2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

HOMER: Learning In-the-Wild Mobile Manipulation via Hybrid Imitation and Whole-Body Control

Priya Sundaresan[1], Rhea Malhotra[1], Phillip Miao[1], Jingyun Yang[1], Jimmy Wu[1], Hengyuan Hu[1], Rika Antonova[1,2], Francis Engelmann[1], Dorsa Sadigh[1], Jeannette Bohg[1]

Stanford University[1], University of Cambridge[2]

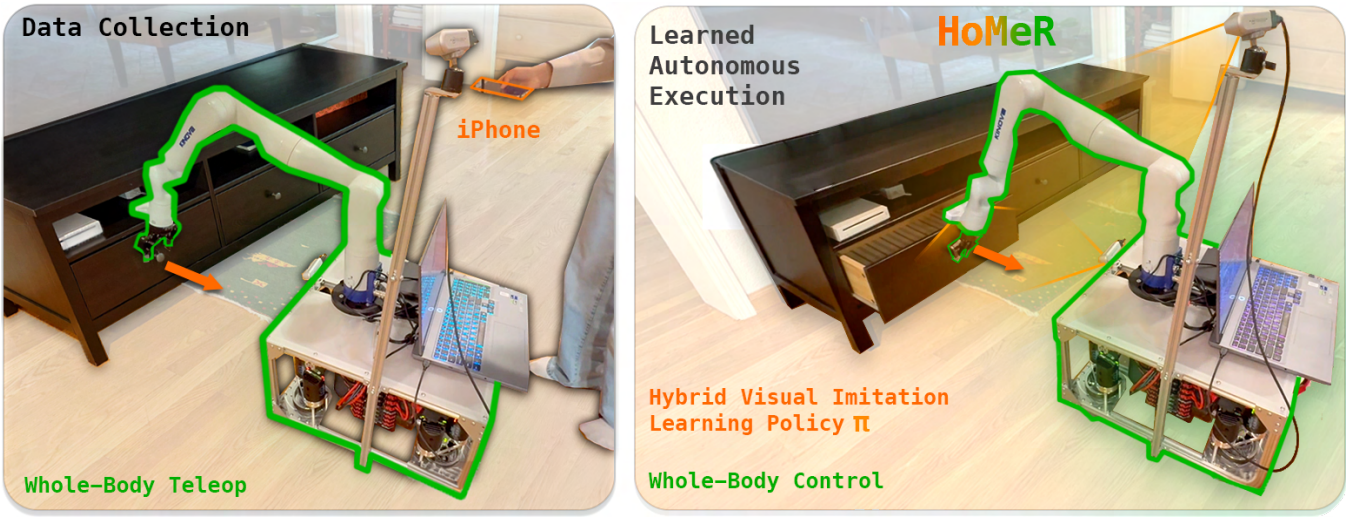

Fig. 1. **HOMER.** Left: A demonstrator uses whole-body iPhone teleoperation to collect data with a mobile manipulator in a real home. Right: From these collected demonstrations, HOMER learns a hybrid imitation learning policy that switches between absolute actions for reaching, and relative actions for fine manipulation. A whole-body controller maps these end-effector commands to arm and base joint commands for execution.

***Abstract*— We introduce HOMER, an imitation learning framework for mobile manipulation that combines whole-body control with hybrid action modes that handle both long-range and fine-grained motion, enabling effective performance on realistic in-the-wild tasks. At its core is a fast, kinematics-based whole-body controller that maps desired end-effector poses to coordinated motion across the mobile base and arm. Within this reduced end-effector action space, HOMER learns to switch between absolute pose predictions for long-range movement and relative pose predictions for fine-grained manipulation, offloading low-level coordination to the controller and focusing learning on task-level decisions. We deploy HOMER on a holonomic mobile manipulator with a 7-DoF arm in a real home. We compare HOMER to baselines without hybrid actions or whole-body control across 3 simulated and 3 real household tasks such as opening cabinets, sweeping trash, and rearranging pillows. Across tasks, HOMER achieves an overall success rate of 79.17% using just 20 demonstrations per task, outperforming the next best baseline by 29.17% on average. HOMER is also compatible with vision-language models and can leverage their internet-scale priors to better generalize to novel object appearances, layouts, and cluttered scenes. In summary, HOMER moves beyond tabletop settings and demonstrates a scalable path toward sample-efficient, generalizable mobile manipulation in everyday indoor spaces. Code, videos, and supplementary material are available at: https://homer-manip.github.io/.**

I. INTRODUCTION

To unlock the full potential of robots, we must move beyond controlled lab spaces and into the diverse, unstructured environments of everyday life. Unlike stationary tabletop robots, which lack the mobility to perform the wide range of tasks found in homes, offices, and warehouses, mobile manipulators are capable of navigating and interacting in these human-centric spaces. Tasks like watering a row of plants, wiping a spill on a long table, or moving to open a cabinet (Fig. 1) require not only precision, but also manipulation and mobility working hand-in-hand.

Modern mobile embodiments such as wheeled base-arm platforms [1–7], humanoids [8–16], and quadrupeds [17–21] enable robots to operate beyond fixed workspaces. However, these embodiments introduce significant control complexity. They require careful coordination between the base and arm in wheeled embodiments, or the limbs and torso in legged systems. One common approach to managing this complexity is to use whole-body controllers (WBCs) [22–25], which map high-level end-effector commands to coordinated whole-body motion using analytical or learning-based approaches. Yet even with WBCs addressing embodiment-level challenges, mobile manipulation remains difficult at the task level. Many mobile manipulation tasks are inherently multi-phase and demand seamless switching between long-range movements (e.g., reaching) and fine-grained local object manipulation. Several recent imitation learning (IL) methods have shown the benefits of hybrid action spaces to handle switching between long-range pose predictions or fine-grained local actions [26–28]. While these approaches achieve strong performance and generalization from limited demonstrations, their scope has so far been restricted to tabletop settings, leaving open the challenge of extending them

to the far more complex domain of mobile manipulation.

Our key insight is that scalable mobile manipulation requires not only an effective strategy for managing control complexity, but also a means of generalizing to novel scenarios. As mobile robots move through diverse human environments, they are exposed to far greater variability in objects, spatial configurations, and task conditions than static arms. To address both challenges, we propose **HOMER**: Hybrid whole-body policies for Mobile Robots. HOMER (Fig. 1) combines a fast, kinematics-based whole-body controller (which maps end-effector actions to coordinated base-arm motion) with a hybrid IL policy that switches between keypose predictions for long-range movement and dense delta actions for fine-grained manipulation. This structured approach addresses high-dimensional control and multi-phase execution. Additionally, we show that HOMER is modular enough to incorporate task-relevant keypoints derived from vision-language models (VLMs), providing a path towards improved generalization in unfamiliar environments.

We deploy **HOMER** in a real home and evaluate it on a suite of challenging mobile manipulation tasks reflecting everyday household demands. Overall, our contributions are:

1) A **sample-efficient IL framework for mobile manipulation** that leverages WBC and hybrid action representations to outperform strong non-hybrid and non-WBC baselines from only 20 demos per task.
2) A **modular policy architecture that can be conditioned on VLM keypoints**, enabling generalization to novel object geometries, appearances, and clutter.
3) A **practical whole-body controller that supports intuitive teleoperation**, facilitating efficient demonstration collection in real household settings.

II. Related Work

Mobile Manipulation and Control.
Legged platforms typically focus on navigation across diverse terrains (*e.g.*, with Boston Dynamics Spot [20] and ANYmal [29] quadrupeds). A few recent works equip quadrupeds with lightweight arms and develop whole-body controllers, utilizing either model-based motion planning (*e.g.*, RoLoMa [30]) or learning-based manipulation policies (*e.g.*, DeepWBC [23], Visual WBC [24], UMI-on-Legs [22]). Our work draws inspiration from these works, but our framework is agnostic to the exact implementation of the WBC (*e.g.*, inverse kinematics (IK) or learning-based). In our work, we use a task-agnostic IK-based WBC that is reusable across many scenarios. Furthermore, in contrast to these prior works, we learn policies with hybrid action modes to encourage both spatial generalization and precision. More recently, humanoid research has advanced rapidly, with most efforts centered on whole-body control for expressive behaviors such as dancing, walking, or jumping [9, 10]. Although some humanoid systems perform manipulation, they often rely on using high-dimensional joint-space actions (*e.g.*, 50+ DoF for HumanPlus [11]) for imitation learning. In contrast, our work adopts a whole-body control strategy with hybrid action modes to enable more tractable learning.

Wheeled platforms consist of a wheeled mobile base with onboard arms, and include examples like TidyBot [1], TidyBot++ [2], mobile Franka Pandas [4, 5], the Fetch robot [6], the HSR [7], Mobile Aloha [31], and the Everyday Robot [32]. Although these embodiments are physically capable of performing many mobile manipulation tasks, they typically assume decoupled control of the base and arm. Having to switch between different movement modes adds complexity to teleoperation, and often requires the use of ad hoc and task-specific strategies.

A variety of other works study *navigation* with legged or wheeled embodiments [33–35]. Our work is different and complementary in that we focus on the "last mile" of manipulation, where mobility is necessary for task completion, but not at the scale or complexity of full-scene navigation.

Visual Imitation Learning. Visual imitation learning (IL) refers to learning from demonstrations using visual observations [36–38]. Recent methods explore IL agents with various action granularities and input/output spaces.

Dense policies, such as Diffusion Policy [39], ACT [40], or Visual-Language-Action (VLA) models (Gemini [41], π_0 [42], RT-X [43], OpenVLA [44], etc.) predict low-level actions (*e.g.*, 6-DoF deltas or joint velocities) at every timestep. While effective for reactive manipulation, dense policies struggle with long-horizon tasks and spatial generalization, since even simple movements like reaching can involve hundreds of consecutive actions.

Keypose-based policies, such as PerAct [45], RVT [46], and KITE [47], predict 6-DoF end-effector poses that are executed via low-level controllers or motion planners. While sample-efficient, these keypose actions can be too sparse to handle precise or reactive control.

Recently proposed *hybrid policies* combine keypose and dense actions for both long-range and precise local manipulation. Hydra [28] and AWE [27] use keypose and/or dense actions but rely solely on images, limiting spatial generalization. SPHINX [26], most similar to our approach, uses images and point clouds with learned attention to task-relevant keypoints to switch between modes. Critically, all these methods are limited to static, tabletop-mounted manipulators. We extend SPHINX to the mobile manipulation setting by incorporating whole-body control, enabling mobility while retaining an end-effector–centric action space. We further support conditioning on object keypoints from VLMs, allowing generalization to unseen objects in clutter.

III. HOMER: Hybrid whOle-body policies for Mobile Robots

A. Problem Formulation

We consider a mobile manipulator composed of a holonomic mobile base and an N-DoF robotic arm, with joint configuration $\mathbf{q}_t = (\mathbf{q}_t^{\text{base}}, \mathbf{q}_t^{\text{arm}}) \in \mathbb{R}^{3+N}$, where $\mathbf{q}_t^{\text{base}} = (x, y, \theta) \in SE(2)$ represents the base pose, and $\mathbf{q}_t^{\text{arm}} \in \mathbb{R}^N$ represents the arm joints.

At each timestep t, an observation $o_t = \left(\mathbf{q}_t, g_t, \{\mathbf{I}_t^k, \mathbf{D}_t^k\}_{k=1}^K\right)$ includes the joint configuration $\mathbf{q}_t$, gripper state $g_t \in \mathbb{R}$, and

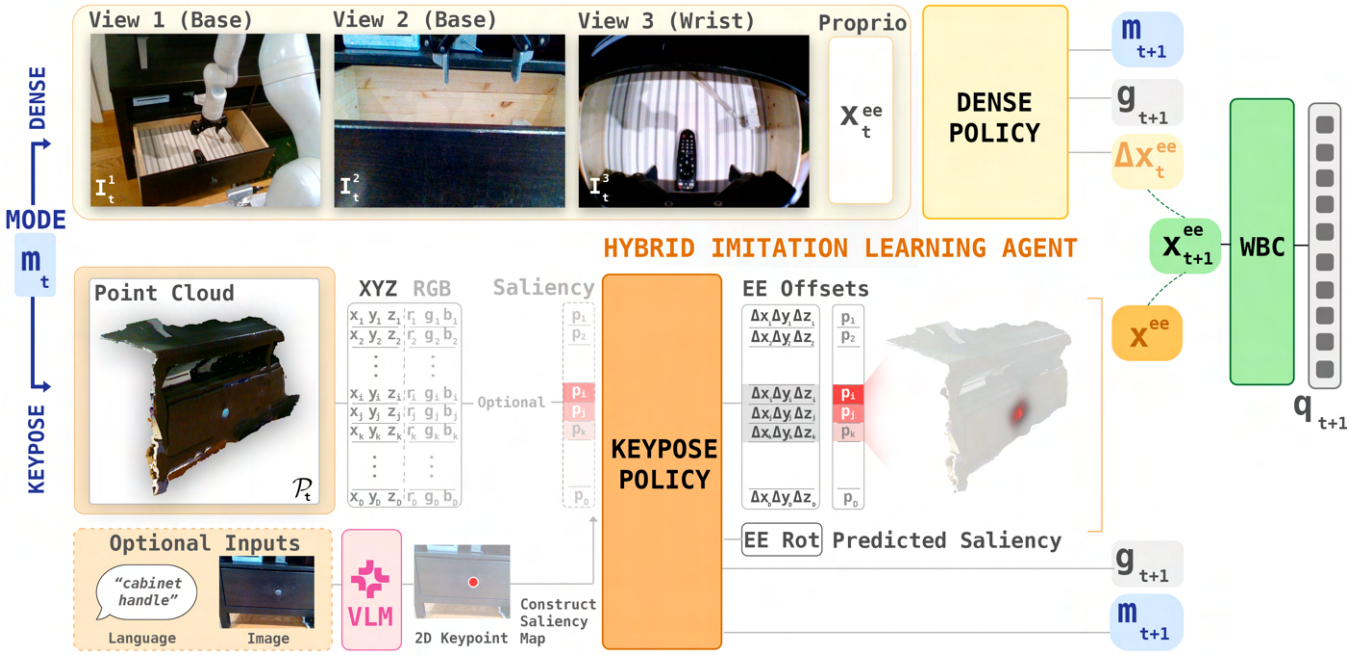

Fig. 2: **HOMER Policy Architecture:** HOMER consists of a *dense policy* that uses RGB images to predict relative actions for fine-grained manipulation, and a *keypose policy* that uses point clouds to predict absolute end-effector poses for long-range motion. Each policy also predicts the next control mode, enabling learned transitions. Optionally, the keypose policy can be conditioned on externally provided salient points derived from a VLM to support dynamic goal specification (HOMER-COND). Finally, a whole-body controller (WBC) converts predicted end-effector actions into joint commands for the mobile base and arm.

RGB-D images from K cameras (with at least one wrist-mounted and one third-person view). To get 3D point clouds, we assume known camera intrinsics and extrinsics.

Rather than learning actions directly in joint space, our goal is to train an imitation learning (IL) policy $\pi(o_t) = (\mathbf{x}^{\text{ee}}_{t+1}, g_{t+1})$ that predicts a 6-DoF end-effector target pose $\mathbf{x}^{\text{ee}}_{t+1} \in SE(3)$, which can then be executed by a whole-body controller (WBC). With this formulation, the IL policy reasons in task space while delegating low-level control and embodiment-specific constraints to the WBC.

B. Whole-Body Controller

We implement a kinematics-based WBC that maps high-level end-effector poses into joint position commands for the full embodiment, delegating joint-space coordination, constraints, and redundancy resolution to the controller rather than the IL agent. Though HOMER is agnostic to the exact WBC implementation, ours is based on MuJoCo [48] and the mink IK library [49].

Formally, the WBC is a mapping $\mathscr{W}: SE(3) \to \mathbb{R}^{3+N}$ from a desired end-effector pose $\mathbf{x}^{\text{ee}} \in SE(3)$ to a joint position command $\mathbf{q} \in \mathbb{R}^{3+N}$ for the base and arm. We implement an iterative IK solver that finds $\mathbf{q}$ minimizing the pose error. To compute a velocity that moves the end-effector toward $\mathbf{x}^{\text{ee}}$, we define pose error as a body-frame twist [50]:

$$\mathbf{e}^{\text{ee}} = \log\left((\mathbf{x}^{\text{ee}}_t)^{-1}\mathbf{x}^{\text{ee}}\right)$$

where $\mathbf{x}^{\text{ee}}_t \in SE(3)$ is the current end-effector pose from forward kinematics. The geometric Jacobian $\mathbf{J}_{\text{ee}}(\mathbf{q}_t) \in \mathbb{R}^{6\times(3+N)}$ maps joint velocities to the induced end-effector twist. At each iteration, the IK solver finds $\dot{\mathbf{q}}$ that minimizes the discrepancy between the Jacobian-induced twist and $\mathbf{e}^{\text{ee}}$, moving the end-effector toward the desired pose. Specifically, the IK solver optimizes the following:

$$\min_{\dot{\mathbf{q}}} \quad \underbrace{\|\mathbf{J}_{\text{ee}}(\mathbf{q}_t)\dot{\mathbf{q}} - \mathbf{e}^{\text{ee}}\|^2_{W_{\text{ee}}}}_{\text{End-effector term}} + \underbrace{\|\mathbf{q}_t + \dot{\mathbf{q}}\cdot\Delta t - \mathbf{q}_{\text{retract}}\|^2_{W_{\text{posture}}}}_{\text{Posture term}} + \underbrace{\|\dot{\mathbf{q}}^{\text{base}}\|^2_{W_{\text{damping}}}}_{\text{Damping term}}$$

$$\text{s.t.} \quad \dot{\mathbf{q}}_{\min} \le \dot{\mathbf{q}} \le \dot{\mathbf{q}}_{\max} \tag{1a}$$

$$\mathbf{q}_{\min} \le \mathbf{q}_t + \dot{\mathbf{q}}\cdot\Delta t \le \mathbf{q}_{\max} \tag{1b}$$

$$-\mathbf{n}_i^\top \mathbf{J}_i(\mathbf{q}_t)\dot{\mathbf{q}} \le \underbrace{\frac{\gamma(d_i - d_{\min})}{\Delta t} + \varepsilon}_{\text{collision margin}}, \quad \forall i \in \mathscr{C} \tag{1c}$$

Objective: The first term encourages the solved joint motion to move towards the target pose $\mathbf{x}^{\text{ee}}$. The second term encourages the joint configuration $\mathbf{q}_t + \dot{\mathbf{q}}\cdot\Delta t$ to stay close to a neutral resting posture $\mathbf{q}_{\text{retract}} \in \mathbb{R}^{3+N}$, shown in Fig. 3. The third term damps the motion of the base. Weights W_{ee}, W_{posture}, and W_{damping} specify the influence of each term.

Constraints: The optimization is subject to constraints that ensure safe and feasible execution. We impose joint velocity Eq. (1a) and position limits Eq. (1b), $\dot{\mathbf{q}}_{\min} \le \dot{\mathbf{q}} \le \dot{\mathbf{q}}_{\max}$ and $\mathbf{q}_{\min} \le \mathbf{q}_t + \dot{\mathbf{q}}\cdot\Delta t \le \mathbf{q}_{\max}$, to satisfy hardware bounds. In constraint Eq. (1c), we enforce velocity-based collision avoidance between selected pairs of robot components, each modeled as geometric primitives (*geoms*) in the MuJoCo simulator [48]. For each pair $i \in \mathscr{C}$, we identify the closest points between geoms and compute the signed distance d_i,

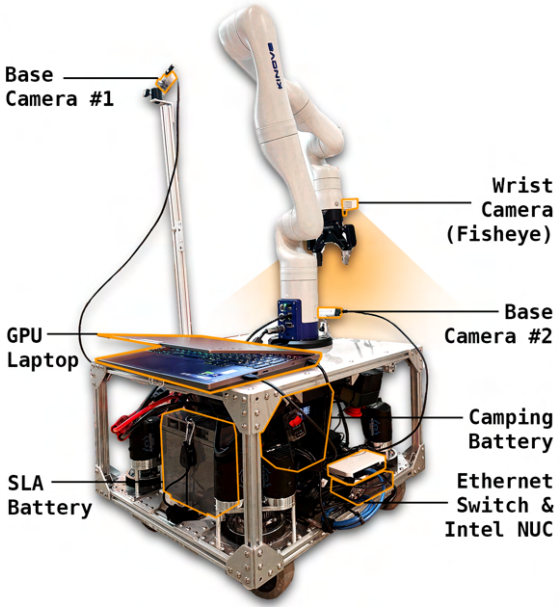

Fig. 3: **Hardware:** We use the TidyBot++ holonomic mobile manipulator [2] with two base cameras and a wrist-mounted fisheye camera. An onboard NUC handles real-time control, and an onboard GPU laptop runs policy inference.

the contact normal $\mathbf{n}_i$, and the Jacobian $\mathbf{J}_i$ of the contact point with respect to joint motion. The constraint aims to slow the robot's motion as the clearance d_i between geoms approaches a minimum threshold, effectively acting as a velocity damper. In our setup, $\mathscr{C}$ includes the (arm, base) and (arm, camera mount) pairs, where camera mounts are represented as cylinders.

The IK solver iteratively integrates the optimized joint velocities using a fixed timestep Δt to obtain the final joint position command: $\mathbf{q}_{t+1} = \mathbf{q}_t + \dot{\mathbf{q}} \cdot \Delta t$. The joint position commands are subsequently executed on hardware using low-level controllers. All WBC hyperparameters are detailed at: homer-manip.github.io/supp.html#sec-wbc and are held constant across all tasks without any per-task retuning.

C. Hybrid Imitation Learning Agent

Built on the WBC's end-effector control space, our hybrid imitation learning (IL) agent consists of two sub-policies: a **keypose sub-policy** for long-range motion and a **dense sub-policy** for fine-grained manipulation. Each sub-policy predicts both the next end-effector action and next control mode $m_{t+1} \in \{\texttt{keypose}, \texttt{dense}, \texttt{terminate}\}$, indicating the sub-policy choice for the next action.

Teleoperation We collect data for HOMER using the iPhone-based interface from [2], which streams the phone's real-time 6-DoF pose via the WebXR API and maps it to the robot's end-effector. Gripper commands are issued through swipe gestures. The WBC from Section III-B solves for joint-space actions at each timestep. We record observations and actions at 10 Hz during teleoperation.

Keypose Sub-policy The keypose sub-policy π^{keypose} handles long-range movements such as reaching, where predicting an absolute end-effector pose provides greater stability than step-wise deltas. It takes as input a third-person point cloud $\mathscr{P}_t \subset \mathbb{R}^3$, constructed by deprojecting RGB-D images using known intrinsics and extrinsics, and outputs a 6-DoF end-effector pose $\mathbf{x}^{\text{ee}}$, gripper state g_{t+1}, and next control mode m_{t+1}. Following SPHINX [26], we avoid directly regressing the target end-effector pose. Instead, the policy predicts per-point saliency probabilities over the input cloud and per-point 3D offsets to the ground-truth end-effector position. The point with the highest saliency defines the *salient point*—a task-relevant 3D location such as a keypoint on a cabinet handle (Fig. 2). During training, we supervise offset predictions only at points with high predicted or ground-truth saliency, encouraging the model to focus on task-relevant regions (Fig. 2, shaded offsets). End-effector orientation, gripper state, and control mode are predicted using additional learnable tokens. At test time, we apply the predicted offset at the highest-saliency point to obtain the positional component, and combine it with the predicted orientation and gripper state to form the full end-effector action $\mathbf{x}^{\text{ee}}$. We then execute interpolated poses from the current pose $\mathbf{x}_t^{\text{ee}}$ to reach the keypose $\mathbf{x}^{\text{ee}}$. Training the keypose policy requires action labels, mode labels, and salient point annotations. For a given dataset of 20 demos, we post-hoc annotate salient points and modes using a lightweight GUI interface, which takes $\sim$15 minutes (see homer-manip.github.io/supp.html#sec-annotation). We train the policy using the Transformer-based architecture from SPHINX [26].

Conditioned Keypose Sub-policy We additionally extend the keypose sub-policy to a salient point-conditioned variant, HOMER-COND, which optionally accepts an externally provided 3D keypoint. This enables us to tap into the internet-scale visual and semantic knowledge encoded in vision-language models (VLMs) by prompting them to localize unseen objects in cluttered scenes, and conditioning HOMER-COND on the resulting deprojected 3D keypoints (Fig. 2). Taking the original point cloud, we first construct a distance-weighted saliency map, where each point's value is inversely proportional to its distance from the provided keypoint, and the map is normalized to represent probabilities of saliency. We concatenate this saliency map as an additional channel at the input. During training, we apply a masked supervision strategy: in 50% of samples, the conditioned saliency map is masked out, and the model learns to predict both saliency and actions as in the unconditioned setting; in the remaining 50%, we pass the unmodified conditioned saliency map and supervise only the action predictions, with offsets penalized only for points with high ground-truth saliency in the conditioned map. This formulation allows the policy to leverage external guidance when available. We further apply train-time visual augmentations to promote generalization: adding randomly generated clusters of points to the input point cloud to mimic distractors, and omitting the RGB channel to reduce overfitting to object appearance.

Dense Sub-policy The dense sub-policy π^{dense} is intended for fine-grained manipulation near salient points, such as inserting, aligning, or grasping objects. The input consists of both third-person and wrist-mounted RGB images $\{\mathbf{I}_t^k \in \mathbb{R}^{H\times W\times 3}\}_{k=1}^K$, along with the current end-effector state $\mathbf{x}_t^{\text{ee}} \in$

$SE(3)$ computed via forward kinematics from $\mathbf{q}_t$. The dense policy predicts a 6D delta action $\Delta\mathbf{x}_t^{\text{ee}} \in \mathbb{R}^6$ relative to the current end-effector pose (Fig. 2), as well as the next control mode m_{t+1}. We obtain the target pose as $\mathbf{x}_{t+1}^{\text{ee}} = \mathbf{x}_t^{\text{ee}} + \Delta\mathbf{x}_t^{\text{ee}}$. We instantiate π^{dense} using Diffusion Policy [39], which in practice predicts a horizon of 16 future actions and executes 8 before re-planning, rather than predicting a single delta action at each timestep.

a) Execution.: The agent automatically switches between sub-policies based on the current mode m_t:

$$(\mathbf{x}_{t+1}^{\text{ee}}, m_{t+1}) = \begin{cases} \pi^{\text{keypose}}(\mathscr{P}_t) & \text{if } m_t = \texttt{keypose} \\ \pi^{\text{dense}}(\{\mathbf{I}_t^k\}) & \text{if } m_t = \texttt{dense} \end{cases}$$

We assume that m_1 corresponds to keypose mode, as nearly all manipulation tasks involve first reaching before performing fine-grained manipulation. For each timestep thereafter, the predicted action $\mathbf{x}_{t+1}^{\text{ee}}$ is passed to the WBC to solve for and execute $\mathbf{q}_{t+1} = \mathscr{W}(\mathbf{x}_{t+1}^{\text{ee}})$ (Section III-B). HOMER uses the predicted mode m_{t+1} to select the next sub-policy, enabling dynamic alternation between reaching and manipulation based on learned transitions.

IV. EXPERIMENTS

We deploy HOMER on the TidyBot++ robot [2], consisting of a 7-DoF Kinova arm and holonomic base (Fig. 3). With this platform, we evaluate a diverse set of challenging manipulation tasks in both simulation and real-world to investigate three core questions, focusing on the benefits of HOMER's imitation learning (IL) agent, whole-body controller (WBC), and generalization capabilities:

(Q1) *Do hybrid actions help with multi-step tasks combining reaching and fine manipulation?*
(Q2) *Does the WBC action space improve performance compared to decoupled base-arm actions?*
(Q3) *Can* HOMER *generalize to novel object instances and spatial configurations?*

	Task	Description	R-R	R-O	P	LH
Sim	*Cube*	Pick up cube placed randomly across large workspace		✔	✔	
	Dishwasher	Open randomly placed dishwasher door		✔	✔	
	Cabinet Opening	Open randomly placed side-hinged cabinet door		✔	✔	
Real	*Pillow*	Move pillow placed randomly on carpet to target couch position	✔	✔	✔	✔
	TV Remote	Open cabinet, retrieve remote, place on stand	✔	✔	✔	✔
	Sweep Trash	Grasp brush, sweep at least 3/4 trash clumps into bin.	✔	✔	✔	✔

TABLE I: **Mobile Manipulation Tasks.** We evaluate our approach on 3 simulated and 3 real-world tasks, covering randomized robot poses (R-R), randomized object poses (R-O), need for precision (P), and long-horizon reasoning (LH). Darker checkmarks indicate greater emphasis on the corresponding aspect.

A. Q1-2: Are hybrid actions & WBC beneficial?

Baselines: We compare HOMER to baselines varying along two axes: hybrid vs. dense-only action spaces, and whole-body vs. decoupled base-arm control. Hybrid variants are trained on data annotated post-hoc with control modes and salient points. WBC baselines are trained from whole-body teleoperation demonstrations (Fig. 1), while base+arm (B+A) baselines use decoupled teleoperation, in which the base has to be directly teleoperated separately from the arm [2].

Diffusion Policy (B+A): A dense policy that predicts 10-DoF relative poses: 3-DoF base pose, 6-DoF end-effector pose, and 1-DoF gripper command. This is comparable to the dense, base-arm policies used in [2, 31, 51] which notably were trained with 50-200 demos.

HOMER (B+A): A hybrid policy identical to HOMER, but predicting either a 3-DoF base keypose, a 6-DoF arm keypose, or a 10-DoF relative pose (as above).

Diffusion Policy (WBC): A dense policy that predicts 7-DoF relative poses: 6-DoF end-effector pose and 1-DoF gripper command, executed through the WBC.

HOMER: Our hybrid policy that predicts either a 6-DoF end-effector keypose or a 6-DoF relative end-effector pose, plus a 1-DoF gripper command, executed through the WBC.

We expect hybrid action modes and WBC to each provide advantages in tasks involving wide workspaces, precise phases, and long horizons. HOMER (B+A) and DP (WBC) each capture one of these components, and may perform competitively by partially addressing these challenges. In contrast, DP (B+A), which lacks both hybrid and whole-body action abstractions, must learn base-arm coordination and long-horizon planning from scratch without structural priors, making the learning problem significantly harder.

Our benchmark tasks are described in Table I. We train and evaluate all methods using 20 demonstrations on 3 simulated and 3 real-world tasks. In Fig. 4, we show illustrations of each task *(top)* and benchmark results *(bottom)*. Across tasks, DP (B+A) struggles the most with reaching and aligning to targets, particularly in tasks like *Cube*, *Dishwasher*, *Cabinet Opening*, and *Pillow*, with significant randomization in either initial robot poses (R-R) or object placements (R-O). The dense end-effector deltas output by the policy often veer off course, leading to failures in reaching pre-manipulation configurations. DP (WBC) exhibits similar limitations without exploiting keyposes, but performs slightly better. We posit that the simplified end-effector action space enabled by the WBC can be beneficial in the low-data regime. HOMER (B+A) is the strongest baseline, using base and arm keyposes to move to favorable poses before manipulation. This highlights the value of our hybrid IL architecture. However, it struggles when smooth base-arm coordination is required (*Cabinet*, *Dishwasher*, *Sweep Trash*), or when base misalignment affects arm reachability.

HOMER achieves the highest success rates across tasks (Fig. 4). Specifically, HOMER performs challenging maneuvers like manipulating appliances larger than the robot itself

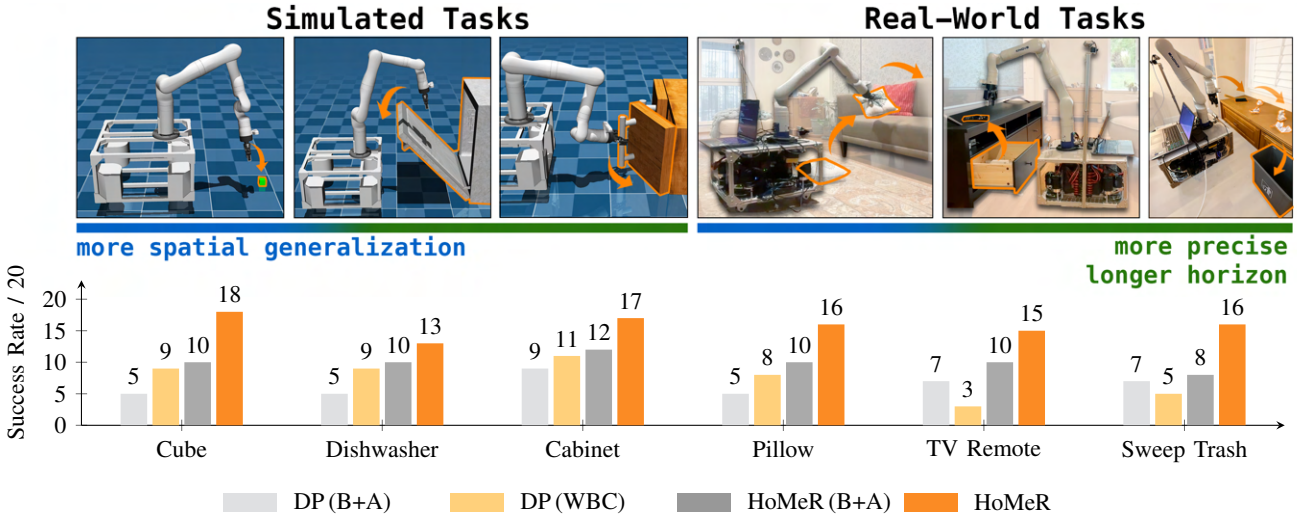

Fig. 4: **Benchmarking Results.** We evaluate HoMeR on six simulated and real-world tasks (top) that require spatial generalization, precision, and long-horizon reasoning. *TV Remote* and *Sweep Trash* are particularly challenging due to their multi-step nature. HoMeR consistently outperforms baselines that use only dense actions or decoupled base-arm control, highlighting the benefits of hybrid action modes and whole-body coordination. The performance of all methods is best understood through videos: homer-manip.github.io/#benchmark.

(*Cabinet*, *Dishwasher*), and executing smooth long-horizon motions (*Sweep Trash*) with precision (*TV Remote*) while generalizing to randomized initial object and robot poses.

B. Q3: Generalization to Novel Scenarios

To assess generalization, we evaluate HoMeR-Cond (Section III-C), a variant of HoMeR that (1) conditions the keypose policy on external salient points from a VLM, and (2) trains on point clouds without color and with randomly generated distractor points for visual robustness.

Simulated Resuts: We first evaluate HoMeR-Cond in simulation on three challenging variants of the *Cube* environment: (1) randomizing cube sizes, (2) adding distractors, and (3) retrieving different-colored cubes. We use MolMo 7B-D [52], a VLM capable of detecting pixel-level keypoints from language prompts to localize the target cube from images. We see that in Fig. 5, both HoMeR and HoMeR-Cond-NoAugs perform well in simple settings, but struggle with distractors and novel object appearances. In contrast, HoMeR-Cond maintains high performance, highlighting the combined importance of salient point conditioning and augmentations for handling clutter and unseen objects.

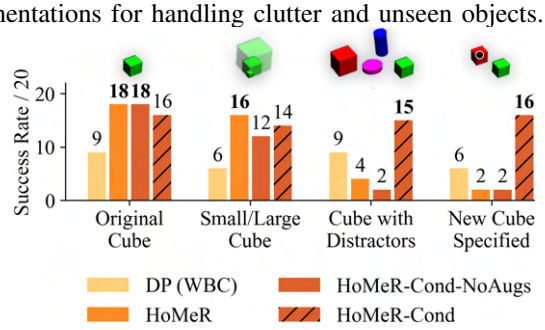

Fig. 5: **Simulated Generalization:** HoMeR-Cond achieves strong generalization to unseen scenarios by combining salient point conditioning with point cloud augmentations (videos at homer-manip.github.io/#generalization). Without augmentations (HoMeR-Cond-NoAugs) or conditioning (HoMeR), performance drops on distractors & novel appearances.

Fig. 6: **Real-World Generalization: Left:** Given a specified target stuffed animal ("brown bear"), and an image from the base-mounted camera, Gemini identifies an appropriate 2d keypoint (●). **Right:** Conditioned on this deprojected salient point, HoMeR-Cond successfully picks up the selected animal and places it in the basket. This illustrates HoMeR's ability to leverage off-the-shelf VLMs for open-set, language-conditioned manipulation.

Real-World Results: We further evaluate HoMeR-Cond on a real-world, language-conditioned decluttering task. We collect 20 demonstrations of picking up various stuffed animals and placing them into a basket, with salient points labeled on the picked items. We then train a hybrid HoMeR-Cond policy that uses waypoint mode to approach a target item and dense mode to grasp and place it. At test time, we use Gemini-2.5-Pro [53], an off-the-shelf VLM, to identify a user-specified animal from a language query. Like MolMo, Gemini provides a 2D click prediction, which we deproject into a 3D salient point that conditions HoMeR-Cond for execution (Fig. 6). In Table II, we report HoMeR-Cond's strong performance on both seen and unseen stuffed animals, subject to clutter and visual distractors.

	Seen	Unseen	Overall
Success Rate	9/10	7/10	16/20

TABLE II: Real-world stuffed animal task. HoMeR-Cond succeeds on both seen and unseen objects, achieving 16/20 overall.

Together, these results underscore HoMeR's seamless integration with SoTA foundation models to perform open-set manipulation in unstructured real-world environments.

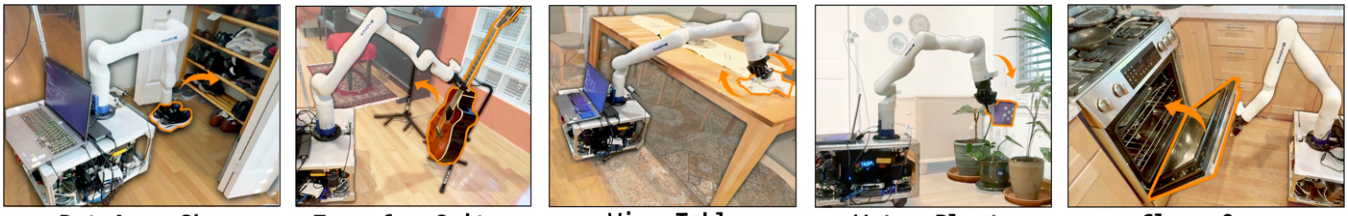

Fig. 7: **Teleoperated Tasks:** We demonstrate a range of teleoperated tasks enabled by our WBC interface in a real household environment.

Our policy's generalization capabilities are best viewed at homer-manip.github.io#homer-cond-real, with VLM prompts at homer-manip.github.io/supp.html#sec-salient.

C. Qualitative Results: Whole-Body Teleoperation

We also qualitatively assess our WBC through teleoperation in a real home. The WBC enables smooth, reliable teleoperation of diverse tasks, including opening/closing cabinets, doors, blinds, and ovens; coordinated motions such as wiping tables and watering plants; and precise tasks like stowing shoes or moving a guitar between stands (Fig. 7). The WBC optionally avoids collisions between the arm, base, and camera mounts, viewable at homer-manip.github.io/#data-collection-video, highlighting our WBC's potential for practical teleop in-the-wild.

V. Discussion

We present HoMeR, a hybrid imitation learning framework for mobile manipulators that combines spatially grounded policy learning with a whole-body controller for executing end-effector actions. By switching between keypose and dense control modes, and operating within a lower-dimensional action space, HoMeR enables generalizable and precise manipulation. Through rigorous evaluations in a *real home*, HoMeR demonstrates strong performance on diverse, everyday household tasks. HoMeR's compatiblity with VLMs further enables language prompting and generalization to unseen object instances in clutter, even having been trained on just *20 demonstrations*. Our results highlight the promise of leveraging hybrid actions, whole-body control, and internet-scale priors from VLMs towards in-the-wild deployment of mobile manipulators. Future work will extend our approach in several ways, namely integrating collision handling with the environment to augment the already implemented self-collision avoidance, leveraging actuated cameras to handle broader workspaces with active perception, chaining navigation with HoMeR policies for extended-horizon tasks, co-training on mobile and non-mobile datasets as enabled by our end-effector centric action space, and extending HoMeR to the multi-task setting.

VI. Acknowledgements

Toyota Research Institute, Intrinsic, the Stanford Institute for Human-Centered Artificial Intelligence (HAI) and the Sloan Foundation provided funds to support this work. This work is also in part supported by funds from NSF Awards 2132847, 2327974, and 2006388. Priya Sundaresan is supported by an NSF GRFP. Francis Engelmann is supported by an SNSF PostDoc Mobility Fellowship. Use and Disclosure of AI: We used GPT-4o for minor code completion (e.g., syntax and boilerplate).

References

[1] 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*, 2023.

[2] 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.

[3] H. P. Brøndmo, "Introducing the everyday robot project," https://blog.x.company/introducing-the-everyday-robot-project-27860f3461a4, 2019, accessed: 2025-04-06.

[4] 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.

[5] K. Rana, J. Haviland, S. Garg, J. Abou-Chakra, I. Reid, and N. Suenderhauf, "Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning," *arXiv preprint arXiv:2307.06135*, 2023.

[6] M. Wise, M. Ferguson, D. King, E. Diehr, and D. Dymesich, "Fetch and freight: Standard platforms for service robot applications," in *Workshop on autonomous mobile service robots*, 2016.

[7] T. Yamamoto, T. Nishino, H. Kajima, M. Ohta, and K. Ikeda, "Human Support Robot (HSR)," in *ACM SIGGRAPH 2018 emerging technologies*, 2018.

[8] 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.

[9] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, "Expressive Whole-body Control for Humanoid Robots," *arXiv preprint arXiv:2402.16796*, 2024.

[10] R.-Z. Qiu, S. Yang, X. Cheng, C. Chawla, J. Li, T. He, G. Yan, L. Paulsen, G. Yang, S. Yi, *et al.*, "Humanoid policy˜ human policy," *arXiv preprint arXiv:2503.13441*, 2025.

[11] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "Humanplus: Humanoid shadowing and imitation from humans," *arXiv preprint arXiv:2406.10454*, 2024.

[12] 1X Technologies, "1x world model challenge for humanoid robots," 2025. [Online]. Available: https://github.com/1x-technologies/1xgpt

[13] C. Lu, X. Cheng, J. Li, S. Yang, M. Ji, C. Yuan, G. Yang, S. Yi, and X. Wang, "Mobile-TeleVision: Predictive Motion Priors for Humanoid Whole-Body Control," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2025.

[14] M. Ji, X. Peng, F. Liu, X. Cheng, R. Yang, G. Yang, and X. Wang, "Exbody2: Advanced expressive humanoid whole-body control," *arXiv preprint arXiv:2412.13196*, 2024.

[15] Unitree Robotics, "Unitree h1: Full-size universal humanoid robot," 2025. [Online]. Available: https://www.unitree.com/h1

[16] C. Sferrazza, D.-M. Huang, X. Lin, Y. Lee, and P. Abbeel, "HumanoidNench: Simulated Humanoid Benchmark for Whole-body Locomotion and Manipulation," *arXiv preprint arXiv:2403.10506*, 2024.

[17] D. Shah, B. Osinski, B. Ichter, and S. Levine, "LM-Nav: Robotic Navigation with Large Pre-Trained Models of Language, Vision, and Action," *arXiv preprint arXiv:2207.04429*, 2022.

[18] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning Quadrupedal Locomotion over Challenging Terrain," *Science Robotics*, 2020.

[19] A. Kumar, Z. Xu, D. Pathak, and J. Malik, "RMA: Rapid Motor Adaptation for Legged Robots," *arXiv preprint arXiv:2107.04034*, 2021.

[20] A. Agarwal, A. Kumar, J. Malik, and D. Pathak, "Legged locomotion in challenging terrains using egocentric vision," in *Conference on robot learning*, 2023.

[21] G. B. Margolis, T. Chen, K. Paigwar, X. Fu, D. Kim, S. Kim, and P. Agrawal, "Learning to Jump from Pixels," *Proceedings of the 5th Conference on Robot Learning (CoRL)*, 2022.

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

[23] Z. Fu, X. Cheng, and D. Pathak, "Deep whole-body control: Learning a unified policy for manipulation and locomotion," in *Conference on Robot Learning (CoRL)*, 2022.

[24] M. Liu, Z. Chen, X. Cheng, Y. Ji, R. Qiu, R. Yang, and X. Wang, "Visual whole-body control for legged loco-manipulation," *The 8th Conference on Robot Learning*, 2024.

[25] J. Brüdigam, A. A. Abbas, M. Sorokin, K. Fang, B. Hung, M. Guru, S. Sosnowski, J. Wang, S. Hirche, and S. Le Cleac'h, "Jacta: A versatile planner for learning dexterous and whole-body manipulation," *arXiv preprint arXiv:2408.01258*, 2024.

[26] P. Sundaresan, H. Hu, Q. Vuong, J. Bohg, and D. Sadigh, "What's the Move? Hybrid Imitation Learning via Salient Points," *Proc. Int. Conf. on Learning Representations*, 2024.

[27] L. X. Shi, A. Sharma, T. Z. Zhao, and C. Finn, "Waypoint-Based Imitation Learning for Robotic Manipulation," *Conference on Robot Learning (CoRL)*, 2023.

[28] S. Belkhale, Y. Cui, and D. Sadigh, "Hydra: Hybrid Robot Actions for Imitation Learning," in *Conference on Robot Learning (CoRL)*, 2023.

[29] M. Hutter, C. Gehring, A. Lauber, F. Gunther, C. D. Bellicoso, V. Tsounis, P. Fankhauser, R. Diethelm, S. Bachmann, M. Blösch, *et al.*, "ANYMal - Toward Legged Robots for Harsh Environments," *Advanced Robotics*, 2017.

[30] H. Ferrolho, V. Ivan, W. Merkt, I. Havoutis, and S. Vijayakumar, "Roloma: Robust loco-manipulation for quadruped robots with arms," *Autonomous Robots*, vol. 47, no. 8, pp. 1463–1481, 2023.

[31] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation," in *Conference on Robot Learning (CoRL)*, 2024.

[32] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[33] N. Hirose, C. Glossop, A. Sridhar, D. Shah, O. Mees, and S. Levine, "LeLan: Learning a Language-conditioned Navigation Policy from In-the-wild Videos," *Conference on Robot Learning (CoRL)*, 2024.

[34] N. Yokoyama, S. Ha, D. Batra, J. Wang, and B. Bucher, "VLFM: Vision-language Frontier Maps for Zero-shot Semantic Navigation," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2024.

[35] S. Jauhri, S. Lueth, and G. Chalvatzaki, "Active-perceptive Motion Generation for Mobile Manipulation," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2024.

[36] S. Schaal, "Learning from demonstration," *Advances in neural information processing systems*, 1996.

[37] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[38] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," in *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[39] 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*, 2023.

[40] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning Fine-grained Bimanual Manipulation with Low-cost Hardware," in *Proc. Robotics: Science and Systems (RSS)*, 2023.

[41] G. R. Team, S. Abeyruwan, J. Ainslie, J.-B. Alayrac, M. G. Arenas, T. Armstrong, A. Balakrishna, R. Baruch, M. Bauza, M. Blokzijl, *et al.*, "Gemini Robotics: Bringing AI into the Physical World," *arXiv preprint arXiv:2503.20020*, 2025.

[42] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, *et al.*, "*pi*_0: A Vision-Language-Action Flow Model for General Robot Control," *arXiv preprint arXiv:2410.24164*, 2024.

[43] O. X.-E. Collaboration, A. O'Neill, A. Rehman, A. Gupta, A. Maddukuri, A. Gupta, *et al.*, "Open X-Embodiment: Robotic Learning Datasets and RT-X Models," https://arxiv.org/abs/2310.08864, 2023.

[44] M. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. 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," *arXiv preprint arXiv:2406.09246*, 2024.

[45] M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-Actor: A Multi-task Transformer for Robotic Manipulation," in *Conference on Robot Learning (CoRL)*, 2023.

[46] A. Goyal, J. Xu, Y. Guo, V. Blukis, Y.-W. Chao, and D. Fox, "RVT: Robotic View Transformer for 3D Object Manipulation," in *Conference on Robot Learning (CoRL)*, 2023.

[47] P. Sundaresan, S. Belkhale, D. Sadigh, and J. Bohg, "KITE: Keypoint-Conditioned Policies for Semantic Manipulation," in *Conference on Robot Learning*, 2023.

[48] E. Todorov, "MuJoCo: A Physics Engine for Model-Based Control," http://www.mujoco.org, 2012, accessed: 2025-04-15.

[49] K. Zakka, "Mink: Python inverse kinematics based on MuJoCo," July 2024. [Online]. Available: https://github.com/kevinzakka/mink

[50] R. M. Murray, Z. Li, and S. S. Sastry, *A Mathematical Introduction to Robotic Manipulation*. CRC Press, 1994.

[51] A. Prasad, K. Lin, J. Wu, L. Zhou, and J. Bohg, "Consistency policy: Accelerated visuomotor policies via consistency distillation," *arXiv preprint arXiv:2405.07503*, 2024.

[52] M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini, *et al.*, "Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models," *arXiv preprint arXiv:2409.17146*, 2024.

[53] G. Team, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," 2025. [Online]. Available: https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf