{\text{base}}$, $\mathbf{a}_{\text{torso}}$, and $\mathbf{E}^a$ together predict the arms and grippers' trajectory $\mathbf{a}_{\text{arms}} \in \mathbb{R}^{T_a \times 14}$. WB-VIMA jointly learns three independent denoising diffusion networks [64–66] for the mobile base, torso, and arms, denoted ϵ_{base}, ϵ_{torso}, and ϵ_{arms}. Whole-body actions $\mathbf{a}_{\text{whole-body}} \in \mathbb{R}^{T_a \times 21}$ are autoregressively decoded through iterative denoising:

$$
\begin{aligned}
\mathbf{a}_{\text{base}}^{k-1} &\sim \mathcal{N}\left(\mu_k\left(\mathbf{a}_{\text{base}}^{k}, \epsilon_{\text{base}}\left(\mathbf{a}_{\text{base}}^{k} | \mathbf{E}^a, k\right)\right), \sigma_k^2 I\right), \\
\mathbf{a}_{\text{torso}}^{k-1} &\sim \mathcal{N}\left(\mu_k\left(\mathbf{a}_{\text{torso}}^{k}, \epsilon_{\text{torso}}\left(\mathbf{a}_{\text{torso}}^{k} | \mathbf{a}_{\text{base}}^{0}, \mathbf{E}^a, k\right)\right), \sigma_k^2 I\right), \\
\mathbf{a}_{\text{arms}}^{k-1} &\sim \mathcal{N}\left(\mu_k\left(\mathbf{a}_{\text{arms}}^{k}, \epsilon_{\text{arms}}\left(\mathbf{a}_{\text{arms}}^{k} | \mathbf{a}_{\text{torso}}^{0}, \mathbf{a}_{\text{base}}^{0}, \mathbf{E}^a, k\right)\right), \sigma_k^2 I\right).
\end{aligned} \tag{1}
$$

To achieve efficient inference for high-frequency control, only action readout tokens are used for whole-body decoding

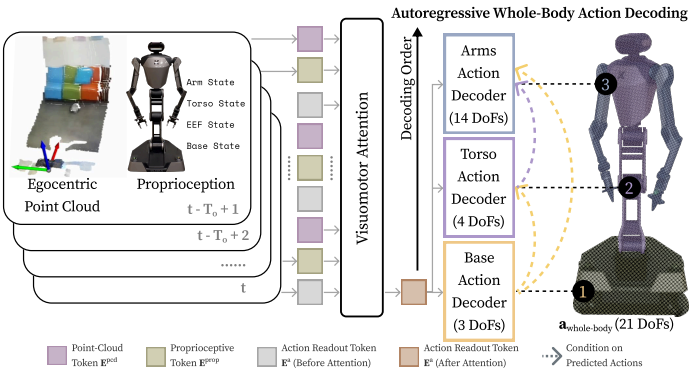


Fig. 4: **WB-VIMA architecture.** It autoregressively decodes whole-body actions by leveraging the hierarchical interdependencies within the embodiment space, and dynamically aggregates multi-modal observations using self-attention.

via diffusion, allowing lightweight UNet-based [67] action heads with a heavier transformer backbone for observation encoding. This balances expressivity and latency.

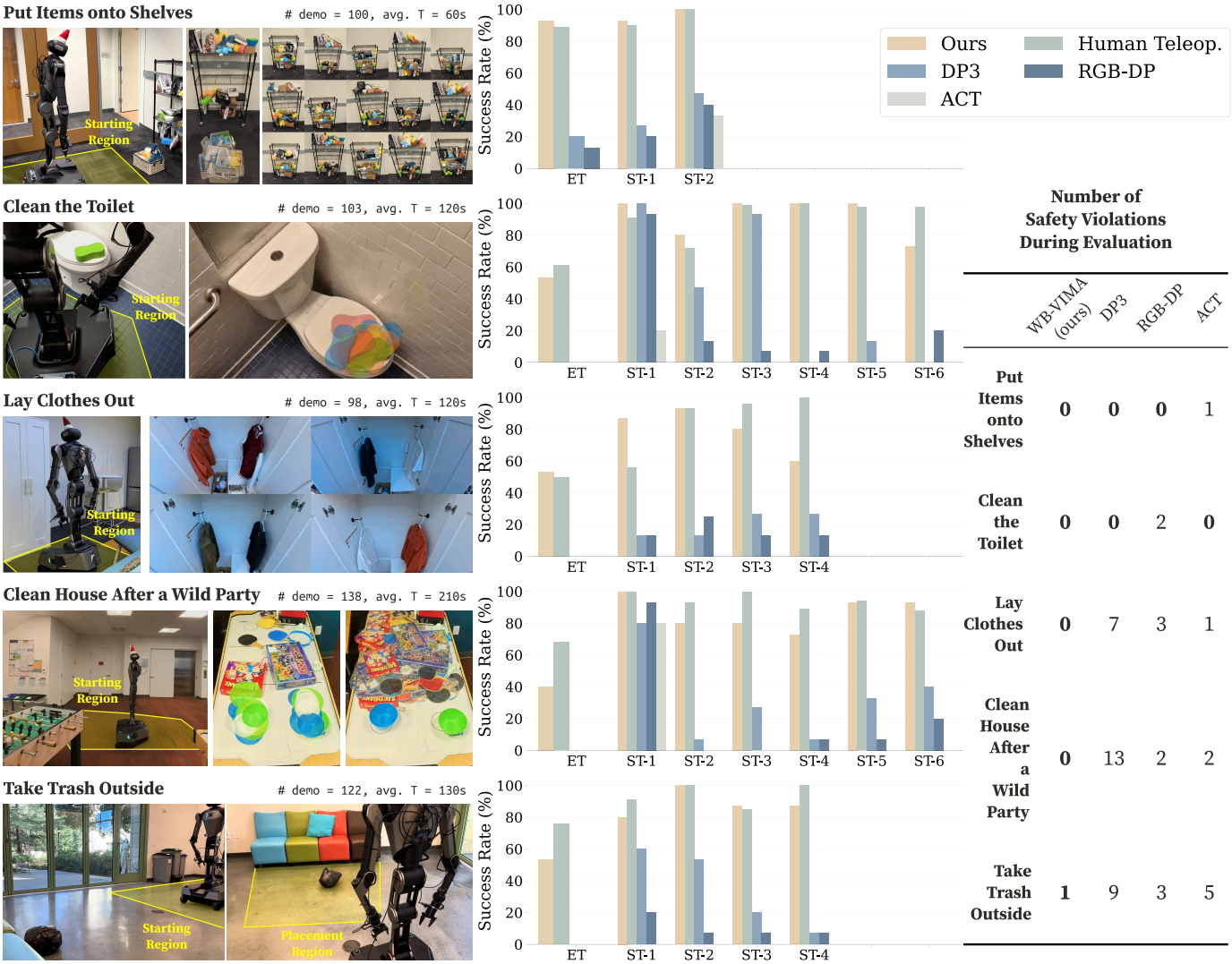
b) Multi-Modal Observation Attention: Observations from multiple modalities are crucial for autonomous robots in complex environments. In WB-VIMA, egocentric colored point clouds and robot proprioception (joint positions and mobile base velocities) are fused via a visuomotor attention network, avoiding overfitting to any single source of information. Concretely, a PointNet [68] encodes the point cloud into a point-cloud token $\mathbf{E}^{\text{pcd}}$, and an MLP encodes proprioception into a proprioceptive token $\mathbf{E}^{\text{prop}}$. Tokens from current and past T_o steps, along with action readout tokens $\mathbf{E}^{\text{a}}$, form a visuomotor sequence: $\mathbf{S} = [\mathbf{E}^{\text{pcd}}_{t-T_o+1}, \mathbf{E}^{\text{prop}}_{t-T_o+1}, \mathbf{E}^{\text{a}}_{t-T_o+1}, \ldots, \mathbf{E}^{\text{pcd}}_{t}, \mathbf{E}^{\text{prop}}_{t}, \mathbf{E}^{\text{a}}_{t}] \in \mathbb{R}^{3T_o \times E}$. $\mathbf{S}$ is then processed through causal self-attention, ensuring action tokens attend only to earlier observations. The final action readout token $\mathbf{E}^a_t$ is used for autoregressive whole-body decoding.

c) Training and Deployment: Following Ho et al. [69], WB-VIMA is trained to predict added noise, minimizing $\mathcal{L} = MSE(\epsilon^k, \epsilon_\theta(\cdot|k))$ for each action decoder, with the total loss aggregated across all three action decoders. Here, ϵ^k and ϵ_θ represent the ground-truth and predicted noise. Deployment uses NVIDIA RTX 4090 GPUs with $0.02\,\text{s}$ effective latency. Data is collected at $10\,\text{Hz}$ with the robot controller running at $100\,\text{Hz}$. A new policy action is issued every $0.1\,\text{s}$ and repeated 10 times.

IV. Experiments

We conduct experiments to answer the following questions.

- $\mathcal{Q}1$: What household tasks are enabled by BRS, and how does WB-VIMA compare to baselines?
- $\mathcal{Q}2$: How different components contribute to WB-VIMA's effectiveness?
- $\mathcal{Q}3$: How does JoyLo compare to other interfaces in efficiency and policy learning suitability?
- $\mathcal{Q}4$: What other insights can be drawn about the system's capabilities?



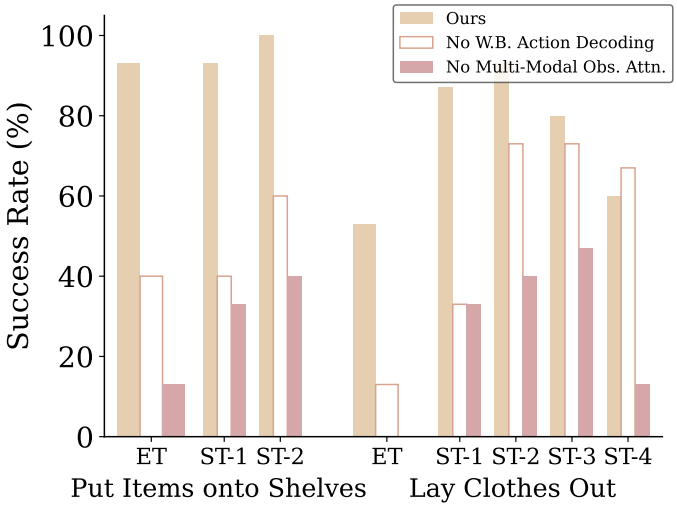
Number of Safety Violations During Evaluation

	WB-VIMA (ours)	DP3	RGB-DP	ACT
Put Items onto Shelves	**0**	**0**	**0**	1
Clean the Toilet	**0**	**0**	2	**0**
Lay Clothes Out	**0**	7	3	1
Clean House After a Wild Party	**0**	13	2	2
Take Trash Outside	**1**	9	3	5

Fig. 5: **Evaluation results for five household tasks. Left:** Initial randomization. **Middle:** Success rates over 15 runs ("ET" = entire task, "ST" = sub-task). **Right:** Number of safety violations.

a) Experiment Settings: We evaluate BRS on five real-world household tasks (see Fig. 1 and Appendix D-A for details), inspired by the everyday activities defined in BEHAVIOR-1K [8]. We collect **100**, **103**, **98**, **138**, and **122** trajectories using JoyLo for these long-horizon tasks, each ranging from 60 s to 210 s. Each task is segmented into multiple sub-tasks (**"ST"**). During evaluation, if a sub-task fails, we reset to the start of the *next* sub-task and *continue* evaluation. We also report the end-to-end success rates for entire tasks (**"ET"**). Baselines include **DP3** [70], **RGB-DP** [65], and **ACT** [39]. We additionally report human teleoperation success and policy safety violations, defined as robot collisions or motor power losses due to excessive force. Each policy is evaluated 15 times with randomized **robot starting position**, **target object placement**, **target object instance**, and **distractors**. Each task covers at least two types of randomization. Task videos are available at `behavior-robot-suite.github.io`.

b) BRS enables various household activities, on which WB-VIMA consistently outperforms baseline methods (Q1): As shown in Fig. 5, WB-VIMA achieves an average sub-task success rate of 88%, and average and peak entire-task success rates of 58% and 93%. On contact-rich sub-tasks involving articulated objects, where human operators often struggle with uncoordinated whole-body motions—such as opening the toilet cover (ST-2) in "clean the toilet" and opening the wardrobe (ST-1) in "lay clothes out"—WB-VIMA even outperforms human teleoperation, suggesting that training on successful demonstrations enables it to learn precise, coordinated maneuvers for reliably completing such tasks. Moreover, WB-VIMA shows an emergent capability for completing long-horizon, multi-stage tasks, enabled by the synergy between its multi-modal observation attention—extracting salient, task-relevant features—and autoregressive whole-body action decoding—generating coherent actions that rarely lead to out-of-distribution states. Finally, WB-VIMA maintains a near-zero safety violation rate, which we attribute



# Safety Violations	Ours	No W.B. Action Decoding	No Multi-Modal Obs. Attn.
Put Items onto Shelves	0	0	0
Lay Clothes Out	0	0	4

Fig. 6: **Real-world ablation results for "put items onto shelves" and "lay clothes out."**

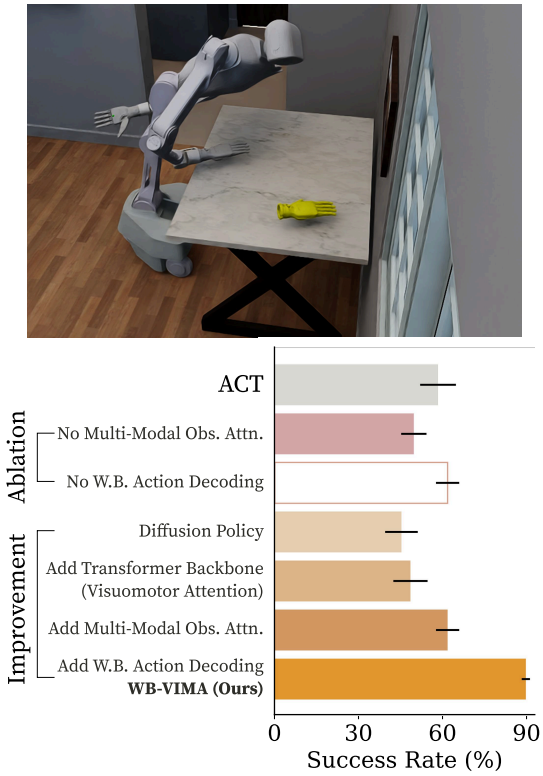


Fig. 7: **Simulation ablation results for "wiping table."** The robot must wipe toward the goal using whole-body motions while maintaining continuous hand contact. Results are averaged over **five** runs with **100** rollouts each; error bars indicate standard deviation.

to its use of colored point-cloud observations that provide explicit 3D perception and semantic understanding, ensuring coordinated actions that inherently respect safety constraints.

For end-to-end task success, WB-VIMA achieves 13× and 21× higher success rates than DP3 and RGB-DP, respectively. For average sub-task performance, it outperforms them by 1.6× and 3.4×. ACT fails to complete any full tasks and rarely succeeds in sub-tasks. These baselines struggle because they directly predict flattened 21-DoF actions, ignoring hierarchical dependencies within the action space. As a result, modeling errors [71] in mobile base or torso predictions cannot be corrected by arm actions, leading to amplified end-effector drift, pushing the robot into out-of-distribution states, and eventually resulting in task failures. Uncoordinated whole-body actions also increase safety violations (Fig. 5), such as DP3 colliding with tables, RGB-DP losing arm power from excessive force, and ACT hitting doorframes during trash disposal. We also observe that WB-VIMA and DP3 outperform RGB-DP and ACT, underscoring the importance of explicit 3D perception in complex environments. Egocentric point clouds provide unified spatial understanding critical for accurate mobile base navigation. While both WB-VIMA and DP3 leverage point clouds, only WB-VIMA incorporates task semantic information through color, whereas DP3 often overfits to proprioception, stitching actions based purely on joint positions without regard to the environment.

*c) Synergistic whole-body action prediction and multi-modal feature extraction are key to WB-VIMA's strong performance ($\mathcal{Q}$**2**):* Can models based solely on explicit 3D perception match WB-VIMA 's performance? Ablation studies show they cannot. We evaluate two WB-VIMA variants: one without **autoregressive whole-body action decoding** and one without **multi-modal observation attention**. As shown in Fig. 6, removing either significantly degrades performance. Tasks like "put items onto shelves" and "open wardrobe" (ST-1) in "lay clothes out" critically depend on coordinated whole-body actions; removing autoregressive action decoding leads to up to a 53% performance drop. Removing multi-modal attention reduces performance across all tasks, causing the model to ignore visual inputs and overfit to proprioception. Four collisions are also observed due to poor visual awareness. The same conclusions hold in a simulated table wiping task (Fig. 7). Furthermore, starting from a vanilla diffusion policy, we provide a roadmap improving the model success by progressively adding components: multi-modal observation attention improves by 27% and surpasses ACT; adding autoregressive whole-body action decoding further boosts success by 45%, culminating in WB-VIMA 's strong final performance.

*d) JoyLo is an efficient, user-friendly interface that provides high-quality data for policy learning ($\mathcal{Q}$**3**):* We conducted a user study with 10 participants to evaluate JoyLo against two IK-based interfaces: **VR controllers** [25] and **Apple Vision Pro** [36, 72]. The study was performed in the OmniGibson simulator [8] on the "clean house after a wild party" task, with randomized interface exposure to eliminate bias. We measured *success rate*, *completion time*, *replay success rate*, and *singularity ratio* across entire tasks and

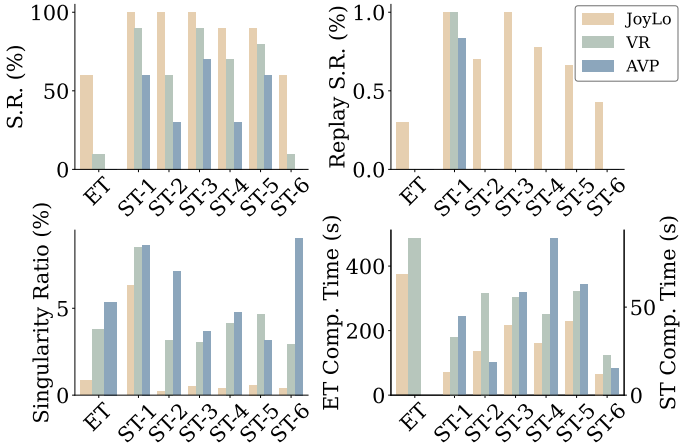


Fig. 8: **User study results.** "S.R." is success rate. "ET Comp. Time" and "ST Comp. Time" refer to entire and sub-task completion times.

sub-tasks. Replay success measures the open-loop execution of collected robot trajectories, where higher values indicate higher-quality, verified data that allows imitation learning policies to better model trajectories [33, 34, 73–75]. Further setup details are provided in Appendix D-D.

As shown in Fig. 8, JoyLo achieves the highest success rate and fastest completion time across all interfaces. It delivers a 5× higher task success rate and 23% shorter median completion time than VR controllers, while no participants completed the entire task with Apple Vision Pro. JoyLo particularly excels at articulated object manipulation (e.g., 67% higher success in "open dishwasher" (ST-2) than VR controllers), enabling users to generate smooth and accurate actions, which is consistent with findings that leader-follower arm control improves fine-grained manipulation [39]. It also significantly reduces sub-task times (e.g., 71% faster navigation and 67% faster bowl picking) compared to Apple Vision Pro, whose reliance on head movement for mobile base control leads to poor coordination and tracking [34]. Moreover, JoyLo provides the highest data quality, achieving the lowest singularity ratio (78% and 85% lower than VR controllers and Apple Vision Pro, respectively) and consistently replaying successful trajectories. Unlike IK-based methods that suffer from suboptimal IK solutions and jerky motions, JoyLo 's direct joint mapping and kinematic-twin arm constraints ensure smooth, stable whole-body teleoperation. In user surveys (Fig. A.4), all participants rated JoyLo the most user-friendly. Although 70% of participants initially believed IK-based interfaces would be more intuitive, after the study they unanimously preferred JoyLo. This shift underscores a key distinction between tabletop data collection and mobile whole-body manipulation: while IK-based methods may suffice for static setups, they struggle to effectively control the mobile base and torso, making high-quality data collection much harder in mobile manipulation settings.

e) Coordinated torso and mobile base movements enhance maneuverability beyond stationary arms (Q4): As shown in Fig. 9, coordinated whole-body movements are critical for tasks involving heavy articulated object interactions, such as "open the door" (ST-3) in "take trash outside" and "open the dishwasher" (ST-2) in "clean house after a wild party." To open a door, the robot bends its hip forward while advancing the base to generate enough inertia; to open a dishwasher, it moves the base backward, using its whole body to pull the door open smoothly. Without hip or base movement, both objects remain closed and the arm joint effort would surge, generating excessive force that is potentially harmful to the hardware. Additional emergent behaviors such as failure recovery are showcased in videos on **behavior-robot-suite.github.io**, demonstrating WB-VIMA 's robustness.

V. Related Work

Robots for Everyday Household Activities Daily household activities have become a major focus for human-centered robotics [1–4, 14], with efforts mainly in: 1) defining benchmarks [5–12, 76–83], and 2) building robotic systems, usually with learning-based methods, to automate tasks [21, 30–32, 34, 84–97]. Unlike field [98], rescue [99], or surgical robots [100], household robots must generalize across diverse, complex home environments. Prior works typically address either data collection or policy learning separately (Table I). In contrast, BRS offers a synergistic framework combining a low-cost, whole-body interface for data collection and a general, competent algorithm for whole-body visuomotor policy learning. Moreover, many household tasks require bimanual coordination and extensive end-effector reachability. Prior systems often rely on a single arm and lifting bodies [26, 80, 92], whereas BRS unleashes the mobile manipulation capabilities to perform broader real-world household tasks.

Low-Cost Hardware for Robot Learning Cost-effective hardware has accelerated robot learning, including: 1) low-cost robots—arms [39], hands [101–103], mobile manipulators [21, 30–32, 84], and humanoids [104–110]; 2) teleoperation interfaces—puppeteering devices [34, 39, 40, 111], exoskeletons [33, 74, 112], and AR/VR devices [25, 36, 113]; and 3) wearable or portable data collection devices [75, 114–119]. Our JoyLo falls under teleoperation interfaces, providing a cost-effective, whole-body solution for mobile, dual-arm robots with torsos. Unlike prior interfaces for stationary arms [40, 74] or mobile bases without independent torso control [30, 34], JoyLo enables efficient, untethered teleoperation of dual-arm mobile manipulators without needing a second operator. Additionally, compared to common puppeteering devices [40], JoyLo offers rich haptic feedback via bilateral teleoperation without requiring force sensors [60, 61] or extra real-robot arms [120].

Learning Whole-Body Manipulation Whole-body manipulation uses the full robot body, including arms [13, 14, 30, 121, 122], torso [123–126], and base [29, 31, 92, 127–132], to interact with objects. Traditional approaches rely on motion planning [97, 124, 125, 133–137], while recent learning-based

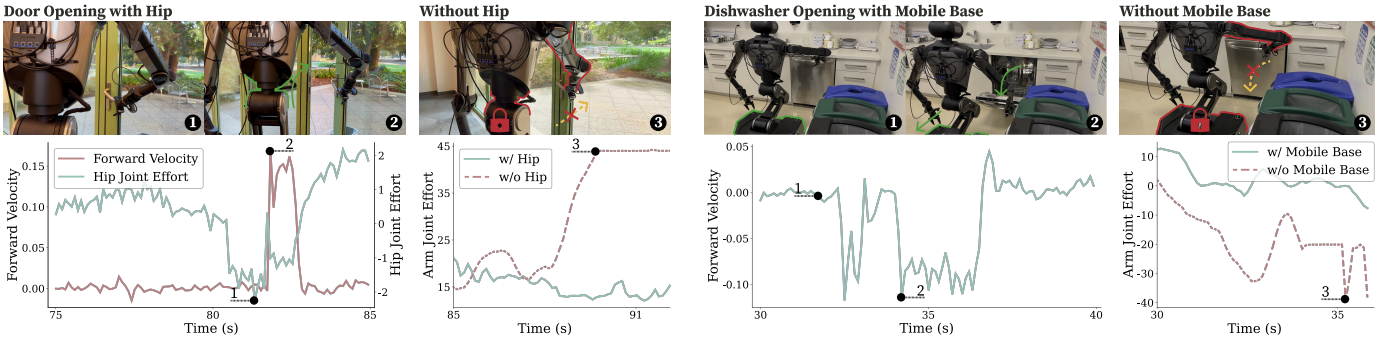


Fig. 9: **Coordinated torso and mobile base movements enhance maneuverability.** WB-VIMA policies use the hip and mobile base to open a door and dishwasher; if the torso or mobile base is locked, opening fails and arm joint effort surges, risking hardware damage.

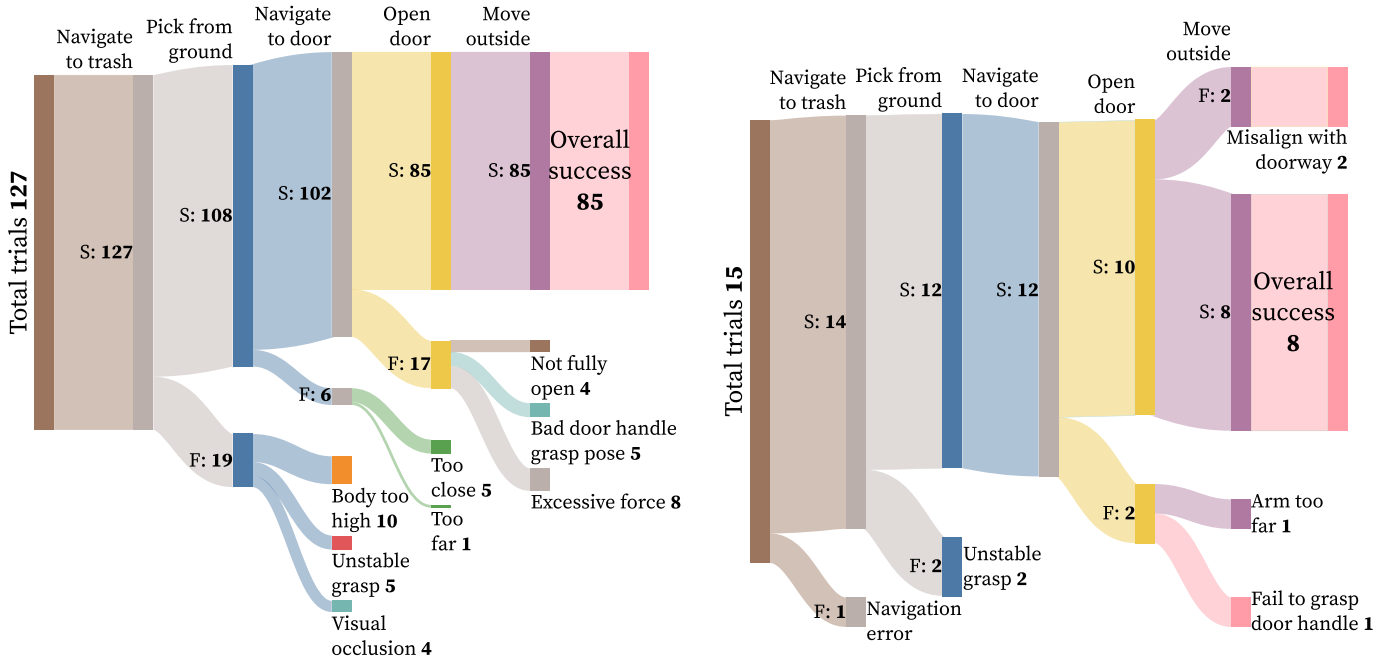


Fig. 10: **Failure modes in the "take trash outside" task. Left:** Failure analysis during data collection using JoyLo. **Right:** Failure analysis during autonomous WB-VIMA policy rollouts. "S" indicates number of successful trials. "F" indicates number of failed trials.

methods use reinforcement learning [27, 29, 30, 92, 127, 129–132, 138–142], behavior cloning [30, 32, 36, 94, 143–147], or large pretrained models [28, 89, 91, 128, 148–150]. Our WB-VIMA introduces a novel algorithm for learning whole-body manipulation on a high-DoF, wheeled, dual-arm robot with a torso. Unlike prior methods that ignore action hierarchy [30, 32, 144] or embodiment interdependencies [27, 130, 139], WB-VIMA explicitly models them through autoregressive whole-body action decoding, enabling coordinated policies for challenging real-world tasks. Additionally, WB-VIMA dynamically fuses multi-modal observations via visuomotor attention, extracting salient task-relevant information, which prior works [94, 131, 146] often neglect.

VI. Limitations

While BRS demonstrates strong performance across real-world household tasks, several limitations remain. In this section, we discuss limiting assumptions, analyze failure modes (Fig. 10), and suggest directions for future work.

a) Mismatched camera field of view between robot and operator.: During data collection with JoyLo, the operator observes the robot from a third-person perspective using their own vision. To collect data efficiently, they must position themselves to maintain a clear view of the workspace without appearing in the robot's field of view. Additionally, the operator must ensure that target objects are visible to the robot's cameras; otherwise, the resulting data will be partially observable, complicating policy training. Future work could

incorporate active perception [36, 151] so that the operator sees exactly what the robot sees.

b) Compounding errors in long-horizon, multi-stage tasks.: In complex tasks like "clean house after a wild party," WB-VIMA experiences compounding errors across multiple sub-tasks and over long horizons. While sub-task success rates remain high, these accumulated errors can significantly reduce overall task success. This limitation could be mitigated by learning on human correction data [35, 71, 93] or integrating model-based task planning [152] to improve robustness over extended execution.

c) Imperfect point cloud observations.: WB-VIMA relies on point cloud data from onboard cameras, which can be degraded by lighting conditions or reflective surfaces. For example, policies trained on data collected during the day may not generalize well to nighttime environments due to visual discrepancies. Since our robot is equipped with stereo cameras, future work could incorporate FoundationStereo [153] to improve point cloud quality.

d) Robot-specific training data.: WB-VIMA is trained on data collected exclusively with the R1 robot. It is intriguing to explore how multi-embodiment data and cross-embodiment transfer can benefit the training [22, 96, 154–156]. The current dataset may also be insufficient for scene-level generalization. Future work could integrate large pre-trained models, such as VLA [157–159], to enhance scene understanding. Finally, it would be valuable to study how whole-body manipulation can benefit from synthetic data [160, 161] or human data [38, 162–164].

VII. Conclusion

This paper presents BRS, a holistic framework for learning whole-body manipulation to tackle diverse real-world household tasks. We identify three core capabilities essential for household activities: **bimanual** coordination, stable **navigation**, and extensive end-effector **reachability**. Achieving these with learning-based methods requires overcoming challenges in both data and modeling. BRS addresses them through two innovations: 1) JoyLo, a cost-effective whole-body interface for efficient data collection, and 2) WB-VIMA, a novel algorithm that leverages embodiment hierarchy and models interdependent whole-body actions. The BRS system demonstrates strong performance across real-world household tasks with unmodified objects in natural, unstructured environments, marking a step toward greater autonomy and reliability in household robotics.

Acknowledgments

We thank Chengshu (Eric) Li, Wenlong Huang, Mengdi Xu, Ajay Mandlekar, Haoyu Xiong, Haochen Shi, Jingyun Yang, Toru Lin, Jim Fan, and the SVL PAIR group for their invaluable technical discussions. We also thank Tianwei Li and the development team at Galaxea.ai for timely hardware support, Yingke Wang for helping with the figures, Helen Roman for processing hardware purchase, Frank Yang, Yihe Tang, Yushan Sun, Chengshu (Eric) Li, Zhenyu Zhang, Haoyu Xiong for participating in user studies, and the Stanford Gates Building community for their patience and support during real-robot experiments. This work is in part supported by the Stanford Institute for Human-Centered AI (HAI), the Schmidt Futures Senior Fellows grant, NSF CCRI #2120095, ONR MURI N00014-21-1-2801, ONR MURI N00014-22-1-2740, and ONR MURI N00014-24-1-2748.

References

[1] M. L. Littman, I. Ajunwa, G. Berger, C. Boutilier, M. Currie, F. Doshi-Velez, G. Hadfield, M. C. Horowitz, C. Isbell, H. Kitano, K. Levy, T. Lyons, M. Mitchell, J. Shah, S. Sloman, S. Vallor, and T. Walsh, "Gathering strength, gathering storms: The one hundred year study on artificial intelligence (ai100) 2021 study panel report," *arXiv preprint arXiv: 2210.15767*, 2022.

[2] M. O. Riedl, "Human-centered artificial intelligence and machine learning," *Human Behavior and Emerging Technologies*, vol. 1, no. 1, pp. 33–36, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/hbe2.117

[3] W. Xu, "Toward human-centered ai: a perspective from human-computer interaction," *Interactions*, vol. 26, no. 4, p. 42–46, Jun. 2019. [Online]. Available: https://doi.org/10.1145/3328485

[4] B. Shneiderman, "Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered ai systems," *ACM Trans. Interact. Intell. Syst.*, vol. 10, no. 4, Oct. 2020. [Online]. Available: https://doi.org/10.1145/3419764

[5] D. Batra, A. X. Chang, S. Chernova, A. J. Davison, J. Deng, V. Koltun, S. Levine, J. Malik, I. Mordatch, R. Mottaghi, M. Savva, and H. Su, "Rearrangement: A challenge for embodied ai," *arXiv preprint arXiv: 2011.01975*, 2020.

[6] S. Srivastava, C. Li, M. Lingelbach, R. Martín-Martín, F. Xia, K. E. Vainio, Z. Lian, C. Gokmen, S. Buch, C. K. Liu, S. Savarese, H. Gweon, J. Wu, and L. Fei-Fei, "BEHAVIOR: benchmark for everyday household activities in virtual, interactive, and ecological environments," in *Conference on Robot Learning, 8-11 November 2021, London, UK*, ser. Proceedings of Machine Learning Research, A. Faust, D. Hsu, and G. Neumann, Eds., vol. 164. PMLR, 2021, pp. 477–490. [Online]. Available: https://proceedings.mlr.press/v164/srivastava22a.html

[7] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. S. Chaplot, O. Maksymets, A. Gokaslan, V. Vondruš, S. Dharur, F. Meier, W. Galuba, A. Chang, Z. Kira, V. Koltun, J. Malik, M. Savva, and D. Batra, "Habitat 2.0: Training home assistants to rearrange their habitat," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 251–266. [Online]. Available:

https://proceedings.neurips.cc/paper_files/paper/2021/file/021bbc7ee20b71134d53e20206bd6feb-Paper.pdf

[8] C. Li, C. Gokmen, G. Levine, R. Martín-Martín, S. Srivastava, C. Wang, J. Wong, R. Zhang, M. Lingelbach, J. Sun, M. Anvari, M. Hwang, M. Sharma, A. Aydin, D. Bansal, S. Hunter, K.-Y. Kim, A. Lou, C. R. Matthews, I. Villa-Renteria, J. H. Tang, C. Tang, F. Xia, S. Savarese, H. Gweon, K. Liu, J. Wu, and L. Fei-Fei, "BEHAVIOR-1k: A benchmark for embodied AI with 1,000 everyday activities and realistic simulation," in *6th Annual Conference on Robot Learning*, 2022. [Online]. Available: https://openreview.net/forum?id=_8DoIe8G3t

[9] M. Heo, Y. Lee, D. Lee, and J. J. Lim, "Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation," in *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, K. E. Bekris, K. Hauser, S. L. Herbert, and J. Yu, Eds., 2023. [Online]. Available: https://doi.org/10.15607/RSS.2023.XIX.041

[10] S. Yenamandra, A. Ramachandran, K. Yadav, A. S. Wang, M. Khanna, T. Gervet, T.-Y. Yang, V. Jain, A. Clegg, J. M. Turner, Z. Kira, M. Savva, A. X. Chang, D. S. Chaplot, D. Batra, R. Mottaghi, Y. Bisk, and C. Paxton, "Homerobot: Open-vocabulary mobile manipulation," in *7th Annual Conference on Robot Learning*, 2023. [Online]. Available: https://openreview.net/forum?id=b-cto-fetlz

[11] A. Shukla, S. Tao, and H. Su, "Maniskill-hab: A benchmark for low-level manipulation in home rearrangement tasks," *arXiv preprint arXiv: 2412.13211*, 2024.

[12] S. Nasiriany, A. Maddukuri, L. Zhang, A. Parikh, A. Lo, A. Joshi, A. Mandlekar, and Y. Zhu, "Robocasa: Large-scale simulation of everyday tasks for generalist robots," *arXiv preprint arXiv: 2406.02523*, 2024.

[13] C. Smith, Y. Karayiannidis, L. Nalpantidis, X. Gratal, P. Qi, D. V. Dimarogonas, and D. Kragic, "Dual arm manipulation—a survey," *Robotics and Autonomous Systems*, vol. 60, no. 10, pp. 1340–1353, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092188901200108X

[14] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, p. eaat8414, 2019. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aat8414

[15] G. Desouza and A. Kak, "Vision for mobile robot navigation: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 237–267, 2002.

[16] T. Kruse, A. K. Pandey, R. Alami, and A. Kirsch, "Human-aware robot navigation: A survey," *Robotics and Autonomous Systems*, vol. 61, no. 12, pp. 1726–1743, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921889013001048

[17] X. Xiao, B. Liu, G. Warnell, and P. Stone, "Motion planning and control for mobile robot navigation using machine learning: a survey," *Autonomous Robots*, vol. 46, pp. 569–597, 2022. [Online]. Available: https://link.springer.com/article/10.1007/s10514-022-10039-8/fulltext.html

[18] L. Peterson, D. Austin, and D. Kragic, "High-level control of a mobile manipulator for door opening," in *Proceedings. 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2000) (Cat. No.00CH37113)*, vol. 3, 2000, pp. 2333–2338 vol.3.

[19] N. Banerjee, X. Long, R. Du, F. Polido, S. Feng, C. G. Atkeson, M. Gennert, and T. Padir, "Human-supervised control of the atlas humanoid robot for traversing doors," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, 2015, pp. 722–729.

[20] M. DeDonato, F. Polido, K. Knoedler, B. P. W. Babu, N. Banerjee, C. P. Bove, X. Cui, R. Du, P. Franklin, J. P. Graff, P. He, A. Jaeger, L. Li, D. Berenson, M. A. Gennert, S. Feng, C. Liu, X. Xinjilefu, J. Kim, C. G. Atkeson, X. Long, and T. Padır, "Team wpi-cmu: Achieving reliable humanoid behavior in the darpa robotics challenge," *Journal of Field Robotics*, vol. 34, no. 2, pp. 381–399, 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21685

[21] M. Bajracharya, J. Borders, R. Cheng, D. Helmick, L. Kaul, D. Kruse, J. Leichty, J. Ma, C. Matl, F. Michel, C. Papazov, J. Petersen, K. Shankar, and M. Tjersland, "Demonstrating mobile manipulation in the wild: A metrics-driven approach," *Robotics: Science and Systems*, 2023. [Online]. Available: https://arxiv.org/abs/2401.01474v1

[22] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, A. Tung, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Gupta, A. Wang, A. Singh, A. Garg, A. Kembhavi, A. Xie, A. Brohan, A. Raffin, A. Sharma, A. Yavary, A. Jain, A. Balakrishna, A. Wahid, B. Burgess-Limerick, B. Kim, B. Schölkopf, B. Wulfe, B. Ichter, C. Lu, C. Xu, C. Le, C. Finn, C. Wang, C. Xu, C. Chi, C. Huang, C. Chan, C. Agia, C. Pan, C. Fu, C. Devin, D. Xu, D. Morton, D. Driess, D. Chen, D. Pathak, D. Shah, D. Büchler, D. Jayaraman, D. Kalashnikov, D. Sadigh, E. Johns, E. Foster, F. Liu, F. Ceola, F. Xia, F. Zhao, F. Stulp, G. Zhou, G. S. Sukhatme, G. Salhotra, G. Yan, G. Feng, G. Schiavi, G. Berseth, G. Kahn, G. Wang, H. Su, H.-S. Fang, H. Shi, H. Bao, H. Ben Amor, H. I. Christensen, H. Furuta, H. Walke, H. Fang, H. Ha, I. Mordatch, I. Radosavovic, I. Leal, J. Liang, J. Abou-Chakra, J. Kim, J. Drake, J. Peters, J. Schneider, J. Hsu, J. Bohg, J. Bingham, J. Wu, J. Gao, J. Hu, J. Wu, J. Wu, J. Sun, J. Luo, J. Gu, J. Tan, J. Oh, J. Wu, J. Lu, J. Yang, J. Malik, J. Silvério, J. Hejna, J. Booher, J. Tompson, J. Yang, J. Salvador, J. J. Lim, J. Han, K. Wang, K. Rao, K. Pertsch, K. Hausman, K. Go, K. Gopalakrishnan,

K. Goldberg, K. Byrne, K. Oslund, K. Kawaharazuka, K. Black, K. Lin, K. Zhang, K. Ehsani, K. Lekkala, K. Ellis, K. Rana, K. Srinivasan, K. Fang, K. P. Singh, K.-H. Zeng, K. Hatch, K. Hsu, L. Itti, L. Y. Chen, L. Pinto, L. Fei-Fei, L. Tan, L. J. Fan, L. Ott, L. Lee, L. Weihs, M. Chen, M. Lepert, M. Memmel, M. Tomizuka, M. Itkina, M. G. Castro, M. Spero, M. Du, M. Ahn, M. C. Yip, M. Zhang, M. Ding, M. Heo, M. K. Srirama, M. Sharma, M. J. Kim, N. Kanazawa, N. Hansen, N. Heess, N. J. Joshi, N. Suenderhauf, N. Liu, N. Di Palo, N. M. M. Shafiullah, O. Mees, O. Kroemer, O. Bastani, P. R. Sanketi, P. T. Miller, P. Yin, P. Wohlhart, P. Xu, P. D. Fagan, P. Mitrano, P. Sermanet, P. Abbeel, P. Sundaresan, Q. Chen, Q. Vuong, R. Rafailov, R. Tian, R. Doshi, R. Martín-Martín, R. Baijal, R. Scalise, R. Hendrix, R. Lin, R. Qian, R. Zhang, R. Mendonca, R. Shah, R. Hoque, R. Julian, S. Bustamante, S. Kirmani, S. Levine, S. Lin, S. Moore, S. Bahl, S. Dass, S. Sonawani, S. Song, S. Xu, S. Haldar, S. Karamcheti, S. Adebola, S. Guist, S. Nasiriany, S. Schaal, S. Welker, S. Tian, S. Ramamoorthy, S. Dasari, S. Belkhale, S. Park, S. Nair, S. Mirchandani, T. Osa, T. Gupta, T. Harada, T. Matsushima, T. Xiao, T. Kollar, T. Yu, T. Ding, T. Davchev, T. Z. Zhao, T. Armstrong, T. Darrell, T. Chung, V. Jain, V. Vanhoucke, W. Zhan, W. Zhou, W. Burgard, X. Chen, X. Wang, X. Zhu, X. Geng, X. Liu, X. Liangwei, X. Li, Y. Lu, Y. J. Ma, Y. Kim, Y. Chebotar, Y. Zhou, Y. Zhu, Y. Wu, Y. Xu, Y. Wang, Y. Bisk, Y. Cho, Y. Lee, Y. Cui, Y. Cao, Y.-H. Wu, Y. Tang, Y. Zhu, Y. Zhang, Y. Jiang, Y. Li, Y. Li, Y. Iwasawa, Y. Matsuo, Z. Ma, Z. Xu, Z. J. Cui, Z. Zhang, and Z. Lin, "Open x-embodiment: Robotic learning datasets and rt-x models : Open x-embodiment collaboration0," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 6892–6903.

[23] H. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, T. Zhao, P. Hansen-Estruch, Q. Vuong, A. W. He, V. Myers, K. Fang, C. Finn, and S. Levine, "Bridgedata v2: A dataset for robot learning at scale," *Conference on Robot Learning*, 2023.

[24] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, P. D. Fagan, J. Hejna, M. Itkina, M. Lepert, Y. J. Ma, P. T. Miller, J. Wu, S. Belkhale, S. Dass, H. Ha, A. Jain, A. Lee, Y. Lee, M. Memmel, S. Park, I. Radosavovic, K. Wang, A. Zhan, K. Black, C. Chi, K. B. Hatch, S. Lin, J. Lu, J. Mercat, A. Rehman, P. R. Sanketi, A. Sharma, C. Simpson, Q. Vuong, H. R. Walke, B. Wulfe, T. Xiao, J. H. Yang, A. Yavary, T. Z. Zhao, C. Agia, R. Baijal, M. G. Castro, D. Chen, Q. Chen, T. Chung, J. Drake, E. P. Foster, J. Gao, D. A. Herrera, M. Heo, K. Hsu, J. Hu, D. Jackson, C. Le, Y. Li, K. Lin, R. Lin, Z. Ma, A. Maddukuri, S. Mirchandani, D. Morton, T. Nguyen, A. O'Neill, R. Scalise, D. Seale, V. Son, S. Tian, E. Tran, A. E. Wang, Y. Wu, A. Xie, J. Yang, P. Yin, Y. Zhang, O. Bastani, G. Berseth, J. Bohg, K. Goldberg, A. Gupta, A. Gupta, D. Jayaraman, J. J. Lim, J. Malik, R. Martín-Martín, S. Ramamoorthy, D. Sadigh, S. Song, J. Wu, M. C. Yip, Y. Zhu, T. Kollar, S. Levine, and C. Finn, "Droid: A large-scale in-the-wild robot manipulation dataset," *Robotics: Science and Systems*, 2024.

[25] 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," *arXiv preprint arXiv: 2403.07869*, 2024.

[26] N. M. M. Shafiullah, A. Rai, H. Etukuru, Y. Liu, I. Misra, S. Chintala, and L. Pinto, "On bringing robots home," *arXiv preprint arXiv: 2311.16098*, 2023. [Online]. Available: https://arxiv.org/abs/2311.16098v1

[27] J. Hu, P. Stone, and R. Mart'in-Mart'in, "Causal policy gradient for whole-body mobile manipulation," *Robotics: Science and Systems*, 2023. [Online]. Available: https://arxiv.org/abs/2305.04866v4

[28] Z. Jiang, Y. Xie, J. Li, Y. Yuan, Y. Zhu, and Y. Zhu, "Harmon: Whole-body motion generation of humanoid robots from language descriptions," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=UUZ4Yw3lt0

[29] S. Uppal, A. Agarwal, H. Xiong, K. Shaw, and D. Pathak, "Spin: Simultaneous perception, interaction and navigation," *Computer Vision and Pattern Recognition*, 2024. [Online]. Available: https://arxiv.org/abs/2405.07991v1

[30] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," *arXiv preprint arXiv: 2401.02117*, 2024. [Online]. Available: https://arxiv.org/abs/2401.02117v1

[31] H. Xiong, R. Mendonca, K. Shaw, and D. Pathak, "Adaptive mobile manipulation for articulated objects in the open world," *arXiv preprint arXiv: 2401.14403*, 2024.

[32] 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," *arXiv preprint arXiv: 2412.10447*, 2024. [Online]. Available: https://arxiv.org/abs/2412.10447v1

[33] S. Yang, M. Liu, Y. Qin, R. Ding, J. Li, X. Cheng, R. Yang, S. Yi, and X. Wang, "Ace: A cross-platform visual-exoskeletons system for low-cost dexterous teleoperation," *arXiv preprint arXiv: 2408.11805*, 2024.

[34] K. Shaw, Y. Li, J. Yang, M. K. Srirama, R. Liu, H. Xiong, R. Mendonca, and D. Pathak, "Bimanual dexterity for complex tasks," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=55tYfHvanf

[35] P. Wu, Y. Shentu, Q. Liao, D. Jin, M. Guo, K. Sreenath, X. Lin, and P. Abbeel, "Robocopilot: Human-in-the-loop interactive imitation learning for robot manipulation," *arXiv preprint arXiv: 2503.07771*, 2025.

[36] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang, "Open-television: Teleoperation with immersive active visual feedback," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=Yce2jeILGt

[37] M. Seo, S. Han, K. Sim, S. H. Bang, C. Gonzalez, L. Sentis, and Y. Zhu, "Deep imitation learning for humanoid loco-manipulation through human teleoperation," in *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2023.

[38] NVIDIA, J. Bjorck, F. Castañeda, N. Cherniadev, 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," *arXiv preprint arXiv: 2503.14734*, 2025.

[39] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv: 2304.13705*, 2023. [Online]. Available: https://arxiv.org/abs/2304.13705v1

[40] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel, "Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators," *IEEE/RJS International Conference on Intelligent RObots and Systems*, 2023.

[41] T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, and J. Malik, "Learning visuotactile skills with two multifingered hands," *arXiv preprint arXiv: 2404.16823*, 2024.

[42] J. J. Liu, Y. Li, K. Shaw, T. Tao, R. Salakhutdinov, and D. Pathak, "Factr: Force-attending curriculum training for contact-rich policy learning," *arXiv preprint arXiv: 2502.17432*, 2025.

[43] J. Panero and M. Zelnik, *Human Dimension and Interior Space: A Source Book of Design Reference Standards*. Clarkson Potter/Ten Speed, 2014. [Online]. Available: https://books.google.com/books?id=VaN_AQAAQBAJ

[44] S. C. Petar Kormushev and D. G. Caldwell, "Imitation learning of positional and force skills demonstrated via kinesthetic teaching and haptic input," *Advanced Robotics*, vol. 25, no. 5, pp. 581–603, 2011. [Online]. Available: https://doi.org/10.1163/016918611X558261

[45] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent advances in robot learning from demonstration," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, no. Volume 3, 2020, pp. 297–330, 2020. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev-control-100819-063206

[46] S. Wrede, C. Emmerich, R. Grünberg, A. Nordmann, A. Swadzba, and J. Steil, "A user study on kinesthetic teaching of redundant robots in task and configuration space," *J. Hum.-Robot Interact.*, vol. 2, no. 1, p. 56–81, Feb. 2013. [Online]. Available: https://doi.org/10.5898/JHRI.2.1.Wrede

[47] M. Hagenow, D. Kontogiorgos, Y. Wang, and J. Shah, "Versatile demonstration interface: Toward more flexible robot demonstration collection," *arXiv preprint arXiv: 2410.19141*, 2024. [Online]. Available: https://arxiv.org/abs/2410.19141v1

[48] A. Setapen, M. Quinlan, and P. Stone, "Marionet: motion acquisition for robots through iterative online evaluative training," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1*, ser. AAMAS '10. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2010, p. 1435–1436.

[49] C. Stanton, A. Bogdanovych, and E. Ratanasena, "Teleoperation of a humanoid robot using full-body motion capture, example movements, and machine learning," in *Australasian Conference on Robotics and Automation, ACRA*, 12 2012.

[50] M. Arduengo, A. Arduengo, A. Colomé, J. Lobo-Prat, and C. Torras, "Human to robot whole-body motion transfer," in *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*, 2021, pp. 299–305.

[51] D. Antotsiou, G. Garcia-Hernando, and T.-K. Kim, "Task-oriented hand motion retargeting for dexterous manipulation imitation," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[52] S. Li, X. Ma, H. Liang, M. Görner, P. Ruppel, B. Fang, F. Sun, and J. Zhang, "Vision-based teleoperation of shadow dexterous hand using end-to-end deep neural network," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 416–422.

[53] J. Liang, A. Handa, K. V. Wyk, V. Makoviychuk, O. Kroemer, and D. Fox, "In-hand object pose tracking via contact feedback and gpu-accelerated robotic simulation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 6203–6209.

[54] A. Handa, K. Van Wyk, W. Yang, J. Liang, Y.-W. Chao, Q. Wan, S. Birchfield, N. Ratliff, and D. Fox, "Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 9164–9170.

[55] Y. Qin, H. Su, and X. Wang, "From one hand to multiple hands: Imitation learning for dexterous manipulation from single-camera teleoperation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 873–10 881, 2022.

[56] A. Sivakumar, K. Shaw, and D. Pathak, "Robotic

telekinesis: Learning a robotic hand imitator by watching humans on youtube," *Robotics: Science and Systems*, 2022. [Online]. Available: https://arxiv.org/abs/2202.10448v2

[57] Y. Qin, W. Yang, B. Huang, K. V. Wyk, H. Su, X. Wang, Y.-W. Chao, and D. Fox, "Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system," *Robotics: Science and Systems*, 2023. [Online]. Available: https://arxiv.org/abs/2307.04577v3

[58] B. Hannaford, "A design framework for teleoperators with kinesthetic feedback," *IEEE Transactions on Robotics and Automation*, vol. 5, no. 4, pp. 426–434, 1989.

[59] D. Lawrence, "Stability and transparency in bilateral teleoperation," *IEEE Transactions on Robotics and Automation*, vol. 9, no. 5, pp. 624–637, 1993.

[60] G. Brantner and O. Khatib, "Controlling ocean one: Human–robot collaboration for deep-sea manipulation," *Journal of Field Robotics*, vol. 38, no. 1, pp. 28–51, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21960

[61] H. Li and K. Kawashima, "Bilateral teleoperation with delayed force feedback using time domain passivity controller," *Robotics and Computer-Integrated Manufacturing*, vol. 37, pp. 188–196, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0736584515000654

[62] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Neural Information Processing Systems*, 2017. [Online]. Available: https://arxiv.org/abs/1706.03762v7

[63] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan, "Vima: General robot manipulation with multimodal prompts," *arXiv preprint arXiv: 2210.03094*, 2022.

[64] Z. Wang, J. J. Hunt, and M. Zhou, "Diffusion policies as an expressive policy class for offline reinforcement learning," *International Conference on Learning Representations*, 2022.

[65] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023. [Online]. Available: https://arxiv.org/abs/2303.04137v5

[66] Z. Wang, Z. Li, A. Mandlekar, Z. Xu, J. Fan, Y. Narang, L. Fan, Y. Zhu, Y. Balaji, M. Zhou, M.-Y. Liu, and Y. Zeng, "One-step diffusion policy: Fast visuomotor policies via diffusion distillation," *arXiv preprint arXiv: 2410.21257*, 2024. [Online]. Available: https://arxiv.org/abs/2410.21257v1

[67] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.

[68] C. Qi, H. Su, K. Mo, and L. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," *Computer Vision and Pattern Recognition*, 2016.

[69] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

[70] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," *arXiv preprint arXiv: 2403.03954*, 2024. [Online]. Available: https://arxiv.org/abs/2403.03954v7

[71] S. Ross, G. J. Gordon, and J. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," *International Conference on Artificial Intelligence and Statistics*, 2010.

[72] Y. Park and P. Agrawal, "Using apple vision pro to train and control robots," 2024. [Online]. Available: https://github.com/Improbable-AI/VisionProTeleop

[73] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Mart'in-Mart'in, "What matters in learning from offline human demonstrations for robot manipulation," *Conference on Robot Learning*, 2021.

[74] H. Fang, H. Fang, Y. Wang, J. Ren, J. Chen, R. Zhang, W. Wang, and C. Lu, "Airexo: Low-cost exoskeletons for learning whole-arm manipulation in the wild," *IEEE International Conference on Robotics and Automation*, 2023. [Online]. Available: https://arxiv.org/abs/2309.14975v2

[75] 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," *arXiv preprint arXiv: 2410.08464*, 2024.

[76] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, M. Deitke, K. Ehsani, D. Gordon, Y. Zhu, A. Kembhavi, A. Gupta, and A. Farhadi, "Ai2-thor: An interactive 3d environment for visual ai," *arXiv preprint arXiv: 1712.05474*, 2017.

[77] X. Puig, K. Ra, M. Boben, J. Li, T. Wang, S. Fidler, and A. Torralba, "Virtualhome: Simulating household activities via programs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Puig_VirtualHome_Simulating_Household_CVPR_2018_paper.html

[78] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox, "Alfred: A benchmark for interpreting grounded instructions for everyday tasks," in *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Shridhar_ALFRED_A_Benchmark_for_Interpreting_Grounded_Instructions_for_Everyday_Tasks_CVPR_2020_paper.html

[79] M. Shridhar, X. Yuan, M.-A. Côté, Y. Bisk, A. Trischler, and M. J. Hausknecht, "Alfworld: Aligning text and embodied environments for interactive learning," *International Conference on Learning Representations*, 2020.

[80] J. Pari, N. M. M. Shafiullah, S. P. Arunachalam, and L. Pinto, "The surprising effectiveness of representation learning for visual imitation," *Robotics: Science and Systems*, 2021.

[81] K. Ehsani, W. Han, A. Herrasti, E. VanderBilt, L. Weihs, E. Kolve, A. Kembhavi, and R. Mottaghi, "Manipulathor: A framework for visual object manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 4497–4506.

[82] L. Weihs, M. Deitke, A. Kembhavi, and R. Mottaghi, "Visual room rearrangement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 5922–5931.

[83] C. Li, F. Xia, R. Mart'in-Mart'in, M. Lingelbach, S. Srivastava, B. Shen, K. Vainio, C. Gokmen, G. Dharan, T. Jain, A. Kurenkov, K. Liu, H. Gweon, J. Wu, L. Fei-Fei, and S. Savarese, "igibson 2.0: Object-centric simulation for robot learning of everyday household tasks," *Conference on Robot Learning*, 2021.

[84] M. Bajracharya, J. Borders, D. Helmick, T. Kollar, M. Laskey, J. Leichty, J. Ma, U. Nagarajan, A. Ochiai, J. Petersen, K. Shankar, K. Stone, and Y. Takaoka, "A mobile manipulation system for one-shot teaching of complex tasks in homes," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 11 039–11 045.

[85] S. Bahl, A. Gupta, and D. Pathak, "Human-to-robot imitation in the wild," *Robotics: Science and Systems*, 2022.

[86] N. Abdo, C. Stachniss, L. Spinello, and W. Burgard, "Robot, organize my shelves! tidying up objects by predicting user preferences," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 1557–1564.

[87] C. Wang, L. J. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu, and A. Anandkumar, "Mimicplay: Long-horizon imitation learning by watching human play," *Conference on Robot Learning*, 2023. [Online]. Available: https://arxiv.org/abs/2302.12422v2

[88] R. Zhang, S. Lee, M. Hwang, A. Hiranaka, C. Wang, W. Ai, J. J. R. Tan, S. Gupta, Y. Hao, G. Levine, R. Gao, A. Norcia, F.-F. Li, and J. Wu, "Noir: Neural signal operated intelligent robots for everyday activities," *Conference on Robot Learning*, 2023. [Online]. Available: https://arxiv.org/abs/2311.01454v1

[89] J. Wu, R. Antonova, A. Kan, M. Lepert, A. Zeng, S. Song, J. Bohg, S. Rusinkiewicz, and T. Funkhouser, "Tidybot: personalized robot assistance with large language models," *Autonomous Robots*, vol. 47, pp. 1087–1102, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s10514-023-10139-z/fulltext.html

[90] H. Shi, H. Xu, S. Clarke, Y. Li, and J. Wu, "Robocook: Long-horizon elasto-plastic object manipulation with diverse tools," *Conference on Robot Learning*, 2023. [Online]. Available: https://arxiv.org/abs/2306.14447v2

[91] A. Stone, T. Xiao, Y. Lu, K. Gopalakrishnan, K.-H. Lee, Q. Vuong, P. Wohlhart, S. Kirmani, B. Zitkovich, F. Xia, C. Finn, and K. Hausman, "Open-world object manipulation using pre-trained vision-language models," in *7th Annual Conference on Robot Learning*, 2023. [Online]. Available: https://openreview.net/forum?id=9al6taqfTzr

[92] R. Yang, Y. Kim, A. Kembhavi, X. Wang, and K. Ehsani, "Harmonic mobile manipulation," *IEEE/RJS International Conference on Intelligent RObots and Systems*, 2023.

[93] Y. Jiang, C. Wang, R. Zhang, J. Wu, and L. Fei-Fei, "Transic: Sim-to-real policy transfer by learning from online correction," *arXiv preprint arXiv: 2405.10315*, 2024. [Online]. Available: https://arxiv.org/abs/2405.10315v3

[94] J. Yang, Z. ang Cao, C. Deng, R. Antonova, S. Song, and J. Bohg, "Equibot: Sim(3)-equivariant diffusion policy for generalizable and data efficient learning," *arXiv preprint arXiv: 2407.01479*, 2024.

[95] T. Dai, J. Wong, Y. Jiang, C. Wang, C. Gokmen, R. Zhang, J. Wu, and L. Fei-Fei, "Automated creation of digital cousins for robust policy learning," *arXiv preprint arXiv: 2410.07408*, 2024. [Online]. Available: https://arxiv.org/abs/2410.07408v3

[96] 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," *arXiv preprint arXiv: 2410.24164*, 2024.

[97] C.-C. Hsu, B. Abbatematteo, Z. Jiang, Y. Zhu, R. Martín-Martín, and J. Biswas, "Kinscene: Model-based mobile manipulation of articulated scenes," *arXiv preprint arXiv: 2409.16473*, 2024. [Online]. Available: https://arxiv.org/abs/2409.16473v2

[98] R. Shamshiri, C. Weltzien, I. Hameed, I. Yule, T. Grift, S. Balasundram, L. Pitonakova, D. Ahmad, and G. Chowdhary, "Research and development in agricultural robotics: A perspective of digital farming," *International Journal of Agricultural and Biological Engineering*, vol. 11, pp. 1–14, 07 2018.

[99] J. Delmerico, S. Mintchev, A. Giusti, B. Gromov, K. Melo, T. Horvat, C. Cadena, M. Hutter, A. Ijspeert,

D. Floreano, L. M. Gambardella, R. Siegwart, and D. Scaramuzza, "The current state and future outlook of rescue robotics," *Journal of Field Robotics*, vol. 36, no. 7, pp. 1171–1191, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21887

[100] P. Gomes, "Surgical robotics: Reviewing the past, analysing the present, imagining the future," *Robotics and Computer-Integrated Manufacturing*, vol. 27, no. 2, pp. 261–266, 2011, translational Research – Where Engineering Meets Medicine. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0736584510000608

[101] R. Ma and A. Dollar, "Yale openhand project: Optimizing open-source hand designs for ease of fabrication and adoption," *IEEE Robotics & Automation Magazine*, vol. 24, no. 1, pp. 32–40, 2017.

[102] K. Shaw, A. Agarwal, and D. Pathak, "Leap hand: Low-cost, efficient, and anthropomorphic hand for robot learning," *Robotics: Science and Systems*, 2023.

[103] K. Shaw and D. Pathak, "LEAP hand v2: Dexterous, low-cost anthropomorphic hybrid rigid soft hand for robot learning," in *2nd Workshop on Dexterous Manipulation: Design, Perception and Control (RSS)*, 2024. [Online]. Available: https://openreview.net/forum?id=eQomRzRZEP

[104] D. Gouaillier, V. Hugel, P. Blazevic, C. Kilner, J. Monceaux, P. Lafourcade, B. Marnier, J. Serre, and B. Maisonnier, "Mechatronic design of nao humanoid," in *2009 IEEE International Conference on Robotics and Automation*, 2009, pp. 769–774.

[105] J. Englsberger, A. Werner, C. Ott, B. Henze, M. A. Roa, G. Garofalo, R. Burger, A. Beyer, O. Eiberger, K. Schmid, and A. Albu-Schäffer, "Overview of the torque-controlled humanoid robot toro," *2014 IEEE-RAS International Conference on Humanoid Robots*, pp. 916–923, 2014. [Online]. Available: https://ieeexplore.ieee.org/document/7041473

[106] P. Seiwald, S.-C. Wu, F. Sygulla, T. F. C. Berninger, N.-S. Staufenberg, M. F. Sattler, N. Neuburger, D. Rixen, and F. Tombari, "Lola v1.1 – an upgrade in hardware and software design for dynamic multi-contact locomotion," in *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*, 2021, pp. 9–16.

[107] N. G. Tsagarakis, D. G. Caldwell, F. Negrello, W. Choi, L. Baccelliere, V. Loc, J. Noorden, L. Muratore, A. Margan, A. Cardellino, L. Natale, E. Mingo Hoffman, H. Dallali, N. Kashiri, J. Malzahn, J. Lee, P. Kryczka, D. Kanoulas, M. Garabini, M. Catalano, M. Ferrati, V. Varricchio, L. Pallottino, C. Pavan, A. Bicchi, A. Settimi, A. Rocchi, and A. Ajoudani, "Walk-man: A high-performance humanoid platform for realistic environments," *Journal of Field Robotics*, vol. 34, no. 7, pp. 1225–1259, 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21702

[108] A. SaLoutos, E. Stanger-Jones, Y. Ding, M. Chignoli, and S. Kim, "Design and development of the mit humanoid: A dynamic and robust research platform," in *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, 2023, pp. 1–8.

[109] Q. Liao, B. Zhang, X. Huang, X. Huang, Z. Li, and K. Sreenath, "Berkeley humanoid: A research platform for learning-based control," *arXiv preprint arXiv: 2407.21781*, 2024. [Online]. Available: https://arxiv.org/abs/2407.21781v1

[110] H. Shi, W. Wang, S. Song, and C. K. Liu, "Toddlerbot: Open-source ml-compatible humanoid platform for loco-manipulation," *arXiv preprint arXiv: 2502.00893*, 2025.

[111] Z. Si, K. Zhang, F. Z. Temel, and O. Kroemer, "Tilde: Teleoperation for dexterous in-hand manipulation learning with a deltahand," *ROBOTICS*, 2024. [Online]. Available: https://arxiv.org/abs/2405.18804v2

[112] Y. Ishiguro, T. Makabe, Y. Nagamatsu, Y. Kojio, K. Kojima, F. Sugai, Y. Kakiuchi, K. Okada, and M. Inaba, "Bilateral humanoid teleoperation system using whole-body exoskeleton cockpit tablis," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6419–6426, 2020.

[113] A. Iyer, Z. Peng, Y. Dai, I. Guzey, S. Haldar, S. Chintala, and L. Pinto, "OPEN TEACH: A versatile teleoperation system for robotic manipulation," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=cvAIaS6V2I

[114] S. Song, A. Zeng, J. Lee, and T. Funkhouser, "Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4978–4985, 2020.

[115] S. Young, D. Gandhi, S. Tulsiani, A. Gupta, P. Abbeel, and L. Pinto, "Visual imitation made easy," in *Proceedings of the 2020 Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Kober, F. Ramos, and C. Tomlin, Eds., vol. 155. PMLR, 16–18 Nov 2021, pp. 1992–2005. [Online]. Available: https://proceedings.mlr.press/v155/young21a.html

[116] F. Sanches, G. Gao, N. Elangovan, R. V. Godoy, J. Chapman, K. Wang, P. Jarvis, and M. Liarokapis, "Scalable. intuitive human to robot skill transfer with wearable human machine interfaces: On complex, dexterous tasks," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 6318–6325.

[117] C. Chi, Z. Xu, C. Pan, E. A. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," *ROBOTICS*, 2024.

[118] C. Wang, H. Shi, W. Wang, R. Zhang, F.-F. Li, and K. Liu, "Dexcap: Scalable and portable mocap data collection system for dexterous manipulation," *ROBOTICS*, 2024.

[119] M. Seo, H. A. Park, S. Yuan, Y. Zhu, and L. Sentis,

"Legato: Cross-embodiment imitation using a grasping tool," *arXiv preprint arXiv: 2411.03682*, 2024.

[120] M. Shridhar, Y. L. Lo, and S. James, "Generative image as action models," *arXiv preprint arXiv: 2407.07875*, 2024. [Online]. Available: https://arxiv.org/abs/2407.07875v2

[121] N. Vahrenkamp, M. Przybylski, T. Asfour, and R. Dillmann, "Bimanual grasp planning," *2011 11th IEEE-RAS International Conference on Humanoid Robots*, pp. 493–499, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:14784225

[122] J. Grannen, Y. Wu, B. Vu, and D. Sadigh, "Stabilize to act: Learning to coordinate for bimanual manipulation," in *7th Annual Conference on Robot Learning*, 2023. [Online]. Available: https://openreview.net/forum?id=86aMPJn6hX9F

[123] K. Harada and M. Kaneko, "Whole body manipulation," in *IEEE International Conference on Robotics, Intelligent Systems and Signal Processing, 2003. Proceedings. 2003*, vol. 1, 2003, pp. 190–195 vol.1.

[124] F. Burget, A. Hornung, and M. Bennewitz, "Whole-body motion planning for manipulation of articulated objects," in *2013 IEEE International Conference on Robotics and Automation*, 2013, pp. 1656–1662.

[125] A. Dietrich, T. Wimbock, A. Albu-Schaffer, and G. Hirzinger, "Reactive whole-body control: Dynamic mobile manipulation using a large number of actuated degrees of freedom," *IEEE Robotics & Automation Magazine*, vol. 19, no. 2, pp. 20–33, 2012.

[126] X. Xu, D. Bauer, and S. Song, "Robopanoptes: The all-seeing robot with whole-body dexterity," *arXiv preprint arXiv: 2501.05420*, 2025.

[127] F. Xia, C. Li, R. Martín-Martín, O. Litany, A. Toshev, and S. Savarese, "Relmogen: Integrating motion generation in reinforcement learning for mobile manipulation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 4583–4590.

[128] R. Shah, A. Yu, Y. Zhu, Y. Zhu*, and R. Martín-Martín*, "Bumble: Unifying reasoning and acting with vision-language models for building-wide mobile manipulation," *arXiv preprint*, 2024.

[129] N. Yokoyama, A. Clegg, J. Truong, E. Undersander, T.-Y. Yang, S. Arnaud, S. Ha, D. Batra, and A. Rai, "Asc: Adaptive skill coordination for robotic mobile manipulation," *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 779–786, 2024.

[130] Z. Fu, X. Cheng, and D. Pathak, "Deep whole-body control: Learning a unified policy for manipulation and locomotion," in *Proceedings of The 6th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, K. Liu, D. Kulic, and J. Ichnowski, Eds., vol. 205. PMLR, 14–18 Dec 2023, pp. 138–149. [Online]. Available: https://proceedings.mlr.press/v205/fu23a.html

[131] M. Liu, Z. Chen, X. Cheng, Y. Ji, R.-Z. Qiu, R. Yang, and X. Wang, "Visual whole-body control for legged loco-manipulation," *arXiv preprint arXiv: 2403.16967*, 2024. [Online]. Available: https://arxiv.org/abs/2403.16967v5

[132] H. Ha, Y. Gao, Z. Fu, J. Tan, and S. Song, "Umi on legs: Making manipulation policies mobile with manipulation-centric whole-body controllers," *arXiv preprint arXiv: 2407.10353*, 2024.

[133] Y. Yamamoto and X. Yun, "Coordinating locomotion and manipulation of a mobile manipulator," in *[1992] Proceedings of the 31st IEEE Conference on Decision and Control*, 1992, pp. 2643–2648 vol.3.

[134] L. P. Kaelbling and T. Lozano-Pérez, "Integrated task and motion planning in belief space," *The International Journal of Robotics Research*, vol. 32, no. 9-10, pp. 1194–1227, 2013. [Online]. Available: https://doi.org/10.1177/0278364913484072

[135] Q. Huang, K. Tanie, and S. Sugano, "Coordinated motion planning for a mobile manipulator considering stability and manipulation," *The International Journal of Robotics Research*, vol. 19, no. 8, pp. 732–742, 2000. [Online]. Available: https://doi.org/10.1177/02783640022067139

[136] L. Sentis and O. Khatib, "A whole-body control framework for humanoids operating in human environments," in *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, 2006, pp. 2641–2648.

[137] H. Dai, A. Valenzuela, and R. Tedrake, "Whole-body motion planning with centroidal dynamics and full kinematics," in *2014 IEEE-RAS International Conference on Humanoid Robots*, 2014, pp. 295–302.

[138] D. Honerkamp, T. Welschehold, and A. Valada, "N^2m^2: Learning navigation for arbitrary mobile manipulation motions in unseen and dynamic environments," *IEEE Transactions on robotics*, 2022.

[139] G. Pan, Q. Ben, Z. Yuan, G. Jiang, Y. Ji, S. Li, J. Pang, H. Liu, and H. Xu, "Roboduet: Whole-body legged loco-manipulation with cross-embodiment deployment," *arXiv preprint arXiv: 2403.17367*, 2024. [Online]. Available: https://arxiv.org/abs/2403.17367v4

[140] P. Arm, M. Mittal, H. Kolvenbach, and M. Hutter, "Pedipulate: Enabling manipulation skills using a quadruped robot's leg," *IEEE International Conference on Robotics and Automation*, 2024. [Online]. Available: https://arxiv.org/abs/2402.10837v1

[141] X. He, C. Yuan, W. Zhou, R. Yang, D. Held, and X. Wang, "Visual manipulation with legs," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=E4K3yLQQ7s

[142] C. Zhang, W. Xiao, T. He, and G. Shi, "Wococo: Learning whole-body humanoid control with sequential contacts," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=Czs2xH9114

[143] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar,

J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. J. Joshi, R. C. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K.-H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. Ryoo, G. Salazar, P. R. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "Rt-1: Robotics transformer for real-world control at scale," *Robotics: Science and Systems*, 2022. [Online]. Available: https://arxiv.org/abs/2212.06817v2

[144] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "Humanplus: Humanoid shadowing and imitation from humans," *arXiv preprint arXiv: 2406.10454*, 2024. [Online]. Available: https://arxiv.org/abs/2406.10454v1

[145] J. Li, Y. Zhu, Y. Xie, Z. Jiang, M. Seo, G. Pavlakos, and Y. Zhu, "OKAMI: Teaching humanoid robots manipulation skills through single video imitation," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=URj5TQTAXM

[146] Y. Ze, Z. Chen, W. Wang, T. Chen, X. He, Y. Yuan, X. B. Peng, and J. Wu, "Generalizable humanoid manipulation with improved 3d diffusion policies," *arXiv preprint arXiv:2410.10803*, 2024.

[147] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. M. Kitani, C. Liu, and G. Shi, "Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=oL1WEZQal8

[148] B. Ichter, A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian, D. Kalashnikov, S. Levine, Y. Lu, C. Parada, K. Rao, P. Sermanet, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, M. Yan, N. Brown, M. Ahn, O. Cortes, N. Sievers, C. Tan, S. Xu, D. Reyes, J. Rettinghouse, J. Quiambao, P. Pastor, L. Luu, K. Lee, Y. Kuang, S. Jesmonth, N. J. Joshi, K. Jeffrey, R. J. Ruano, J. Hsu, K. Gopalakrishnan, B. David, A. Zeng, and C. K. Fu, "Do as I can, not as I say: Grounding language in robotic affordances," in *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, ser. Proceedings of Machine Learning Research, K. Liu, D. Kulic, and J. Ichnowski, Eds., vol. 205. PMLR, 2022, pp. 287–318. [Online]. Available: https://proceedings.mlr.press/v205/ichter23a.html

[149] M. Xu, P. Huang, W. Yu, S. Liu, X. Zhang, Y. Niu, T. Zhang, F. Xia, J. Tan, and D. Zhao, "Creative robot tool use with large language models," *arXiv preprint arXiv: 2310.13065*, 2023.

[150] Q. Wu, Z. Fu, X. Cheng, X. Wang, and C. Finn, "Helpful doggybot: Open-world object fetching using legged robots and vision-language models," in *arXiv*, 2024.

[151] R. Bajcsy, "Active perception," *Proceedings of the IEEE*, vol. 76, no. 8, pp. 966–1005, 1988.

[152] W. Liu, N. Nie, R. Zhang, J. Mao, and J. Wu, "Learning compositional behaviors from demonstration and language," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=fR1rCXjCQX

[153] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, "Foundationstereo: Zero-shot stereo matching," *arXiv preprint arXiv: 2501.09898*, 2025.

[154] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, J. Luo, Y. L. Tan, P. R. Sanketi, Q. Vuong, T. Xiao, D. Sadigh, C. Finn, and S. Levine, "Octo: An open-source generalist robot policy," *ROBOTICS*, 2024. [Online]. Available: https://arxiv.org/abs/2405.12213v2

[155] J. Yang, C. Glossop, A. Bhorkar, D. Shah, Q. Vuong, C. Finn, D. Sadigh, and S. Levine, "Pushing the limits of cross-embodiment learning for manipulation and navigation," *Robotics: Science and Systems*, 2024. [Online]. Available: https://arxiv.org/abs/2402.19432v1

[156] R. Doshi, H. R. Walke, O. Mees, S. Dasari, and S. Levine, "Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=AuJnXGq3AL

[157] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, K. Choromanski, T. Ding, D. Driess, K. A. Dubey, C. Finn, P. R. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. J. Joshi, R. C. Julian, D. Kalashnikov, Y. Kuang, I. Leal, S. Levine, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. R. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, T. Xiao, T. Yu, and B. Zitkovich, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *Conference on Robot Learning*, 2023. [Online]. Available: https://arxiv.org/abs/2307.15818v1

[158] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, "OpenVLA: An open-source vision-language-action model," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=ZMnD6QZAE6

[159] Z. Xu, H.-T. L. Chiang, Z. Fu, M. G. Jacob, T. Zhang, T.-W. E. Lee, W. Yu, C. Schenck, D. Rendleman, D. Shah, F. Xia, J. Hsu, J. Hoech, P. Florence, S. Kirmani, S. Singh, V. Sindhwani,

C. Parada, C. Finn, P. Xu, S. Levine, and J. Tan, "Mobility VLA: Multimodal instruction navigation with long-context VLMs and topological graphs," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=JScswMfEQ0

[160] A. Mandlekar, S. Nasiriany, B. Wen, I. Akinola, Y. Narang, L. Fan, Y. Zhu, and D. Fox, "Mimicgen: A data generation system for scalable robot learning using human demonstrations," *arXiv preprint arXiv: 2310.17596*, 2023. [Online]. Available: https://arxiv.org/abs/2310.17596v1

[161] C. R. Garrett, A. Mandlekar, B. Wen, and D. Fox, "Skillmimicgen: Automated demonstration generation for efficient skill learning and deployment," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=YOFrRTDC6d

[162] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu, "Egomimic: Scaling imitation learning via egocentric video," *arXiv preprint arXiv: 2410.24221*, 2024.

[163] G. Papagiannis, N. D. Palo, P. Vitiello, and E. Johns, "R+x: Retrieval and execution from everyday human videos," *arXiv preprint arXiv: 2407.12957*, 2024.

[164] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote, E. Byrne, Z. Chavis, J. Chen, F. Cheng, F.-J. Chu, S. Crane, A. Dasgupta, J. Dong, M. Escobar, C. Forigua, A. Gebreselasie, S. Haresh, J. Huang, M. M. Islam, S. Jain, R. Khirodkar, D. Kukreja, K. J. Liang, J.-W. Liu, S. Majumder, Y. Mao, M. Martin, E. Mavroudi, T. Nagarajan, F. Ragusa, S. K. Ramakrishnan, L. Seminara, A. Somayazulu, Y. Song, S. Su, Z. Xue, E. Zhang, J. Zhang, A. Castillo, C. Chen, X. Fu, R. Furuta, C. Gonzalez, P. Gupta, J. Hu, Y. Huang, Y. Huang, W. Khoo, A. Kumar, R. Kuo, S. Lakhavani, M. Liu, M. Luo, Z. Luo, B. Meredith, A. Miller, O. Oguntola, X. Pan, P. Peng, S. Pramanick, M. Ramazanova, F. Ryan, W. Shan, K. Somasundaram, C. Song, A. Southerland, M. Tateno, H. Wang, Y. Wang, T. Yagi, M. Yan, X. Yang, Z. Yu, S. C. Zha, C. Zhao, Z. Zhao, Z. Zhu, J. Zhuo, P. Arbelaez, G. Bertasius, D. Damen, J. Engel, G. M. Farinella, A. Furnari, B. Ghanem, J. Hoffman, C. Jawahar, R. Newcombe, H. S. Park, J. M. Rehg, Y. Sato, M. Savva, J. Shi, M. Z. Shou, and M. Wray, "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 (CVPR)*, June 2024, pp. 19 383–19 400.

[165] T. F. Gonzalez, "Clustering to minimize the maximum intercluster distance," *Theoretical Computer Science*, vol. 38, pp. 293–306, 1985. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0304397585902245

[166] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[167] M. Han, L. Wang, L. Xiao, H. Zhang, C. Zhang, X. Xu, and J. Zhu, "Quickfps: Architecture and algorithm co-design for farthest point sampling in large-scale point clouds," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2023.

[168] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. The MIT Press, 2018. [Online]. Available: http://incompleteideas.net/book/the-book-2nd.html

[169] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8162–8171. [Online]. Available: https://proceedings.mlr.press/v139/nichol21a.html

[170] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 2256–2265. [Online]. Available: https://proceedings.mlr.press/v37/sohl-dickstein15.html

[171] N. Shazeer, "Glu variants improve transformer," *arXiv preprint arXiv: 2002.05202*, 2020.

[172] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Computer Vision and Pattern Recognition*, 2015.

[173] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *International Conference on Learning Representations*, 2017.

[174] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *International Conference on Learning Representations*, 2020.

[175] B. Sundaralingam, S. K. S. Hari, A. Fishman, C. Garrett, K. V. Wyk, V. Blukis, A. Millane, H. Oleynikova, A. Handa, F. Ramos, N. Ratliff, and D. Fox, "curobo: Parallelized collision-free minimum-jerk robot motion generation," *arXiv preprint arXiv: 2310.17274*, 2023.

APPENDIX A
ROBOT HARDWARE DETAILS

This section provides additional hardware details, including robot specifications, onboard sensors and computing, and the communication scheme.

A. Robot Platform

We select the Galaxea R1 robot as our platform to meet the three critical capabilities essential for household tasks: **bimanual** coordination, stable and precise **navigation**, and extensive end-effector **reachability**. As illustrated in Fig.3, the R1 robot features two 6-DoF arms mounted on a 4-DoF torso. Each arm is equipped with a parallel jaw gripper and has a maximum payload of $5\,\mathrm{kg}$[1], making it well-suited for manipulating most objects encountered in daily household activities. The torso incorporates four revolute joints: two for waist rotation and hip bending, and two additional joints enabling knee-like motions. This design allows the robot to transition smoothly between standing and squatting positions, enhancing its reachability in household environments. By integrating the torso into the kinematic chain of the end-effectors, the R1 robot achieves an effective reach range from ground level to $2\,\mathrm{m}$ vertically and up to $2.06\,\mathrm{m}$ horizontally, covering the workspace shown in Fig. 2. The arms and torso are controlled using joint impedance controllers, with target joint positions as inputs.

To ensure stable navigation in household environments, the robot's torso is mounted on an omnidirectional mobile base, capable of moving in any direction on the ground plane at a maximum speed of $1.5\,\mathrm{m\,s^{-1}}$. Additionally, the base can independently execute yaw rotations at a maximum angular speed of $3\,\mathrm{rad\,s^{-1}}$. This mobility is powered by three wheel motors and three steering motors. With a $30\,\mathrm{mm}$ ground clearance, the mobile base can traverse most household terrains. It also achieves horizontal accelerations of up to $2.5\,\mathrm{m\,s^{-2}}$, enhancing maneuverability for tasks that require simultaneous movement and manipulation, such as opening doors (Fig. 9). The mobile base is controlled via velocity commands corresponding to its three degrees of freedom on the ground plane: forward motion, lateral motion, and yaw rotation.

For perception, we equip the R1 robot with a suite of onboard sensors, including a stereo ZED 2 RGB-D camera as the head camera, two stereo ZED-Mini RGB-D cameras as wrist cameras, and a RealSense T265 tracking camera for visual odometry. All RGB-D cameras operate at $60\,\mathrm{Hz}$, streaming rectified RGB and depth images. The cameras' poses are updated at $500\,\mathrm{Hz}$ via the robot's forward kinematics, enabling the effective fusion of sensory data from all three cameras. This integration supports high-fidelity global and ego-centric 3D perception, such as colored point-cloud observations. Simultaneously, the visual odometry system operates at $200\,\mathrm{Hz}$, providing real-time velocity and acceleration estimates of the mobile base, which is critical feedback for learning precise velocity control for the mobile base.

[1]All numbers related to the robot's hardware capabilities are based on our testing.

B. Hardware Specifications

1) Arms: The Galaxea R1 robot has two 6-DoF arms, each equipped with a parallel jaw gripper. As shown in Fig. A.1a, each arm has a $128\,\mathrm{mm}$ width and a $923\,\mathrm{mm}$ full reach. The arms are mirrored on the robot and are controlled via a joint impedance controller, receiving target joint positions as inputs. We set the following impedance gains: $\mathbf{K_p} = [140, 200, 120, 20, 20, 20]$ and $\mathbf{K_d} = [10, 50, 5, 1, 1, 0.4]$. Each gripper has a stroke range from $0\,\mathrm{mm}$ (fully closed) to $100\,\mathrm{mm}$ (fully open), with a rated gripping force of $100\,\mathrm{N}$. The grippers are controlled by specifying a target opening width, which is converted into the required motor current.

2) Torso: The torso consists of four revolute joints: two joints for waist rotation and hip bending, and two additional joints for knee-like motions. As shown in Fig. A.1b, the torso has a $340\,\mathrm{mm}$ width and a $1223\,\mathrm{mm}$ height (excluding the head) when fully extended. Table A.I lists the motor specifications.

TABLE A.I: **Torso motor specifications.**

Parameter	Value
Waist Joint Range (Yaw)	$\pm\,3.05\,\mathrm{rad}\,(175^\circ)$
Hip Joint Range (Pitch)	$-2.09\,\mathrm{rad}\,(-120^\circ) \sim 1.83\,\mathrm{rad}\,(105^\circ)$
Knee Joint 1 Range	$-2.79\,\mathrm{rad}\,(-160^\circ) \sim 2.53\,\mathrm{rad}\,(145^\circ)$
Knee Joint 2 Range	$-1.13\,\mathrm{rad}\,(-65^\circ) \sim 1.83\,\mathrm{rad}\,(105^\circ)$
Rated Motor Torque	$108\,\mathrm{N\,m}$
Maximum Motor Torque	$304\,\mathrm{N\,m}$

3) Mobile Base: As illustrated in Fig. A.1c, the mobile base is wheeled and omnidirectional, equipped with three steering motors and three wheel motors. The base can move in any direction on the ground plane and perform yaw rotations. It is controlled via a velocity controller with 3-DoF inputs corresponding to forward velocity (x-axis), lateral velocity (y-axis), and rotation velocity (z-axis). Performance parameters are listed in Table A.II.

TABLE A.II: **Mobile base specifications.**

Parameter	Value
Forward Velocity Limit	$\pm\,1.5\,\mathrm{m\,s^{-1}}$
Lateral Velocity Limit	$\pm\,1.5\,\mathrm{m\,s^{-1}}$
Yaw Rotation Velocity Limit	$\pm\,3\,\mathrm{rad\,s^{-1}}$
Forward Acceleration Limit	$\pm\,2.5\,\mathrm{m\,s^{-2}}$
Lateral Acceleration Limit	$\pm\,1.0\,\mathrm{m\,s^{-2}}$
Yaw Rotation Acceleration Limit	$\pm\,1.0\,\mathrm{rad\,s^{-2}}$

C. Onboard Sensors and Computing

As shown in Fig. 3, the robot is equipped with several onboard sensors: a ZED 2 RGB-D camera (head camera), two ZED-Mini RGB-D cameras (wrist cameras), and a RealSense T265 tracking camera (visual odometry). Camera configurations are provided in Table A.III.

The three RGB-D cameras stream colored point clouds at $60\,\mathrm{Hz}$, obtained from rectified RGB images and aligned depth images. These point clouds are fused into a common robot base frame. For each point cloud in the camera frame $\mathbf{P}^{camera}$,

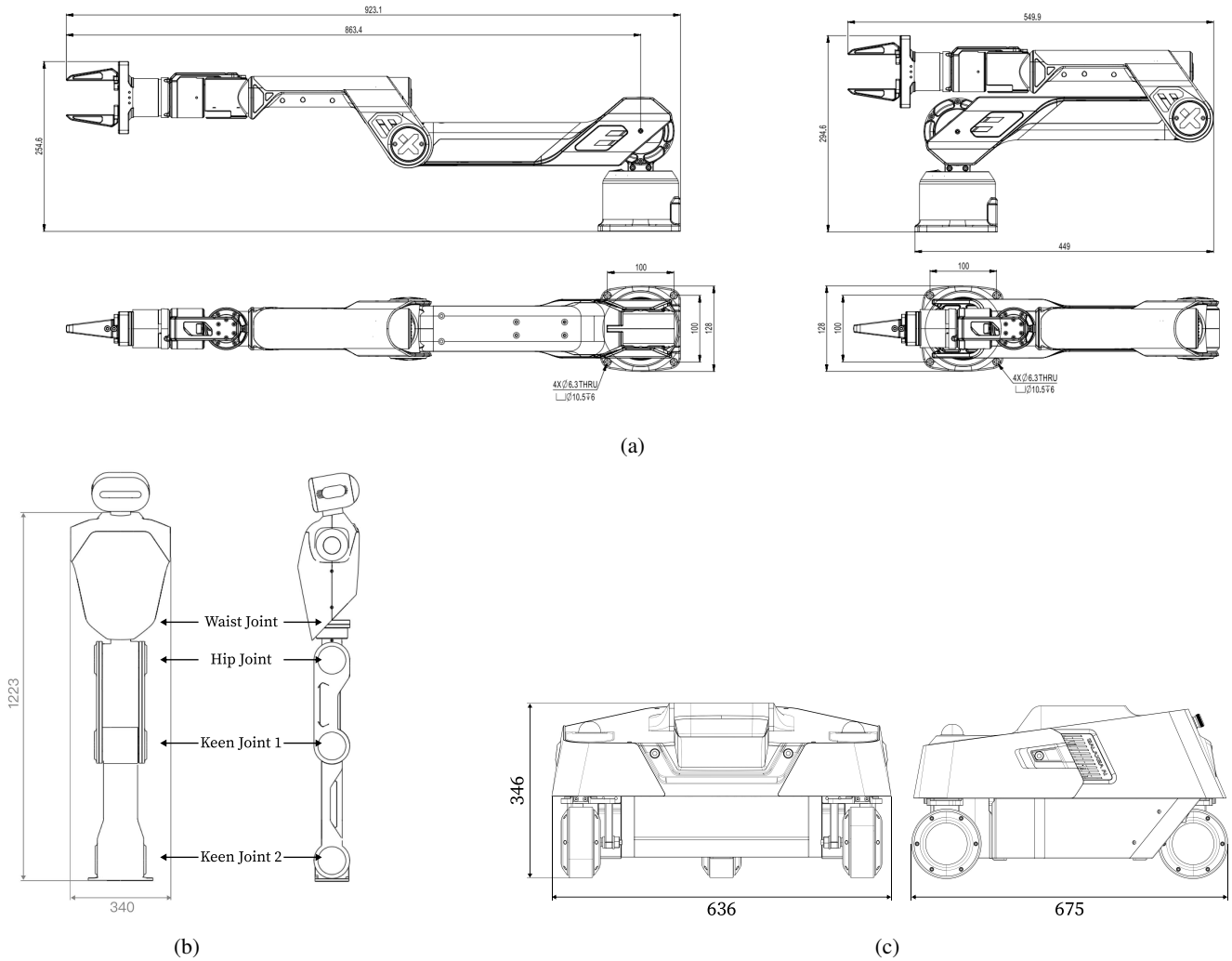


Fig. A.1: **Robot diagrams. (a):** Each arm has six DoFs and a parallel jaw gripper. **(b):** The torso features four revolute joints for waist rotation, hip bending, and knee-like motions. **(c):** The wheeled, omnidirectional mobile base is equipped with three steering motors and three wheel motors.

TABLE A.III: **Configurations for the ZED RGB-D cameras and RealSense T265 tracking camera.**

Parameter	Value
RGB-D Cameras	
Frequency	60 Hz
Image Resolution	1344×376
ZED Depth Mode	PERFORMANCE
Head Camera Min Depth	0.2
Head Camera Max Depth	3
Wrist Camera Min Depth	0.1
Wrist Camera Max Depth	1
Tracking Camera	
Odometry Frequency	200 Hz

where $camera \in$ all cameras $= \{\text{head}, \text{left wrist}, \text{right wrist}\}$, the transformation from the robot base frame to camera frames is computed using forward kinematics at 500 Hz. Denote rotation matrices as $\mathbf{R}^{camera} \in \mathbb{R}^{3\times3}$ and translations as $\mathbf{t}^{camera} \in \mathbb{R}^{3\times1}$, the fused, ego-centric point cloud $\mathbf{P}^{\text{ego-centric}}$ is computed as $\mathbf{P}^{\text{ego-centric}} = \bigcup_{camera}^{\text{all cameras}} \mathbf{P}^{camera} (\mathbf{R}^{camera})^\intercal + (\mathbf{t}^{camera})^\intercal$. An example of the fused ego-centric colored point cloud is shown in Fig. A.2. The point cloud is then spatially cropped and downsampled using farthest point sampling (FPS) [165–167].

The RealSense T265 tracking camera provides 6D velocity and acceleration feedback at 200 Hz. It is mounted on the back of the mobile base using a custom-designed camera mount.

The R1 robot is equipped with an NVIDIA Jetson Orin, dedicated to running cameras and processing observations at a high rate.

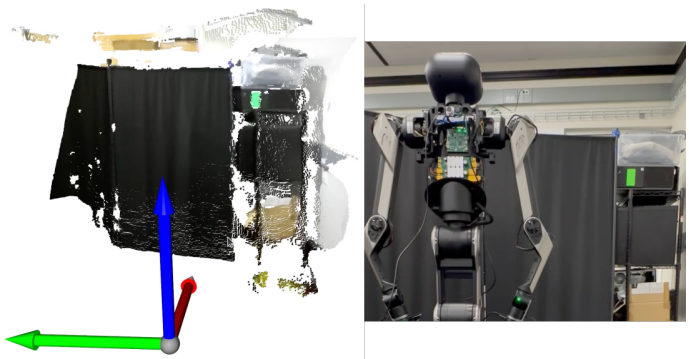


Fig. A.2: **Visualization of the fused, ego-centric colored point clouds. Left:** The colored point cloud observation, aligned with the robot's coordinate frame. **Right:** The robot's orientation and its surrounding environment.

D. Communication Scheme

The robot communicates with a workstation via the Robot Operating System (ROS). Each camera operates as an individual ROS node. The workstation runs the master ROS node, which subscribes to robot state nodes and camera nodes, and issues control commands via ROS topics. To reduce latency, a local area network (LAN) is established between the workstation and the robot.

Appendix B
JoyLo Details

This section provides details on JoyLo, including its hardware components, controller implementation, and data collection process.

A. Hardware Components

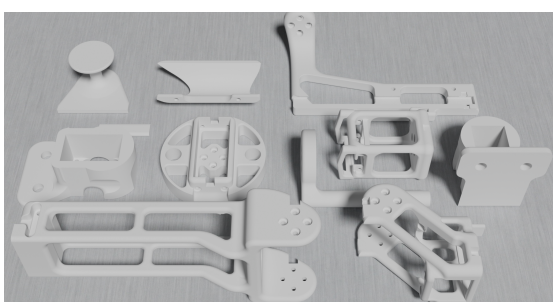


Fig. A.3: **Individual JoyLo arm links.**

The JoyLo system consists of 3D-printable arm links, low-cost Dynamixel motors, and off-the-shelf Joy-Con controllers. The individual arm links are shown in Fig. A.3. Using a Bambu Lab P1S 3D printer, we printed two arms in 13 h, consuming 317 g of PLA filament. The bill of materials is listed in Table A.IV. Once assembled, we use the official Dynamixel SDK to read motor states at 400 Hz - 500 Hz. The Joy-Cons connect to the workstation via Bluetooth, communicating at 66 Hz.

B. Controller Implementation

We provide an intuitive, real-time Python-based controller to operate JoyLo with the R1 robot. As illustrated in Pseudocode 1, the controller includes a joint impedance controller for the torso and arms with target joint positions as inputs, and a velocity controller for the mobile base with target base velocities as inputs. Control commands are converted into waypoints and sent to the robot via ROS topics at 100 Hz, which we find to be sufficient in practice.

To enable bilateral teleoperation of JoyLo arms as discussed in Sec. II, we implement a joint impedance controller using current-based control, where force is proportional to motor current. We set proportional gains $\mathbf{K_p} = [0.5, 0.5, 0.5, 0.5, 0.5, 0.5]$ and derivative gains $\mathbf{K_d} = [0.01, 0.01, 0.01, 0.01, 0.01, 0.01]$. To ensure sufficient stall torque for load-bearing joints in the JoyLo arms, such as the shoulder joints, the two low-cost Dynamixel motors are coupled together, as illustrated in Fig. 3.

C. Data Collection

During data collection, the robot operates at 100 Hz, while samples are recorded at 10 Hz. Functional buttons on the right Joy-Con (Fig. 3) control start, pause, save, and discard actions. Recorded data includes RGB images, depth images, point clouds, joint states, odometry, and action commands.

```
from brs_ctrl.robot_interface import R1Interface

# instantiate the controller
robot = R1Interface(...)
# send a control command
robot.control(
    # the torso and arms commands are target joint
    positions
    arm_cmd={
        "left": left_arm_target_q,
        "right": right_arm_target_q,
    },
    gripper_cmd={
        "left": left_gripper_target_width,
        "right": left_gripper_target_width,
    },
    torso_cmd=torso_target_q,
    # the mobile base commands are target velocities
    base_cmd=mobile_base_target_velocity,
)
```

Pseudocode 1: **Python interface for the R1 robot controller.**

Appendix C
Model Architectures, Policy Training, and Deployment Details

This section provides details on WB-VIMA and baseline model architectures, policy training, and real-robot deployment.

A. Preliminaries

a) Problem Formulation: We formulate robot manipulation as a Markov Decision Process (MDP) $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathcal{T}, \rho_0, R)$, where $s \in \mathcal{S}$ represents states, $a \in \mathcal{A}$ represents actions, $\mathcal{T}$ is the transition function, ρ_0 is

TABLE A.IV: **JoyLo bill of materials.**

Item No.	Part Name	Description	Quantity	Unit Price ($)	Total Price ($)	Supplier
1	Dynamixel XL330-M288-T	JoyLo arm joint motors	16	23.90	382.40	Dynamixel
2	Nintendo Joy-Con	JoyLo hand-held controllers	1	70	70	Nintendo
3	Dynamixel U2D2	USB communication converter for controlling Dynamixel motors	1	32.10	32.10	Dynamixel
4	5V DC Power Supply	Power supply for Dynamixel motors	1	<10	<10	Various
5	3D Printer PLA Filament	PLA filament for 3D printing JoyLo arm links	1	∼5	∼5	Various

Total Cost: ∼$499.5

the initial state distribution, and R is the reward function [168]. A policy π_θ, parameterized by θ, learns the mapping $\mathcal{S} \to \mathcal{A}$.

b) Denoising Diffusion for Policy Learning: A denoising diffusion probabilistic model (DDPM) [69, 169, 170] represents the data distribution $p(x^0)$ as the reverse denoising process of a forward noising process $q(x^k|x^{k-1})$, where Gaussian noise is iteratively applied. Given a noisy sample x^k and timestep k in the forward process, a neural network $\epsilon_\theta(x^k, k)$, parameterized by θ, learns to predict the applied noise ϵ. Starting with a random sample $x^K \sim \mathcal{N}(0, I)$, the reverse denoising process is described as

$$x^{k-1} \sim \mathcal{N}\left(\mu_k\left(x^k, \epsilon_\theta\left(x^k, k\right)\right), \sigma_k^2 I\right), \tag{A.1}$$

where $\mu_k(\cdot)$ maps the noisy sample x^k and the predicted noise ϵ_θ to the mean of the next distribution, and σ_k^2 is the variance obtained from a predefined schedule for $k = 1, \ldots, K$. Recently, DDPMs have been utilized to model policies π_θ, where the denoising network $\epsilon_\theta(a^k|s, k)$ is trained through behavior cloning [64–66].

B. WB-VIMA Architecture

1) Observation Encoder: As introduced in Sec. III-0a, there are two types of observation tokens: the point-cloud token $\mathbf{E}^{\text{pcd}}$ and the proprioceptive token $\mathbf{E}^{\text{prop}}$. A colored point-cloud observation is denoted as $\mathbf{P}^{\text{colored pcd}} \in \mathbb{R}^{N_{\text{pcd}} \times 6}$, where N_{pcd} is the number of points in the point cloud. Each point contains six channels: three for RGB values and three for spatial coordinates. To encode point-cloud tokens, RGB values are normalized to $[0, 1]$ by dividing by 255; spatial coordinates are normalized to $[-1, 1]$ by dividing by task-specific spatial limits; finally, a PointNet encoder [68] processes the point cloud. Proprioceptive observations include the mobile base velocity $v_{\text{mobile base}} \in \mathbb{R}^3$, torso joint positions $q_{\text{torso}} \in \mathbb{R}^4$, arms joint positions $q_{\text{arms}} \in \mathbb{R}^{12}$, and gripper widths $q_{\text{grippers}} \in \mathbb{R}^2$. These values are concatenated and processed through an MLP. Model hyperparameters for the PointNet and proprioception MLP are listed in Table A.V.

TABLE A.V: **Hyperparameters for PointNet and the proprioception MLP.**

Hyperparameter	Value	Hyperparameter	Value
PointNet		Prop. MLP	
N_{pcd}	4096	Input Dim	21
Hidden Dim	256	Hidden Dim	256
Hidden Depth	2	Hidden Depth	3
Output Dim	256	Output Dim	256
Activation	GELU	Activation	ReLU

2) Multi-Modal Observation Attention: To effectively fuse multi-modal observations, WB-VIMA employs a multi-modal observation attention network—a transformer decoder that applies causal self-attention over the input sequence: $\mathbf{S} = [\mathbf{E}^{\text{pcd}}_{t-T_o+1}, \mathbf{E}^{\text{prop}}_{t-T_o+1}, \mathbf{E}^{\text{a}}_{t-T_o+1}, \ldots, \mathbf{E}^{\text{pcd}}_t, \mathbf{E}^{\text{prop}}_t, \mathbf{E}^{\text{a}}_t] \in \mathbb{R}^{3T_o \times E}$, where T_o is the observation window size, E is the token dimension, and $\mathbf{E}^{\text{a}}$ represents the action readout token. The transformer decoder's hyperparameters are listed in Table A.VI. Action readout tokens are passive and do not influence the transformer output; they only attend to previous observation tokens to maintain causality. The final action readout token at time step t, $\mathbf{E}^a_t$, is used for autoregressive whole-body action decoding. We use an observation window size of $T_o = 2$ for all methods.

TABLE A.VI: **Hyperparameters for the transformer decoder used in multi-modal observation attention.**

Hyperparameter	Value
Embed Size	256
Num Layers	2
Num Heads	8
Dropout Rate	0.1
Activation	GEGLU [171]

3) Autoregressive Whole-Body Action Decoding: As discussed in Sec. III-0a, WB-VIMA jointly learns three independent denoising networks for the mobile base, torso, and arms, denoted as ϵ_{base}, ϵ_{torso}, and ϵ_{arms}, respectively. Each denoising network is implemented using a UNet [67], with hyperparameters listed in Table A.VII. The denoising process follows three sequential steps. First, the mobile base denoising network ϵ_{base} takes the action readout token $\mathbf{E}^a$ as input and predicts future mobile base actions $\mathbf{a}_{\text{base}} \in \mathbb{R}^{T_a \times 3}$. Subsequently, the torso denoising network ϵ_{torso} takes $\mathbf{E}^a$ and $\mathbf{a}_{\text{base}}$ as input and predicts future torso actions $\mathbf{a}_{\text{torso}} \in \mathbb{R}^{T_a \times 4}$. Finally, the arms denoising network ϵ_{arms} takes $\mathbf{E}^a$, $\mathbf{a}_{\text{base}}$, and $\mathbf{a}_{\text{torso}}$ as input and predicts future arm and gripper actions $\mathbf{a}_{\text{arms}} \in \mathbb{R}^{T_a \times 14}$. Here T_a is the action prediction horizon, and we use $T_a = 8$ hereafter. To ensure low-latency inference, denoising starts from the encoded action readout tokens, meaning the observation encoders and transformer run only once per inference call.

C. Baselines Architectures

We provide details on baseline methods DP3 [70], RGB-DP [65], and ACT [39]. DP3 uses the same PointNet encoder as WB-VIMA (Table A.V), but ignores RGB channels. Proprioceptive features are processed through the same MLP en-

TABLE A.VII: **Hyperparameters for the UNet models used for denoising.**

Hyperparameter	Value
Hidden Dim	[64,128]
Kernel Size	2
GroupNorm Num Groups	5
Diffusion Step Embd Dim	8

coder. Encoded features are concatenated and passed through a fusion MLP with two hidden layers and 512 hidden units. A UNet denoising network (Table A.VII) predicts a flattened 21-DoF whole-body action trajectory. RGB-DP is similar to DP3 but uses a pre-trained ResNet-18 [172] as the vision encoder. The last classification layer is replaced with a 512-dimensional output layer for policy learning. We use the recommended hyperparameters provided in Zhao et al. [39] for ACT.

D. Policy Training Details

Policies are trained using the AdamW optimizer [173], with hyperparameters in Table A.VIII. 90% of collected data is used for training, and 10% is reserved for validation. Policies are trained for equal steps, using the last checkpoint for evaluation. During training, we use the DDPM noise scheduler [69, 169, 170] with 100 denoising steps. During evaluation and inference, we use the DDIM noise scheduler [174] with 16 denoising steps. Training is performed using Distributed Data Parallel (DDP) on NVIDIA GPUs, including RTX A5000, RTX 4090, and A40.

TABLE A.VIII: **Training hyperparameters.**

Hyperparameter	Value
Learning Rate	7×10^{-4}
Weight Decay	0.1
Learning Rate Warm Up Steps	1000
Learning Rate Cosine Decay Steps	300,000
Minimal Learning Rate	5×10^{-6}

E. Policies Deployment Details

During deployment, observations from the robot's onboard sensors are transmitted to a workstation, where policy inference is performed, and the resulting actions are sent back for execution. To minimize latency, we implement asynchronous policy inference. Concretely, policy inference runs continuously in the background. When switching to a new predicted trajectory, the initial few actions are discarded to compensate for inference latency. This ensures non-blocking execution, preventing delays caused by observation acquisition and controller execution.

APPENDIX D
TASK DEFINITION AND EVALUATION DETAILS

This section provides detailed task definitions, generalization conditions, and evaluation protocols.

A. Task Definition

Activity 1 ***Clean House After a Wild Party*** *(Fig. 1 First Row)*: Starting in the living room, the robot navigates to a dishwasher in the kitchen (**ST-1**) and opens it (**ST-2**). It then moves to a gaming table (**ST-3**) to collect bowls (**ST-4**). Finally, the robot returns to the dishwasher (**ST-5**), places the bowls inside, and closes it (**ST-6**). Stable and accurate navigation is the most critical capability for this task. We collect 138 demonstrations, with an average human completion time of 210 s. We randomize the starting position of the robot, bowl instances and their placements, and distractors on the table.

Activity 2 ***Clean the Toilet*** *(Fig. 1 Second Row)*: In a restroom, the robot picks up a sponge placed on a closed toilet (**ST-1**), opens the toilet cover (**ST-2**), cleans the seat (**ST-3**), closes the cover (**ST-4**), and wipes it (**ST-5**). The robot then moves to press the flush button (**ST-6**). Extensive end-effector reachability is the most critical capability for this task. We collect 103 demonstrations, with an average human completion time of 120 s. We randomize the robot starting position, sponge instances, and placements.

Activity 3 ***Take Trash Outside*** *(Fig. 1 Third Row)*: The robot navigates to a trash bag in the living room, picks it up (**ST-1**), carries it to a closed door (**ST-2**), opens the door (**ST-3**), moves outside, and deposits the trash bag into a trash bin (**ST-4**). Stable and accurate navigation is the most critical capability for this task. We collect 122 demonstrations, with an average human completion time of 130 s. We randomize the robot starting position and the placement of the trash bag.

Activity 4 ***Put Items onto Shelves*** *(Fig. 1 Fourth Row)*: In a storage room, the robot lifts a box from the ground (**ST-1**), moves to a four-level shelf, and places the box on the appropriate level based on available space (**ST-2**). Extensive end-effector reachability is the most critical capability for this task. We collect 100 demonstrations, with an average human completion time of 60 s. We randomize the robot starting position, box placement, objects inside the box, shelf empty spaces, and distractors.

Activity 5 ***Lay Clothes Out*** *(Fig. 1 Fifth Row)*: In a bedroom, the robot moves to a wardrobe, opens it (**ST-1**), picks up a jacket on a hanger (**ST-2**), lays the jacket on a sofa bed (**ST-3**), and then returns to close the wardrobe (**ST-4**). Bimanual coordination is the most critical capability for this task. We collect 98 demonstrations, with an average human completion time of 120 s. We randomize the robot starting position, clothing placements, and clothing instances.

B. Policy Evaluation Results

Numerical results from policy evaluation are presented in Tables A.IX, A.X, A.XI, A.XII, and A.XIII.

C. Simulation Ablation Details

We design a simulated table-wiping task in OmniGibson [8] to perform ablation studies. The robot must use whole-body motions to wipe to a target hand position (marked by the yellow hand in Fig. 7) while maintaining contact with the

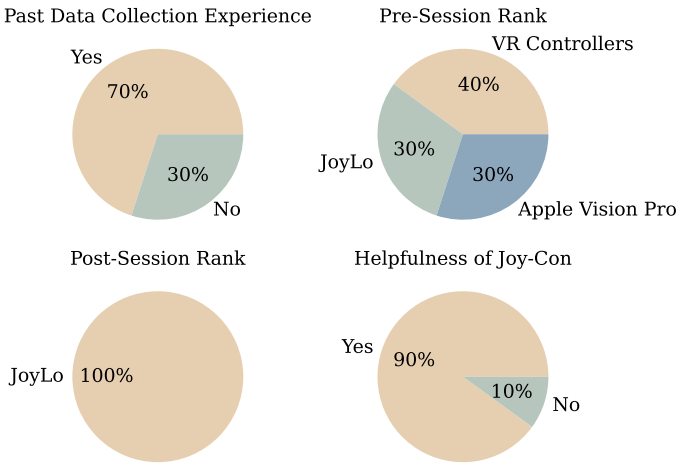


Fig. A.4: **Participant demographics and questionnaire results.**

table surface. To generate training data, we use cuRobo [175] to produce 100,000 whole-body trajectories, constraining the motion space by locking the mobile base and the first two torso joints. To isolate the effects of autoregressive whole-body action decoding and multi-modal observation attention, we replace camera input with a goal position, treated as a separate observation modality alongside robot proprioception.

D. User Study Details

As described in Sec. IV-0d, we conducted a user study with 10 participants to compare JoyLo against two alternative interfaces: VR controllers [25] and Apple Vision Pro [36, 72]. The study was conducted in the OmniGibson simulator [8] on the task "clean house after a wild party". To provide equal depth perception, participants wore a Meta Quest 3 headset while using both JoyLo and VR controllers. To eliminate bias, participants were exposed to the three interfaces in a randomized order. Each participant had a 10-minute practice session for each interface before beginning the formal evaluation. A successful task rollout is shown in Fig. A.6.

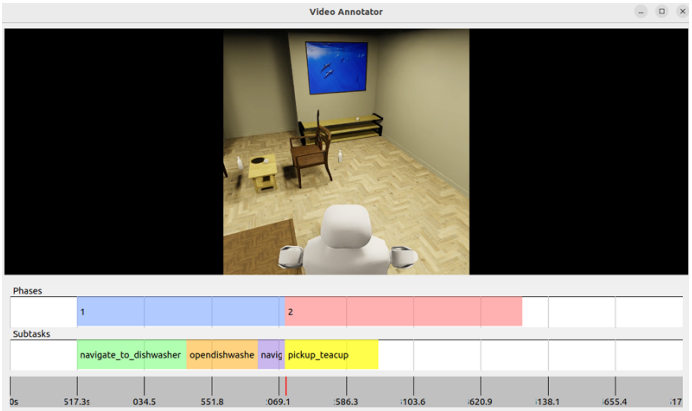


Fig. A.5: **GUI for annotating user study rollouts.**

After the sessions, rollouts were manually segmented, and task and sub-task completions were annotated using a GUI (Fig. A.5). For VR controllers and Apple Vision Pro, which use inverse kinematics (IK) based on end-effector poses, singular configurations were identified when the Jacobian matrix's condition number exceeded a set threshold. For JoyLo, which directly controls joints, excessive joint velocities were used as an indicator of singular or near-singular configurations. The post-session survey questions sent to participants are listed below:

- $\mathcal{Q}1$: Do you have prior data collection experience in robot learning? [Yes/No]
- $\mathcal{Q}2$: Before the session, which device did you expect to be the most user-friendly? [VR/Apple Vision Pro/JoyLo]
- $\mathcal{Q}3$: After the session, which device did you find to be the most user-friendly? [VR/Apple Vision Pro/JoyLo]
- $\mathcal{Q}4$: Did physically holding JoyLo arms help with data collection? [Yes/No]
- $\mathcal{Q}5$: Did using thumbsticks for torso and mobile base movement improve control? [Yes/No]

Fig. A.6: **Successful task completion by a participant.** The robot navigates to a dishwasher and opens it, moves to a table to collect teacups, returns to the dishwasher, places the teacups inside, and closes it.

TABLE A.IX: **Numerical evaluation results for the task "clean house after a wild party".** Success rates are shown as percentages. Values in parentheses indicate the number of successful trials out of the total trials.

	ET	ST-1	ST-2	ST-3	ST-4	ST-5	ST-6	Safety Violations
Human Teleop.	68% (50/73)	100% (73/73)	93% (69/74)	100% (69/69)	89% (64/72)	94% (60/64)	88% (53/60)	N/A
Ours	**40% (6/15)**	**100% (15/15)**	**80% (12/15)**	**80% (12/15)**	**73% (11/15)**	**93% (14/15)**	**93% (14/15)**	**0**
DP3 [70]	0% (0/15)	80% (12/15)	7% (1/15)	27% (4 / 15)	7% (1/15)	33% (5/15)	40% (6/15)	13
RGB-DP [65]	0% (0/15)	93% (14/15)	0% (0/15)	0% (0/15)	7% (1/15)	7% (1/15)	20% (3/15)	2
ACT [39]	0% (0/15)	80% (12/15)	0% (0/15)	0% (0/15)	0% (0/15)	0% (0/15)	0% (0/15)	2

TABLE A.X: **Numerical evaluation results for the task "clean the toilet".** Success rates are shown as percentages. Values in parentheses indicate the number of successful trials out of the total trials.

	ET	ST-1	ST-2	ST-3	ST-4	ST-5	ST-6	Safety Violations
Human Teleop.	61% (100/164)	91% (150/164)	72% (106/148)	99% (104/105)	100% (103/103)	98% (102/104)	98% (100/102)	N/A
Ours	**53% (8/15)**	**100% (15/15)**	**80% (12/15)**	**100% (15/15)**	**100% (15/15)**	**100% (15/15)**	**73% (11/15)**	**0**
DP3 [70]	0% (0/15)	**100% (15/15)**	47% (7/15)	93% (14/15)	0% (0/15)	13% (2/15)	0% (0/15)	**0**
RGB-DP [65]	0% (0/15)	93% (14/15)	13% (2/15)	7% (1/15)	7% (1/15)	0% (0/15)	20% (3/15)	2
ACT [39]	0% (0/15)	20% (3/15)	0% (0/15)	0% (0/15)	0% (0/15)	0% (0/15)	0% (0/15)	**0**

TABLE A.XI: **Numerical evaluation results for the task "take trash outside".** Success rates are shown as percentages. Values in parentheses indicate the number of successful trials out of the total trials.

	ET	ST-1	ST-2	ST-3	ST-4	Safety Violations
Human Teleop.	76% (96/127)	91% (116/128)	100% (124/124)	85% (106/125)	100% (115/115)	N/A
Ours	**53% (8/15)**	**80% (12/15)**	**100% (15/15)**	**87% (13/15)**	**87% (13/15)**	**1**
DP3 [70]	0% (0/15)	60% (9/15)	53% (8/15)	20% (3/15)	7% (1/15)	9
RGB-DP [65]	0% (0/15)	20% (3/15)	7% (1/15)	7% (1/15)	7% (1/15)	3
ACT [39]	0% (0/15)	0% (0/15)	0% (0/15)	0% (0/15)	0% (0/15)	5

TABLE A.XII: **Numerical evaluation results for the task "put items onto shelf".** Success rates are shown as percentages. Values in parentheses indicate the number of successful trials out of the total trials.

	ET	ST-1	ST-2	Safety Violations
Human Teleop.	89% (93/104)	90% (94/104)	100% (93/93)	N/A
Ours	**93% (14/15)**	**93% (14/15)**	**100% (15/15)**	**0**
DP3 [70]	20% (3/15)	27% (4/15)	47% (7/15)	**0**
RGB-DP [65]	13% (2/15)	20% (3/15)	40% (6/15)	**0**
ACT [39]	0% (0/15)	0% (0/15)	33% (5/15)	1
Ours w/o W.B. Action Denoising	40% (6/15)	40% (6/15)	60% (9/15)	**0**
Ours w/o Multi-Modal Obs. Attn.	13% (2/15)	33% (5/15)	40% (6/15)	**0**

TABLE A.XIII: **Numerical evaluation results for the task "lay clothes out".** Success rates are shown as percentages. Values in parentheses indicate the number of successful trials out of the total trials.

	ET	ST-1	ST-2	ST-3	ST-4	Safety Violations
Human Teleop.	50% (54/108)	56% (60/108)	93% (56/60)	96% (54/56)	100% (54/54)	N/A
Ours	**53% (8/15)**	**87% (13/15)**	**93% (14/15)**	**80% (12/15)**	60% (9/15)	**0**
DP3 [70]	0% (0/15)	13% (2/15)	13% (2/15)	27% (4/15)	27% (4/15)	7
RGB-DP [65]	0% (0/8)	13% (1/8)	25% (2/8)	13% (1/8)	13% (1/8)	3
ACT [39]	0% (0/15)	0% (0/15)	0% (0/15)	0% (0/15)	0% (0/15)	1
Ours w/o W.B. Action Denoising	13% (2/15)	33% (5/15)	73% (11/15)	73% (11/15)	**67% (10/15)**	**0**
Ours w/o Multi-Modal Obs. Attn.	0% (0/15)	33% (5/15)	40% (6/15)	47% (7/15)	13% (2/15)	4