MobileManiBench: Simplifying Model Verification for Mobile Manipulation

Wenbo Wang[2] Fangyun Wei[1] Qixiu Li[3] Xi Chen[1] Yaobo Liang[1]
Chang Xu[2⋆] Jiaolong Yang[1] Baining Guo[1]

[1]Microsoft Research Asia [2]University of Sydney [3]Tsinghua University
{fawe,xichen6,yalia,jiaoyan,bainguo}@microsoft.com
wwan0412@uni.sydney.edu.au liqx23@mails.tsinghua.edu.cn
c.xu@sydney.edu.au

Abstract. Vision-language-action models have advanced robotic manipulation but remain constrained by reliance on the large, teleoperation-collected datasets dominated by the static, tabletop scenes. We propose a simulation-first framework to verify VLA architectures before real-world deployment and introduce MobileManiBench, a large-scale benchmark for mobile-based robotic manipulation. Built on NVIDIA Isaac Sim and powered by reinforcement learning, our pipeline autonomously generates diverse manipulation trajectories with rich annotations (language instructions, multi-view RGB–depth–segmentation images, synchronized object/robot states and actions). MobileManiBench features 2 mobile platforms (parallel-gripper and dexterous-hand robots), 2 synchronized cameras (head and right wrist), 630 objects in 20 categories, 5 skills (open, close, pull, push, pick) with over 100 tasks performed in 100 realistic scenes, yielding 300K trajectories. This design enables controlled, scalable studies of robot embodiments, sensing modalities, and policy architectures, accelerating research on data efficiency and generalization. We benchmark representative VLA models and report insights into perception, reasoning, and control in complex simulated environments, with all code, datasets, and models publicly released at our project website: https://dexhand.github.io/MobileManiBench/.

Keywords: Vision–language–action model · Mobile robot manipulation

1 Introduction

Recent advances in Vision-Language-Action models [4–6, 8, 17, 21, 24–26, 31, 35, 36, 40, 43, 46, 50, 51, 53, 56, 58, 60] have substantially advanced robotic manipulation, demonstrating strong generalization to novel objects and diverse visual domains. However, the success of current VLA models heavily dependents on large-scale datasets, among which, the *openX-Embodiment* dataset [35] has become the de facto training resource. Yet, most of its data are collected via teleoperation from static, third-person or head-mounted viewpoints, restricted to indoor tabletop environments populated with a limited set of household objects. These constraints introduce several challenges:

– Any modification to the hardware configuration—such as incorporating new sensors (e.g., depth or wrist-mounted cameras), extending to mobile-based

⋆ Corresponding author.

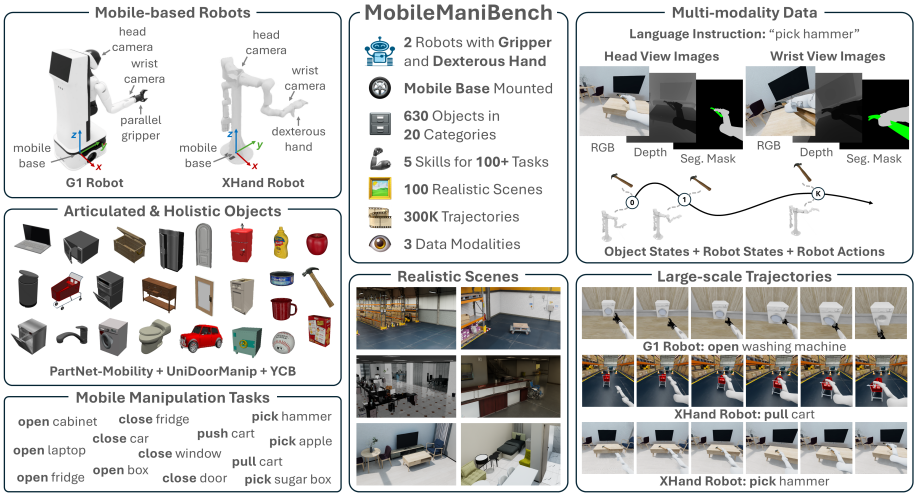

Fig. 1: Overview of **MobileManiBench**. It features 2 mobile-based robots: the G1 robot with a parallel gripper and the XHand robot with a dexterous hand. The benchmark includes 630 articulated and holistic objects across 20 categories and supports 5 mobile manipulation skills—open, close, pull, push, and pick—enabling over 100 tasks. To efficiently scale data generation while ensuring task success, we train a universal **MobileManiRL** policy for each robot–object–skill triplet and generate **MobileManiDataset** across 100 realistic scenes with 300K trajectories and 3 data modalities—language instructions, multi-view RGB–depth–segmentation images, synchronized object/robot states and actions. MobileManiBench offers a flexible testbed to accelerate model innovation and data-efficiency research for VLA models.

manipulation, or replacing grippers with dexterous hands—necessitates recollecting data from scratch. If these additions fail to provide substantial performance gains, the effort becomes wasteful due to the high cost and inefficiency of teleoperation-based data collection.

– Although teleoperation pipelines are effective for simple gripper-based robots, they become cumbersome when scaled to dexterous or mobile platforms, thereby impeding rapid model development and iteration.

To support scalable *verification* of VLA architectures, we propose conducting data generation and evaluation in simulation environments prior to collecting real-world robotic data. The key idea is to offer an efficient way to scale up objects, tasks, and scenes, while flexibly supporting diverse robot configurations—such as dexterous hands, mobile bases, arbitrary camera placements—and easily acquiring multi-modal sensory inputs (e.g., RGB, depth, robot proprioception). In this work, we leverage *NVIDIA Isaac Sim* [34] as the simulation platform and employ reinforcement learning [39] to train agents that can automatically generate manipulation trajectories along with corresponding annotations (e.g., text instructions, camera images, and robot states). Building upon this automated pipeline, we introduce **MobileManiBench**—a large-scale benchmark for mobile manipulation and embodied skill learning.

As shown in Figure 1, our MobileManiBench is designed with several distinguishing features:

- *Multi-Robot Embodiment*: Two robot platforms are provided—one equipped with a parallel gripper and another with a dexterous anthropomorphic hand.
- *Mobile-Base Platform*: Both robots are mounted with a mobile base, allowing for spatially extended tasks that require coordinated navigation and manipulation.
- *Rich Visual Sensing*: Two cameras are mounted on the head and right wrist, each capturing synchronized RGB, depth, and segmentation images to facilitate multi-view and multi-sensory perception.
- *Multi-Modality Trajectories*: Each recorded trajectory includes textual instructions, multi-view visual images, together with object states, robot states and actions, forming comprehensive supervision signals for VLA training.
- *Diverse and Scalable Content*: The dataset contains 630 unique objects spanning 20 categories, 100 realistic and semi-structured scenes, 5 mobile manipulation skills across 100 tasks, and generates over 300K trajectories.

The above design enables controlled experimentation with different robot configurations, sensory modalities, and model architectures, without requiring expensive real-world data collection. Researchers can efficiently evaluate whether new input modalities (e.g., wrist or depth images) contribute to better manipulation performance or whether specific architectural designs improve multi-sensory fusion and reasoning. MobileManiBench provides a flexible testbed for accelerating both model innovation and data-efficiency exploration in embodied AI.

Leveraging MobileManiBench, we systematically compare several existing VLA models [5, 17, 22, 25] with our proposed MobileManiVLA, analyze their strengths and weaknesses across various tasks, and provide a series of empirical takeaways that offer new insights into how large-scale multi-modal architectures perceive and act in complex robotic manipulation environments. We hope that this benchmark will foster reproducible research, enable rapid architectural iteration, and pave the way toward scalable, general-purpose VLA models.

2 Related Works

Robotic Manipulation Benchmarks. Robotic manipulation [1, 2, 9, 10, 13, 14, 18, 20, 28, 33, 48, 55, 57] remains a central challenge in embodied AI. Existing datasets and benchmarks can be broadly categorized into two groups, as summarized in Table 1. The first group comprises real-world datasets and evaluation protocols, such as DROID [20], AgiBotWorld-Beta [1] and Open X-Embodiment [35], which collect large-scale manipulation trajectories through extensive teleoperation or by aggregating data from multiple robot platforms. Despite their impressive scale, these works primarily focus on fixed-base or gripper-equipped robots performing mostly tabletop tasks. Recent efforts like DexGraspVLA [58], Humanoid Everyday [57], and Mobile ALOHA [13] extend to dexterous hands and mobile manipulation, yet still rely on real-world human teleoperation, which limits their diversity and scalability. Overall, real-world data

Benchmarks / Datasets	Object Number	Skill Number†	Scene Number	Trajectory Number	Mobile Based	Parallel Gripper	Dexterous Hand	Articulated Object	Universal Model††
DROID [20]	*	86	564	76K		✓		✓	
Open X-Embodiment [35]	5228	527	311	1400K		✓		✓	✓
AgiBotWorld-Beta [1]	3000	87	106	1000K		✓	✓	✓	✓
DexGraspVLA [58]	36	1	-	2K			✓		✓
Humanoid Everyday [57]	*	221	*	10.3K	✓		✓	✓	✓
Mobile-ALOHA [13]	6	6	6	300	✓	✓		✓	
DexArt [2]	82	2	1	/			✓	✓	
RLBench [18]	*	30	1	/		✓		✓	
Robosuite [59]	10	10	1	/		✓		✓	
SIMPLER [27]	17	4	-	/		✓		✓	✓
RoboTwin 2.0 [10]	731	11	-	100K		✓		✓	✓
VLABench [55]	2164	11	-	5K		✓		✓	✓
LIBERO [29]	75	*	20	6.5K		✓		✓	✓
CALVIN [32]	30	10	4	20K		✓		✓	✓
BEHAVIOR-1K [23]	9318	10	50	10K	✓	✓		✓	
Maniskill2 [14]	2144	12	*	30K	✓	✓		✓	
RoboCasa [33]	2509	8	120	100K	✓	✓		✓	
UniDoorManip [28]	328	1	1	/	✓	✓		✓	✓
OWMM-Agent [9]	157	2	143	21K	✓	✓			✓
GRUtopia [47]	2956	2	100	900	✓	✓	✓		
MobileManiBench (Ours)	**630**	**5**	**100**	**300K**	✓	✓	✓	✓	✓

Table 1: Comparison of existing robotic manipulation benchmarks and datasets. The top rows correspond to real-world datasets, while the bottom rows refer to simulation-based works. *Skill Number*† denotes the set of manipulation verbs included. *Universal Model*†† indicates whether a universal policy has been trained within the benchmark. A "–" in *Scene Number* indicates that the work uses a single tabletop layout with varying textures. A "/" in *Trajectory Number* indicates that users must generate the trajectories themselves. An asterisk "∗" denotes statistics not reported in the original paper. AgiBotWorld-Beta collects real-world G1-robot data through teleoperation, with only a small number of mobile-manipulation examples.

collection and model evaluation remain costly and risky, as they require complex scene setups and pose a high risk of hardware damage. Moreover, although existing real-world benchmarks cover a variety of manipulation skills, their evaluations are often limited to a narrow subset of tasks, leading to imbalanced training and testing coverage across object categories and skills.

The second group [2, 9, 10, 14, 15, 18, 27–29, 32, 33, 47, 49, 55, 59] alleviates the above limitation by leveraging high-fidelity simulation environments to simulate diverse manipulation tasks, collect large-scale trajectories, and perform extensive model evaluations. These methods typically collect data either through simulation-based teleoperation or by designing task-specific agents using rule-based methods or reinforcement learning (RL) policies. While simulation greatly reduces the cost and risk, there remains a lack of large-scale, multi-modality trajectory datasets that support training universal models capable of generalizing across diverse object categories, manipulation skills, and realistic scenes. MobileManiBench follows this simulation-based paradigm, leveraging the high-fidelity *NVIDIA Isaac Sim* [34] and a diverse set of digital assets to train universal RL policies across various robot–object–skill combinations. It supports large-scale trajectory generation in diverse realistic digital scenes, facilitating scalable model training and evaluation for mobile-based robotic manipulation with both parallel grippers and dexterous hands.

Vision-language-action Models. VLA models [4–6, 8, 17, 21, 24–26, 31, 35, 36, 40, 43, 46, 50, 51, 53, 56, 58, 60] have emerged as a promising framework for learning universal robotic policies that generalize across diverse objects, tasks, and scenes. Recent works [4,5,17,25,51] advance this direction by integrating vision–language (VL) backbones [3, 11, 12, 19, 30, 42] with diffusion-based or flow-matching-based action modules. By combining the pretrained VL priors, multi-modality representations, and continuous action modeling, these approaches substantially lead the robotic manipulation performance.

However, existing VLA models are mostly trained and evaluated on fixed-base, gripper-equipped robots performing tabletop manipulation tasks, with limited examples of dexterous hands or mobile-based manipulation [6, 8, 21, 25, 26, 31, 35, 36, 40, 43, 46, 50, 51, 53, 56, 58, 60]. MobileManiBench extends this research by providing large-scale synthetic trajectories for mobile-based robots equipped with either parallel grippers or dexterous hands, covering diverse objects, tasks, and scenes. It systematically studies the effect of different model structures and input modalities on mobile manipulation performance, filling a critical gap in current VLA research.

3 Problem Formulation

Mobile-based Robot. MobileManiBench employs two mobile-based robots: the G1 robot from AGIBOT [38] and the XHand robot from RobotEra [37], as shown in Figure 1. Although both robots are equipped with dual arms, we fix the left arm and activate only the right arm for manipulation. Each robot has a 7-DOF right arm and a 2-DOF mobile base (one rotational DOF around the z-axis and one translational DOF along the y-axis). The G1 robot uses a parallel gripper with 1 DOF as its end effector, whereas the XHand robot is equipped with a dexterous 12-DOF hand. Both robots are controlled using (6+D) dimensional actions, where the first 6 dimensions represent the wrist pose displacement relative to the previous frame, from which inverse kinematics computes the target joint angles of mobile base and right arm. The remaining D dimensions correspond to the target joint angles of end effector. This formulation yields 7-d actions for the G1 robot and 18-d actions for the XHand robot.

Mobile-based Manipulation Objects, Skills, and Tasks. MobileManiBench encompasses both the articulated and holistic objects sourced from the PartNet-Mobility [52], UniDoorManip [28], and YCB [7] datasets, totaling 630 objects across 20 categories (holistic objects such as apples, bottles, and toys are grouped into a single "holistic" category). We define five mobile manipulation skills: *open* and *close* for objects including boxes, toilets, laptops, trashcans, dishwashers, ovens, microwaves, refrigerators, washing machines, cars, safes, fridges, cabinets, windows, lever doors, round doors, faucets, and tables; *pull* and *push* for carts; and *pick* for holistic (YCB) objects. This formulation yields over 100 distinct tasks for large-scale data collection and model training, including tasks such as opening a laptop, closing a cabinet, pushing a cart, pulling a cart, picking an apple, and picking a mustard bottle.

4 Method

Overview. The construction of MobileManiBench consists of two stages. First, in the *MobileManiRL Training* stage (Section 4.1), we train a state-based reinforcement learning (RL) policy for each robot–object–skill combination, covering robots equipped with either a parallel gripper or a dexterous hand. Second, in the *MobileManiDataset Generation* stage (Section 4.2), each trained RL policy is deployed in diverse digital scenes to collect successful manipulation trajectories. Each trajectory includes a natural language instruction paired with synchronized data: multi-view camera images, object states, robot states, and actions.

Building on *MobileManiDataset*, we further introduce *MobileManiVLA* (Section 4.3), where all trajectories across robot–object–skill–scene combinations are aggregated to train a universal vision-language-action model for each robot, enabling strong generalization to unseen objects and unseen scenes.

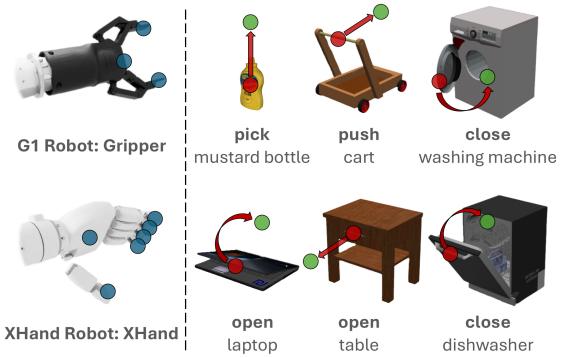

Fig. 2: Definitions of the robot gripper/hand points (blue), object grasp point (red), and goal point (green) across diverse tasks.

4.1 MobileManiRL Training

Given that our benchmark involves 2 mobile-based robots, 630 objects across 20 categories, 5 manipulation skills, and over 100 tasks, teleoperating or manually designing policies for each configuration would be prohibitively time-consuming. To address this, we propose a universal state-based reinforcement learning (RL) policy, termed *MobileManiRL*, which parameterizes each robot–object–skill combination using keypoint-based displacements of the robot gripper/hand points, the object grasp point, and the goal point [16, 54], as illustrated in Figure 2. A universal reward function R encourages the robot's gripper/hand points to reach the object grasp point and transport it to the goal point. This formulation enables a single RL policy to generalize across diverse manipulation scenarios while maintaining task-specific success.

Inputs. Table 2 summarizes the five input types used in MobileManiRL. Specifically, *Time* encodes the timestep embedding; *Object State* represents the states of the object grasp point and goal point; *Robot Proprioception* captures the robot's body and joint states; *Robot–Object Distance* measures the distances between the robot gripper/hand points and the object grasp point; and *Previous Action* records the previous (6+D)-dimensional action.

Input Type	G1 Robot	XHand Robot
Time	30	30
Object State	9	9
Robot Proprioception	78	135
Robot-Object Distance	22	31
Previous Action	7	18

Table 2: Dimensions of the input types used in MobileManiRL for the G1 and XHand robots. Detailed definitions of each input element are provided in the Appendix.

Network Architecture. Each MobileManiRL policy network is a 4-layer MLP with hidden dimensions of {1024, 1024, 512, 512}, followed by an action prediction head implemented as a single fully connected layer. This head outputs a (6+D) dimensional vector (7-d for the G1 robot and 18-d for the XHand robot) representing the action at the current time step. The value network adopts the same architecture as the policy network but outputs a single scalar.

Reward Function. The reward function R is designed as:

$$R = R_d + (1 - f_g)R_a + f_g(R_g + R_m + R_s), \tag{1}$$

where the global distance reward R_d penalizes the distances between the gripper/hand points and the object grasp point, as well as between the object grasp point and the goal point, encouraging the robot to grasp the object and move it toward the goal. The grasp flag f_g is set to 1 when the distance between the gripper/hand points and the object grasp point falls below a predefined threshold. The approach action reward R_a guides the gripper/hand toward the object grasp point via a reference approach action until a grasp is achieved. Once grasped (i.e., $f_g = 1$), the grasp reward R_g provides a bonus; the move action reward R_m encourages the gripper/hand to move to the goal point via a reference move action; finally the success reward R_s provides an extra bonus when the object reaches the goal. Formal definitions of all reward elements are provided in the Appendix. A manipulation is considered successful if the object grasp point reaches the goal point within $T = 300$ steps.

Simplified Scene Training. We place the 20 categories of objects into 2 simplified scene settings in Isaac Sim and train one MobileManiRL policy for each robot-object-skill combination, as shown in Figure 3. The *ground* scene includes toilet, trashcan, refrigerator, washing machine, car, fridge, cabinet, window, lever door, round door, table, and cart; while the *tabletop* scene includes box, laptop, dishwasher, oven, microwave, safe, faucet, and holistic (YCB) objects. Each object is initialized either on the ground or on the tabletop with random heights, facing the negative y axis, and with its grasp point's x-y coordinates set to zero. The robot mobile base is initialized facing the object grasp point with random positional offsets ($1.0m$ to $1.5m$ along the x and y axes) and rotational offsets (-15^o to 15^o around the z axis) to enhance policy robustness. Under this setup, we train MobileManiRL on each of the 1,182 robot-object–skill combinations for both G1 robot and XHand robot, achieving mean success rates of **89.6%** and **92.9%**, respectively.

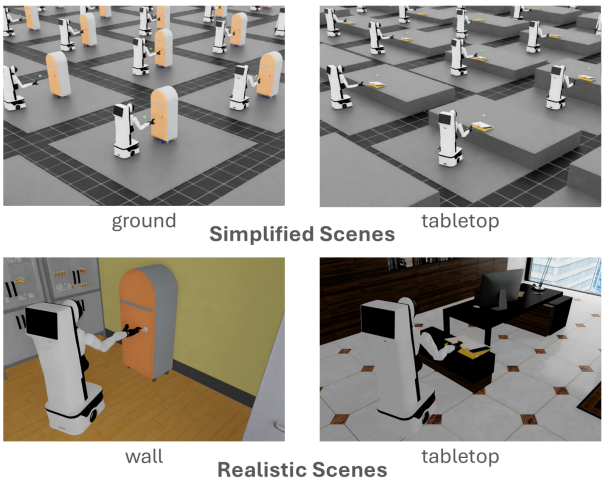

Fig. 3: Illustrations of simplified scenes for MobileManiRL training and realistic scenes for MobileManiDataset generation and MobileManiVLA evaluation.

4.2 MobileManiDataset Generation

Realistic Scene Rendering. Each trained MobileManiRL policy is now able to perform its assigned robot-object-skill across diverse robot initial poses. We then place the 20 categories of objects into 5 realistic scene settings for trajectory rendering: space (cart), wall (toilet, trashcan, refrigerator, washing machine, fridge, cabinet, table), door (window, lever door, round door), outdoor (car), and tabletop (box, laptop, dishwasher, oven, microwave, safe, faucet, holistic objects). For each scene setting, we manually annotate 20 scene placements using digital assets from the Isaac Sim [34] and Genie Sim [45], yielding a total of 80 seen scene placements for VLA training and 20 unseen reserved for testing. These scenes cover a variety of everyday environments, including kitchen, bedroom, hospital, office, warehouse, and parking lot, as shown in the Appendix.

Dataset Analysis. For each of the G1 robot and the XHand robot, MobileManiDataset comprises 630 objects across 20 categories, manipulated through 5 skills and over 100 tasks using 1,182 robot-object-skill combinations of MobileManiRL, which are further distributed across 100 scene placements. The dataset is split into 506 objects and 80 scenes for VLA training, plus 124 objects and 20 scenes for testing, yielding a total of 15,232 robot-object–skill–scene combinations for training and 920 combinations for testing. For each training combination, we generate 10 successful manipulation trajectories in Isaac Sim, resulting in 150K training trajectories. Each manipulation trajectory, $\mathcal{T} = \{L, (I_1, S_1, A_1), \ldots, (I_t, S_t, A_t), \ldots, (I_T, S_T, A_T)\}$, is recorded at 30 FPS with an average length of 160 frames, including one natural language instruction L; synchronized 520×520 RGB, depth, and segmentation images I_t from both head-view and wrist-view cameras; the corresponding object and robot states S_t; and the executed (6+D) dimensional action A_t at each timestep t. All states and actions are recorded in the global world coordinate frame.

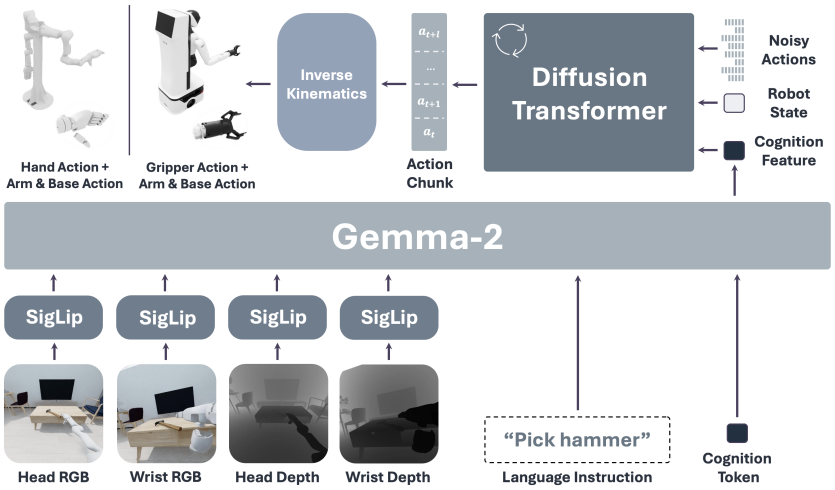

Fig. 4: Model architecture of MobileManiVLA with multi-modality inputs.

4.3 MobileManiVLA Training

The objective is to use the generated MobileManiDataset to train a universal vision-language-action model, MobileManiVLA, for each of the G1 robot and XHand robot. MobileManiVLA is capable of performing mobile-based robotic manipulation of diverse ground and tabletop objects across varying initial poses, and can generalize to both unseen objects and unseen scenes. Following the design of CogACT [25], we structure our MobileManiVLA into three components: the vision, language, and action modules, as shown in Figure 4.

Vision and Language Modules. The vision and language modules are initialized from a pretrained vision–language model, PaliGemma-2 [41]. The vision module processes multi-view image inputs, including two RGB images and two depth images captured from the head-view and wrist-view cameras (each reshaped to $224 \times 224 \times 3$). These images are encoded via a SigLIP vision encoder into dense visual embeddings. In parallel, a language instruction of the form "<skill> <object>" (such as "open faucet" or "close door") is tokenized and projected into language embeddings. The vision and language embeddings are then fused via a linear projection for alignment. Together with a learnable *cognition token* c, they are fed into a Gemma-2 language model [44] to produce an output feature f_t^c that captures both perceptual and instructional context.

Action Module with State Conditions. The action module is implemented as a diffusion-transformer (DiT) that takes as input: 1) a series of noisy actions $(a_t^i, a_{t+1}^i, \ldots, a_{t+N}^i)$, where i denotes the current denoising step; 2) the encoded cognition feature f_t^c; and 3) the state feature f_t^s, obtained by encoding the 6-d robot wrist pose using a lightweight MLP. Conditioned on f_t^c and f_t^s, the DiT iteratively predicts the clean actions $(a_t, a_{t+1}, \ldots, a_{t+N})$ over multiple denoising steps, progressively refining its predictions toward physically consistent and goal-directed motions. All states and actions are expressed in the robot mobile base coordinate frame for consistency across scenes.

Training Objective and Inference Strategy. The vision, language, and action modules are trained end-to-end by minimizing the mean squared error (MSE) between the predicted and the ground truth Gaussian noises, with the loss function defined as:

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1),\, i} \|\hat{\epsilon}^i - \epsilon\|_2, \tag{2}$$

where $\hat{\epsilon}^i$ is the predicted noise for the noisy action sequence $(a_t^i, a_{t+1}^i, \ldots, a_{t+N}^i)$ at the i-th denoising step, and ϵ is the corresponding ground truth Gaussian noise.

During inference, MobileManiVLA predicts a chunk of actions with length $N = 16$ conditioned on the current instruction and observations. we adopt the adaptive ensemble strategy proposed in CogACT [25] to improve trajectory smoothness and robustness, with a window size of $K = 4$.

5 Experiment

For each of the G1 robot and XHand robot, we first train MobileManiRL on each of the 1,182 robot-object-skill combinations, with 1,024 simulation environments, a learning rate of $1e^{-3}$, and 4K iterations. Training is conducted in parallel on 32 NVIDIA V100 GPUs and requires about 4 days per robot. We then collect 15,232 robot-object-skill-scene combinations for training and 920 combinations for testing to generate the MobileManiDataset, resulting in 150K training trajectories. Trajectory generation is conducted in parallel on 8 NVIDIA RTX A6000 GPUs and requires about 6 days per robot. Finally, We train the MobileManiVLA model on MobileManiDataset, with a batch size of 480, a learning rate of $2e^{-5}$, and 320K iterations. Training is conducted on 8 NVIDIA B200 GPUs and takes about 12 days per robot.

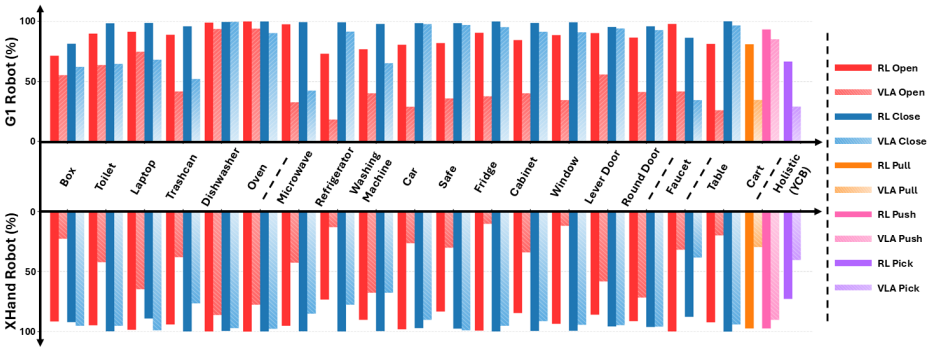

Fig. 5: Success rates of MobileManiRL and MobileManiVLA on the G1 robot and XHand robot across 20 object categories and 5 mobile manipulation skills. In terms of manipulation motion patterns, objects like box and oven require lid-flipping upward or downward; car, fridge, and door require door-handle grasping followed by pivoting left or right; faucet requires handle rotation; table and cart require handle grasping with pulling or pushing; holistic (ycb) objects require object grasping and lifting.

5.1 Main Results

Figure 5 demonstrates the success rates of MobileManiRL and MobileManiVLA across the 20 object categories, while Table 3 summarizes the mean success rates over the 5 mobile manipulation skills. MobileManiRL is evaluated on seen objects with 1024 episodes per robot-object-skill combination, whereas MobileManiVLA is evaluated on unseen objects and scenes, with 10 episodes per robot-object-skill-scene combination. Both experiments are conducted with randomized robot initial poses to assess robustness.

Skill	MobileManiRL (%)		MobileManiVLA (%)	
	G1 Robot	XHand Robot	G1 Robot	XHand Robot
Open	86.6	**91.9**	**42.9**	34.5
Close	96.2	**96.5**	75.8	**77.5**
Pull	80.8	**97.3**	**34.4**	29.4
Push	93.1	**97.2**	85.1	**90.0**
Pick	66.4	**72.6**	28.8	**40.2**
Mean	89.6	**92.9**	56.7	**57.3**

Table 3: Success rates of MobileManiRL and MobileManiVLA on the G1 robot and XHand robot across 5 manipulation skills.

Takeaway 1 — Object structure and skill complexity drive performance variance. For both MobileManiRL and MobileManiVLA, object categories like toilet, laptop, and dishwasher achieve higher success rates due to their relatively simple structures and motion patterns, typically involving lid-flipping without requiring complex handle grasping. In contrast, object categories such as car, door, and table present greater challenges, as they demand precise handle localization and stable grasp control. Regarding the mobile manipulation skills, open, pull, and pick, which require accurate and stable grasping of the object, are generally more difficult to learn than close and push.

Takeaway 2 — Generalization remains challenging for VLA models with the implicit inputs. Compared with MobileManiRL, the success rates of MobileManiVLA exhibit a moderate decline for two primary reasons. First, MobileManiRL is trained and evaluated individually on each seen robot–object–skill combination, whereas MobileManiVLA is trained as a universal model and evaluated on unseen objects and unseen scenes, which increases diversity and generalization difficulty. Second, MobileManiRL leverages explicit state-based observations, such as the world positions of the robot gripper/hand points, the object grasp and goal points, providing precise geometric information. In contrast, MobileManiVLA relies solely on implicit sensory inputs, including a language instruction, multi-view RGB-D images, and the robot wrist pose represented in the mobile base coordinate frame. This transition from explicit to implicit observations introduces additional learning challenges but enables MobileManiVLA to generalize effectively to unseen objects and unseen scenes in both simulation and the real world. The real world inference of MobileManiVLA is demonstrated in the Appendix.

Takeaway 3 — Dexterous hands improve manipulation precision and versatility. As shown in Table 3, with the MobileManiRL, the XHand robot achieves a higher success rate of **92.9%**, compared to **89.6%** for the G1 robot, highlighting the advantages of dexterous manipulation, particularly on the open, pull, and pick skills. While with MobileManiVLA, the G1 robot and XHand robot achieve comparable success rates of **56.7%** and **57.3%**, respectively. In this setting, the XHand robot performs worse on the open and pull skills, where its dexterous fingers often collide with surrounding object surfaces, preventing stable handle grasps. Such disturbance between the object body (e.g., door faces) and the dexterous fingers significantly degrades the VLA performance. Nevertheless, with MobileManiVLA , the XHand robot outperforms the G1 robot on the pick skill, where the dexterous fingers can easily establish the force-closure grasps on the holistic objects.

5.2 Ablation Studies and Model Comparisons

We conduct ablation studies and model comparisons on a subset of MobileManiDataset that includes only the challenging skills: open, pull, and pick for the G1 robot, covering 272 training and 66 testing objects across 6 categories: laptop, cabinet, faucet, table, cart and holistic objects, providing 4,352 robot-object-skill-scene combinations for training and 264 unseen combinations for testing. All VLA models are trained with a batch size of 120 and 160K iterations, which are evaluated on unseen objects and scenes with 10 episodes per combination.

Image Inputs				State Inputs			Success Rate
Head RGB	Head Depth	Wrist RGB	Wrist Depth	Wrist Pose	Grasp Point	Goal Point	(%)
✓				✓			7.9
✓	✓			✓			14.1
✓		✓		✓			14.9
✓	✓	✓	✓	✓			**28.2**
✓	✓	✓	✓				22.4
✓	✓	✓	✓	✓	✓		32.3
✓	✓	✓	✓	✓	✓	✓	**36.6**

Table 4: Ablation studies of MobileManiVLA on the G1 robot with different image inputs (top rows) and state inputs (bottom rows).

Takeaway 4 — Multi-view and multi-modality visual inputs significantly enhance policy generalization. Table 4 shows the success rates of MobileManiVLA with different image inputs. Using only the head-view RGB image yields a success rate of only 7.9%, indicating that a single global view is insufficient for precise manipulation in mobile-based settings. Adding either the head-view depth image or the wrist-view RGB image improves the performance to 14.1% and 14.9% respectively, suggesting that additional geometric or visual cues enhance spatial understanding. Finally, combining both RGB-D

images from the head-view and wrist-view cameras achieves the highest success rate of 28.2%, confirming that multi-view and multi-modality visual inputs significantly enhance manipulation accuracy and generalization ability.

Takeaway 5 — State inputs from robot proprioception or simple object detection further improve performance. Table 4 reports the impact of different state inputs on the performance of MobileManiVLA. Compared with no state input, encoding the robot wrist pose in its mobile base coordinate frame increases the success rate from 22.4% to 28.2%, highlighting the importance of proprioceptive information for precise manipulation. Further incorporating the pseudo-observations, such as the first-frame positions of the object grasp point and goal point expressed in the robot mobile base coordinate frame, leads to additional performance gains, confirming that even coarse spatial priors about the object can enhance policy grounding and performance.

Uneen Object	Unseen Scene	Seen Object	Seen Scene	Success Rate (%)
		✓	✓	59.6
	✓	✓		51.3
✓			✓	39.2
✓	✓			28.2

Table 5: Ablation studies of MobileManiVLA on the G1 robot with seen or unseen objects and scenes.

Model	Success Rate (%)
OpenVLA [22] (7B)	4.5
CogACT [25] (7B)	6.8
π_0 [5] (3B)	11.2
$\pi_{0.5}$ [17] (3B)	18.8
MobileManiVLA (3B)	**28.2**

Table 6: Model comparisons of representative VLA models with their default architectures and input modalities on the G1 robot.

Takeaway 6 — Unseen objects present greater challenges than unseen scenes. Table 5 indicates the success rates of MobileManiVLA on seen or unseen objects and scenes. The model achieves the highest success rate of 59.6% when both objects and scenes are seen during testing. When evaluated with seen objects in unseen scenes or unseen objects in seen scenes, the success rates drop to 51.3% and 39.2% respectively, indicating that generalizing to unseen object structures presents greater challenges than adapting to new scene layouts. The lowest success rate of 28.2% occurs when both objects and scenes are unseen, reflecting the compounded difficulty of transferring manipulation skills to entirely novel environments.

Takeaway 7 — MobileManiBench provides a unified platform for training and evaluating VLA models. By offering standardized training and evaluation protocols across diverse robots, objects, tasks, scenes and input modalities, MobileManiBench enables consistent benchmarking of existing VLA models with their default architectures and inputs, as shown in Table 6. OpenVLA and CogACT receive only head-view RGB image, resulting in low success rates of 4.5% and 6.8%, respectively. π_0 and $\pi_{0.5}$ process both head-view and wrist-view RGB images together with wrist pose states, achieving improved success rates of 11.2% and 18.8%, respectively. The stronger performance of $\pi_{0.5}$ arises from its pretraining on more mobile manipulation trajectories. Overall, MobileManiVLA achieves the highest success rate of 28.2% by jointly leveraging head-view and wrist-view RGB-D images along with wrist pose states, highlighting the advantages of multi-view and multi-modality inputs.

Fixed Base	Mobile Base	Success Rate (%)
	✓	**82.8**
✓		25.4

Table 7: Success rates of MobileManiRL on the G1 robot with fixed or mobile base.

Takeaway 8 — Base mobility is crucial for effective manipulation beyond tabletop tasks. To evaluate the importance of base mobility during manipulation, we fix the G1 robot base and initialize it closer to the object grasp point than before ($0.5m$ to $1.0m$ along the x and y axes). We train MobileManiRL under this fixed-base setting on several robot-object-skill combinations, including open laptop, open cabinet, open faucet, open table, pull cart and pick holistic (YCB) objects. As shown in Table 7, the success rates of MobileManiRL drop drastically when the robot base is fixed, confirming that base mobility is essential for spatial manipulation beyond static tabletop interactions.

6 Conclusion

We introduce MobileManiBench, a large-scale simulation-based benchmark designed to simplify model verification for mobile-based robotic manipulation. MobileManiBench encompasses 2 mobile-based robots with either parallel grippers or dexterous hands, 630 objects spanning 20 categories, and 5 mobile manipulation skills across 100+ tasks. By training a universal MobileManiRL policy on each robot-object-skill triplet, the benchmark autonomously generates the MobileManiDataset in 100 realistic scenes, which contains 300K trajectories with 3 data modalities: language instructions, multi-view RGB-depth-segmentation images, synchronized object/robot states and actions. Experiments demonstrate that MobileManiBench serves as a unified and scalable platform for developing and evaluating next-generation vision–language–action models, accelerating research toward general-purpose, embodied intelligence.

References

1. AgiBot-World-Contributors, Bu, Q., Cai, J., Chen, L., Cui, X., Ding, Y., Feng, S., Gao, S., He, X., Hu, X., Huang, X., Jiang, S., Jiang, Y., Jing, C., Li, H., Li, J., Liu, C., Liu, Y., Lu, Y., Luo, J., Luo, P., Mu, Y., Niu, Y., Pan, Y., Pang, J., Qiao, Y., Ren, G., Ruan, C., Shan, J., Shen, Y., Shi, C., Shi, M., Shi, M., Sima, C., Song, J., Wang, H., Wang, W., Wei, D., Xie, C., Xu, G., Yan, J., Yang, C., Yang, L., Yang, S., Yao, M., Zeng, J., Zhang, C., Zhang, Q., Zhao, B., Zhao, C., Zhao, J., Zhu, J.: Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems (2025), https://arxiv.org/abs/2503.06669
2. Bao, C., Xu, H., Qin, Y., Wang, X.: Dexart: Benchmarking generalizable dexterous manipulation with articulated objects (2023), https://arxiv.org/abs/2305.05706
3. Beyer, L., Steiner, A., Pinto, A.S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., et al.: Paligemma: A versatile 3b vlm for transfer. arXiv preprint arXiv:2407.07726 (2024)
4. Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., Fang, Y., Fox, D., Hu, F., Huang, S., et al.: Gr00t n1: An open foundation model for generalist humanoid robots. arXiv preprint arXiv:2503.14734 (2025)
5. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: π_0: A vision-language-action flow model for general robot control. arXiv:2410.24164 (2024)
6. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al.: RT-1: Robotics transformer for real-world control at scale. arXiv:2212.06817 (2022)
7. Calli, B., Singh, A., Walsman, A., Srinivasa, S.S., Abbeel, P., Dollar, A.M.: The ycb object and model set: Towards common benchmarks for manipulation research. In: Proceedings of the IEEE International Conference on Advanced Robotics (ICAR). pp. 510–517 (2015)
8. Cheang, C., Chen, S., Cui, Z., Hu, Y., Huang, L., Kong, T., Li, H., Li, Y., Liu, Y., Ma, X., et al.: Gr-3 technical report. arXiv preprint arXiv:2507.15493 (2025)
9. Chen, J., Liang, H., Du, L., Wang, W., Hu, M., Mu, Y., Wang, W., Dai, J., Luo, P., Shao, W., Shao, L.: Owmm-agent: Open world mobile manipulation with multimodal agentic data synthesis (2025), https://arxiv.org/abs/2506.04217
10. Chen, T., Chen, Z., Chen, B., Cai, Z., Liu, Y., Li, Z., Liang, Q., Lin, X., Ge, Y., Gu, Z., Deng, W., Guo, Y., Nian, T., Xie, X., Chen, Q., Su, K., Xu, T., Liu, G., Hu, M., ang Gao, H., Wang, K., Liang, Z., Qin, Y., Yang, X., Luo, P., Mu, Y.: Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation (2025), https://arxiv.org/abs/2506.18088
11. Chen, X., Djolonga, J., Padlewski, P., Mustafa, B., Changpinyo, S., Wu, J., Ruiz, C.R., Goodman, S., Wang, X., Tay, Y., et al.: Pali-x: On scaling up a multilingual vision and language model. arXiv preprint arXiv:2305.18565 (2023)
12. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., et al.: Palm-e: An embodied multimodal language model (2023)
13. Fu, Z., Zhao, T.Z., Finn, C.: Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation (2024), https://arxiv.org/abs/2401.02117

14. Gu, J., Xiang, F., Li, X., Ling, Z., Liu, X., Mu, T., Tang, Y., Tao, S., Wei, X., Yao, Y., Yuan, X., Xie, P., Huang, Z., Chen, R., Su, H.: Maniskill2: A unified benchmark for generalizable manipulation skills (2023), https://arxiv.org/abs/2302.04659
15. Han, S., Qiu, B., Liao, Y., Huang, S., Gao, C., Yan, S., Liu, S.: Robocerebra: A large-scale benchmark for long-horizon robotic manipulation evaluation (2025), https://arxiv.org/abs/2506.06677
16. Huang, W., Wang, C., Li, Y., Zhang, R., Fei-Fei, L.: Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation (2024), https://arxiv.org/abs/2409.01652
17. Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., et al.: $\pi_{0.5}$: a vision-language-action model with open-world generalization. arXiv:2504.16054 (2025)
18. James, S., Ma, Z., Arrojo, D.R., Davison, A.J.: Rlbench: The robot learning benchmark & learning environment (2019), https://arxiv.org/abs/1909.12271
19. Karamcheti, S., Nair, S., Balakrishna, A., Liang, P., Kollar, T., Sadigh, D.: Prismatic vlms: Investigating the design space of visually-conditioned language models. In: Forty-first International Conference on Machine Learning (2024)
20. Khazatsky, A., Pertsch, K., Nair, S., Balakrishna, A., Dasari, S., Karamcheti, S., Nasiriany, S., Srirama, M.K., Chen, L.Y., Ellis, K., Fagan, P.D., Hejna, J., Itkina, M., Lepert, M., Ma, Y.J., Miller, P.T., Wu, J., Belkhale, S., Dass, S., Ha, H., Jain, A., Lee, A., Lee, Y., Memmel, M., Park, S., Radosavovic, I., Wang, K., Zhan, A., Black, K., Chi, C., Hatch, K.B., Lin, S., Lu, J., Mercat, J., Rehman, A., Sanketi, P.R., Sharma, A., Simpson, C., Vuong, Q., Walke, H.R., Wulfe, B., Xiao, T., Yang, J.H., Yavary, A., Zhao, T.Z., Agia, C., Baijal, R., Castro, M.G., Chen, D., Chen, Q., Chung, T., Drake, J., Foster, E.P., Gao, J., Guizilini, V., Herrera, D.A., Heo, M., Hsu, K., Hu, J., Irshad, M.Z., Jackson, D., Le, C., Li, Y., Lin, K., Lin, R., Ma, Z., Maddukuri, A., Mirchandani, S., Morton, D., Nguyen, T., O'Neill, A., Scalise, R., Seale, D., Son, V., Tian, S., Tran, E., Wang, A.E., Wu, Y., Xie, A., Yang, J., Yin, P., Zhang, Y., Bastani, O., Berseth, G., Bohg, J., Goldberg, K., Gupta, A., Gupta, A., Jayaraman, D., Lim, J.J., Malik, J., Martín-Martín, R., Ramamoorthy, S., Sadigh, D., Song, S., Wu, J., Yip, M.C., Zhu, Y., Kollar, T., Levine, S., Finn, C.: Droid: A large-scale in-the-wild robot manipulation dataset (2025), https://arxiv.org/abs/2403.12945
21. Kim, M.J., Finn, C., Liang, P.: Fine-tuning vision-language-action models: Optimizing speed and success. arXiv:2502.19645 (2025)
22. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: OpenVLA: An open-source vision-language-action model. arXiv:2406.09246 (2024)
23. Li, C., Zhang, R., Wong, J., Gokmen, C., Srivastava, S., Martín-Martín, R., Wang, C., Levine, G., Ai, W., Martinez, B., Yin, H., Lingelbach, M., Hwang, M., Hiranaka, A., Garlanka, S., Aydin, A., Lee, S., Sun, J., Anvari, M., Sharma, M., Bansal, D., Hunter, S., Kim, K.Y., Lou, A., Matthews, C.R., Villa-Renteria, I., Tang, J.H., Tang, C., Xia, F., Li, Y., Savarese, S., Gweon, H., Liu, C.K., Wu, J., Fei-Fei, L.: Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation (2024), https://arxiv.org/abs/2403.09227
24. Li, Q., Deng, Y., Liang, Y., Luo, L., Zhou, L., Yao, C., Zeng, L., Feng, Z., Liang, H., Xu, S., et al.: Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. arXiv preprint arXiv:2510.21571 (2025)

25. Li, Q., Liang, Y., Wang, Z., Luo, L., Chen, X., Liao, M., Wei, F., Deng, Y., Xu, S., Zhang, Y., et al.: CogACT: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. arXiv:2411.19650 (2024)
26. Li, X., Liu, M., Zhang, H., Yu, C., Xu, J., Wu, H., Cheang, C., Jing, Y., Zhang, W., Liu, H., et al.: Vision-language foundation models as effective robot imitators. In: ICLR (2022)
27. Li, X., Hsu, K., Gu, J., Pertsch, K., Mees, O., Walke, H.R., Fu, C., Lunawat, I., Sieh, I., Kirmani, S., Levine, S., Wu, J., Finn, C., Su, H., Vuong, Q., Xiao, T.: Evaluating real-world robot manipulation policies in simulation. arXiv preprint arXiv:2405.05941 (2024)
28. Li, Y., Zhang, X., Wu, R., Zhang, Z., Geng, Y., Dong, H., He, Z.: Unidoormanip: Learning universal door manipulation policy over large-scale and diverse door manipulation environments (2024), https://arxiv.org/abs/2403.02604
29. Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., Stone, P.: Libero: Benchmarking knowledge transfer for lifelong robot learning (2023), https://arxiv.org/abs/2306.03310
30. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS (2023)
31. Liu, S., Wu, L., Li, B., Tan, H., Chen, H., Wang, Z., Xu, K., Su, H., Zhu, J.: RDT-1B: A diffusion foundation model for bimanual manipulation. ICLR (2025)
32. Mees, O., Hermann, L., Rosete-Beas, E., Burgard, W.: Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks (2022), https://arxiv.org/abs/2112.03227
33. Nasiriany, S., Maddukuri, A., Zhang, L., Parikh, A., Lo, A., Joshi, A., Mandlekar, A., Zhu, Y.: Robocasa: Large-scale simulation of everyday tasks for generalist robots (2024), https://arxiv.org/abs/2406.02523
34. NVIDIA: Isaac Sim, https://github.com/isaac-sim/IsaacSim
35. O'Neill, A., Rehman, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., et al.: Open X-Embodiment: Robotic learning datasets and RT-X models. In: ICRA (2024)
36. Qu, D., Song, H., Chen, Q., Yao, Y., Ye, X., Ding, Y., Wang, Z., Gu, J., Zhao, B., et al.: SpatialVLA: Exploring spatial representations for visual-language-action model. arXiv:2501.15830 (2025)
37. Robotera: Xhand (2025), https://www.robotera.com/en/
38. Robotics, A.: Agibot g1 mobile manipulator (2025), https://www.agibot.com/products/G1
39. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms (2017), https://arxiv.org/abs/1707.06347
40. Shi, H., Xie, B., Liu, Y., Sun, L., Liu, F., Wang, T., Zhou, E., Fan, H., Zhang, X., Huang, G.: Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation. arXiv preprint arXiv:2508.19236 (2025)
41. Steiner, A., Pinto, A.S., Tschannen, M., Keysers, D., Wang, X., Bitton, Y., Gritsenko, A., Minderer, M., Sherbondy, A., Long, S., et al.: Paligemma 2: A family of versatile vlms for transfer. arXiv preprint arXiv:2412.03555 (2024)
42. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv:2312.11805 (2023)
43. Team, G.R., Abeyruwan, S., Ainslie, J., Alayrac, J.B., Arenas, M.G., Armstrong, T., Balakrishna, A., Baruch, R., Bauza, M., Blokzijl, M., et al.: Gemini robotics: Bringing ai into the physical world. arXiv preprint arXiv:2503.20020 (2025)

44. Team, G., Riviere, M., Pathak, S., Sessa, P.G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al.: Gemma 2: Improving open language models at a practical size. arXiv preprint arXiv:2408.00118 (2024)
45. Team, G.S.: Genie sim assets. https://github.com/AgibotTech/genie_sim (2025)
46. Team, O.M., Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Kreiman, T., Xu, C., et al.: Octo: An open-source generalist robot policy. arXiv:2405.12213 (2024)
47. Wang, H., Chen, J., Huang, W., Ben, Q., Wang, T., Mi, B., Huang, T., Zhao, S., Chen, Y., Yang, S., Cao, P., Yu, W., Ye, Z., Li, J., Long, J., Wang, Z., Wang, H., Zhao, Y., Tu, Z., Qiao, Y., Lin, D., Pang, J.: Grutopia: Dream general robots in a city at scale (2024), https://arxiv.org/abs/2407.10943
48. Wang, W., Li, G., Zamora, M., Coros, S.: Trtm: Template-based reconstruction and target-oriented manipulation of crumpled cloths (2024), https://arxiv.org/abs/2308.04670
49. Wang, W., Wei, F., Zhou, L., Chen, X., Luo, L., Yi, X., Zhang, Y., Liang, Y., Xu, C., Lu, Y., Yang, J., Guo, B.: Unigrasptransformer: Simplified policy distillation for scalable dexterous robotic grasping (2025), https://arxiv.org/abs/2412.02699
50. Wang, Y., Li, X., Wang, W., Zhang, J., Li, Y., Chen, Y., Wang, X., Zhang, Z.: Unified vision-language-action model. arXiv:2506.19850 (2025)
51. Wen, J., Zhu, Y., Li, J., Tang, Z., Shen, C., Feng, F.: DexVLA: Vision-language model with plug-in diffusion expert for general robot control. In: CoRL (2025)
52. Xiang, F., Mo, K., Xia, Y., Liu, H., Zhang, F., Han, L., Guibas, L.J., Xiao, J., Dong, H., Yuan, Y., et al.: Partnet-mobility: A large-scale database for articulated objects. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 11359–11368 (2020)
53. Yang, J., Tan, R., Wu, Q., Zheng, R., Peng, B., Liang, Y., Gu, Y., Cai, M., Ye, S., Jang, J., et al.: Magma: A foundation model for multimodal ai agents. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14203–14214 (2025)
54. Zhang, H., Christen, S., Fan, Z., Hilliges, O., Song, J.: GraspXL: Generating grasping motions for diverse objects at scale. In: European Conference on Computer Vision (ECCV) (2024)
55. Zhang, S., Xu, Z., Liu, P., Yu, X., Li, Y., Gao, Q., Fei, Z., Yin, Z., Wu, Z., Jiang, Y.G., Qiu, X.: Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks (2024), https://arxiv.org/abs/2412.18194
56. Zhang, W., Liu, H., Qi, Z., Wang, Y., Yu, X., Zhang, J., Dong, R., He, J., Lu, F., Wang, H., et al.: Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. arXiv preprint arXiv:2507.04447 (2025)
57. Zhao, Z., Jing, H., Liu, X., Mao, J., Jha, A., Yang, H., Xue, R., Zakharor, S., Guizilini, V., Wang, Y.: Humanoid everyday: A comprehensive robotic dataset for open-world humanoid manipulation (2025), https://arxiv.org/abs/2510.08807
58. Zhong, Y., Huang, X., Li, R., Zhang, C., Chen, Z., Guan, T., Zeng, F., Lui, K.N., et al.: DexGraspVLA: A vision-language-action framework towards general dexterous grasping. arXiv:2502.20900 (2025)
59. Zhu, Y., Wong, J., Mandlekar, A., Martín-Martín, R., Joshi, A., Lin, K., Maddukuri, A., Nasiriany, S., Zhu, Y.: robosuite: A modular simulation framework and benchmark for robot learning. In: arXiv preprint arXiv:2009.12293 (2020)
60. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: RT-2: Vision-language-action models transfer web knowledge to robotic control. In: CoRL (2023)