RoboMIND: Benchmark on Multi-embodiment Intelligence Normative Data for Robot Manipulation

Kun Wu[1,*], Chengkai Hou[2,3,*], Jiaming Liu[2,3,*], Zhengping Che[1,*,†], Xiaozhu Ju[1,*,†],
Zhuqin Yang[1], Meng Li[1], Yinuo Zhao[1], Zhiyuan Xu[1], Guang Yang[1], Shichao Fan[1], Xinhua Wang[1], Fei Liao[1],
Zhen Zhao[1], Guangyu Li[1], Zhao Jin[1], Lecheng Wang[1], Jilei Mao[1], Ning Liu[1], Pei Ren[1], Qiang Zhang[1],
Yaoxu Lyu[2], Mengzhen Liu[2,3], Jingyang He[2,3], Yulin Luo[2,3], Zeyu Gao[3], Chenxuan Li[2], Chenyang Gu[2,3],
Yankai Fu[2], Di Wu[2], Xingyu Wang[2], Sixiang Chen[2,3], Zhenyu Wang[2,3], Pengju An[2,3], Siyuan Qian[2,3],
Shanghang Zhang[2,3,✉], Jian Tang[1,✉]
[1]Beijing Innovation Center of Humanoid Robotics
[2]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University
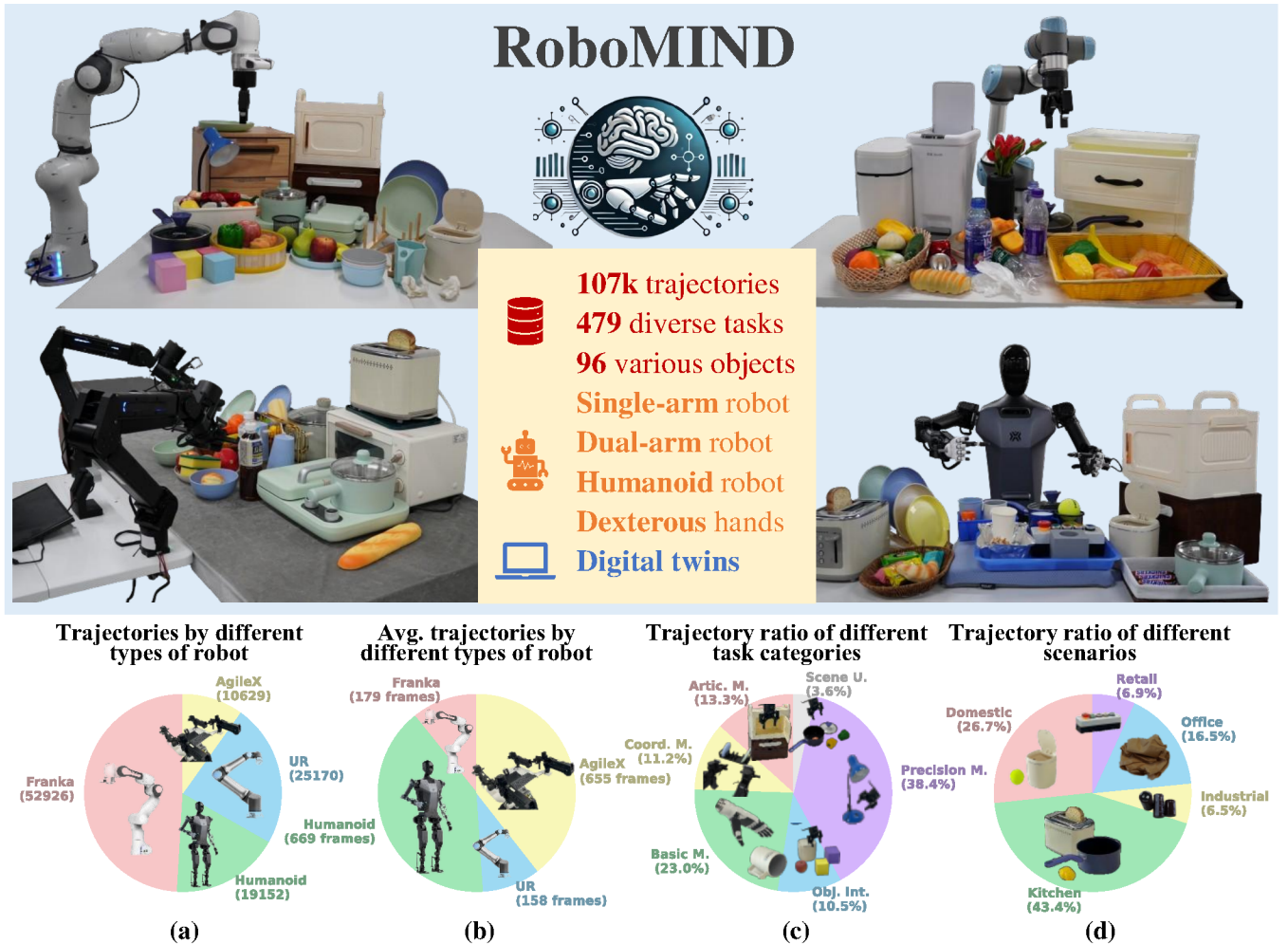[3]Beijing Academy of Artificial Intelligence

Fig. 1: **Overview of RoboMIND. We introduce RoboMIND (Multi-embodiment Intelligence Normative Data for Robot Manipulation), comprising 107k demonstration trajectories across 479 diverse tasks involving 96 distinct object classes. To ensure consistency and reliability during policy learning, RoboMIND is gathered through human teleoperation and structured around a unified data collection standard. The four pie charts represent (a) the total trajectory numbers categorized by different types of robots, (b) average trajectory lengths (frames) categorized by different types of robots, (c) trajectory ratio of different task categories (Artic. M.: Articulated Manipulations; Coord. M.: Coordination Manipulations; Basic M.: Basic Manipulations; Obj. Int.: Multiple Object Interactions; Precision M.: Precision Manipulations; Scene U.: Scene Understanding), and (d) trajectory ratio of different scenarios.**

Abstract—Developing robust and general-purpose manipulation policies is a key goal in robotics. To achieve effective generalization, it is essential to construct comprehensive datasets that encompass a large number of demonstration trajectories and diverse tasks. Unlike vision or language data, which can be sourced from the internet, robotic datasets require detailed observations and manipulation actions, necessitating significant investments in both hardware-software infrastructure and human labor. While existing works have focused on assembling various individual robot datasets, there is still a lack of a unified data collection standard and insufficient high-quality data across diverse tasks, scenarios, and robot types. In this paper, we introduce RoboMIND (Multi-embodiment Intelligence Normative Data for Robot Manipulation), a dataset containing 107k demonstration trajectories across 479 diverse tasks involving 96 object classes. RoboMIND is collected through human teleoperation and encompasses comprehensive robotic-related information, including multi-view observations, proprioceptive robot state information, and linguistic task descriptions. To ensure data consistency and reliability for imitation learning, RoboMIND is built on a unified data collection platform and a standardized protocol, covering four distinct robotic embodiments: the Franka Emika Panda, the X-Humannoid Tien Kung humanoid robot with dual dexterous hands, the AgileX dual-arm robot, and the UR5e. Our dataset also includes 5k real-world failure demonstrations, each accompanied by detailed causes, enabling failure reflection and correction during policy learning. Additionally, we created a digital twin environment in the Isaac Sim simulator, replicating the real-world tasks and assets, which facilitates the low-cost collection of additional training data and enables efficient evaluation. To demonstrate the quality and diversity of our dataset, we conducted extensive experiments using various imitation learning methods for single-task settings and state-of-the-art Vision-Language-Action (VLA) models for multi-task scenarios. By leveraging RoboMIND, the VLA models achieved high manipulation success rates and demonstrated strong generalization capabilities. To the best of our knowledge, RoboMIND is the largest multi-embodiment teleoperation dataset collected on a unified platform, providing large-scale and high-quality robotic training data. Our project is at https://x-humanoid-robomind.github.io/.

I. INTRODUCTION

One of the aspirations of any professional in the field of robotics is to develop a versatile, general-purpose robotic model capable of performing a broad spectrum of real-world tasks. Specifically, such models should be generalizable in order to execute the intended manipulation tasks under varying conditions, such as a new robot, unfamiliar environments, or different objects [77, 50, 51, 65, 64, 14]. To achieve this level of generalization, researchers have drawn inspiration from the training of large models in computer vision and natural language processing, where rich and diverse datasets have proven essential [1, 61, 96, 101, 32, 55]. They concluded that for training generalizable robotic models, one of the most

critical elements is the access to rich and diverse training data that encompass varied scenes, tasks, and robot types. Such diversity ensures that models learn to perform reliably under different conditions and environments [70, 77, 92, 14, 29, 97]. Therefore, in this work, we aim to *construct comprehensive datasets that capture a broad spectrum of robotic interactions and experiences to facilitate training models capable of mastering various manipulation policies.*

However, the curation of large-scale datasets for training general-purpose robotic models poses significant challenges. In contrast to the acquisition of vision or language data, which can often be sourced through web-based collection methods [32, 55], collecting robotic data is difficult because such data cannot be easily obtained in the same way, as it requires controlled environments where the joints and end-effector information of robotic systems are meticulously recorded. Moreover, scaling up data collection efforts necessitates considerable investment in both hardware and software infrastructure and human labor for oversight, particularly when it comes to acquiring and curating high-quality demonstration data [77, 102, 50]. Consequently, even the most versatile robotic manipulation policies currently in use are predominantly trained on datasets gathered within constrained conditions that offer limited diversity in robot types [77, 50].

Our dataset, called RoboMIND (**M**ulti-embodiment **I**ntelligence **N**ormative **D**ata for **Robot** manipulation), is an extensive dataset that encompasses a broad range of robotic interactions and experiences. RoboMIND features **107k** demonstration trajectories amounting to 305.5 hours of interaction data of 4 kinds of robotic embodiments including Franka Emika Panda [34], a humanoid robot (i.e., X-Humanoid Tien Kung [12]), AgileX Cobot Magic V2.0 [87], and UR5e [88], as shown in Figure 1. Unlike the Open X-Embodiment dataset [77], which was compiled from various laboratories with differing data collection standards and diverse combinations of robotic platforms, RoboMIND is gathered within the same standardized setting, adhering to a standardized data collection protocol to ensure consistency and reliability. By maintaining uniform data collection standards, all data points are captured under similar conditions, reducing variability and noise, which is crucial for training models that can generalize well across different tasks and environments. The standardized procedures also enhance the reliability of the dataset, making it easier to validate and reproduce experimental results, thereby building trust in the trained models and ensuring their consistent performance in real-world applications.

Moreover, RoboMIND covers a wide range of robot environments and spans 479 diverse tasks involving 96 various object classes. Additionally, we provide a dataset from our real-world tasks simulated in the Nvidia Isaac Sim [76]. RoboMIND incorporates data from various robot types, including 26,856 motion trajectories from Franka Emika Panda single-arm robots, 15,187 from Tien Kung humanoid robots, 10,269 from AgileX Cobot Magic V2.0 dual-arm robots, 25,170 from UR5e single-arm robots, and 30,035 from simulation. All these

[1]Beijing Innovation Center of Humanoid Robotics, Beijing, China {*Gongda.Wu, z.che, jason.ju, jian.tang*}@x-humanoid.com
[2]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University, Beijing, China *shanghang@pku.edu.cn*
*Co-first authors: Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, and Xiaozhu Ju
[†]Project leaders: Zhengping Che and Xiaozhu Ju
[✉]Corresponding authors: Shanghang Zhang and Jian Tang

TABLE I: Comparison to existing real-world datasets for robot manipulation. All data is drawn from the original paper or from the DROID paper [50]. We divide robot types into three categories: single-arm, dual-arm, and humanoid. We report the number of unique multi-view trajectories and highlight the advantages of RoboMIND in orange. ‡non-robot, tool-based data collections. §not a dataset in itself, but an *aggregation* of existing datasets.

Dataset	Trajectory	Task	Skill	Arm	Dexterous Hand	Detailed Annotation	Robot Type	Public Robot	Failure Data	Digital Twin	Collection
Pinto and Gupta [83]	50k	n/a	1	Dual	✗	✗	1	✓	✓	✗	Scripted
Home-LCA [38]	28k	n/a	1	Single	✗	✗	1	✗	✗	✗	Scripted
BrainRobotData [54]	800k	n/a	1	Single	✗	✗	1	✗	✓	✗	Scripted
Roboturk [70]	2.1k	3	2	Single	✗	✗	1	✗	✗	✗	Human Teleoperation
MIME [93]	8.2k	20	20	Single+Dual	✗	✓	1	✗	✗	✗	Human Teleoperation
Sketchy [10]	74.4k	5	n/a	Single	✗	✓	1	✗	✓	✗	12% Human / 78% Scripted
RoboNet [22]	162k	n/a	n/a	Single	✗	✗	1	✓	✗	✗	Scripted
BridgeData [26]	7.2k	71	4	Single	✗	✗	1	✓	✗	✗	Human Teleoperation
MT-Opt [47]	800k	12	1	Single	✗	✗	1	✗	✗	✗	Scripted
RT-1 [9]	130k	700	8	Single	✗	✗	1	✗	✗	✗	Human Teleoperation
BC-Z [45]	26k	100	3	Single	✗	✗	1	✗	✗	✗	Human Teleoperation
BridgeData V2 [99]	60.1k	n/a	13	Single	✗	✗	1	✓	✗	✗	85% Human / 15% Scripted
RoboSet [8]	98.5k	38	6	Single	✗	✗	1	✓	✗	✗	30% Human / 70% Scripted
RH20T [29]	13k	140	33	Single	✗	✗	1	✓	✗	✗	Human Teleoperation
DROID [50]	76k	n/a	86	Single	✗	✗	1	✓	✗	✗	Human Teleoperation
BRMData [115]	0.5k	10	7	Dual	✗	✗	1	✓	✗	✗	Human Teleoperation
Dobb-E [92]‡	5.6k	109	6	Single	✗	✗	1	✓	✗	✗	Human Tool-based
Open X-Embodiment [77]§	1.4M	160k	217	Single+Dual	✗	✗	2	✓	✗	✗	Dataset Aggregation
RoboMIND	107k	479	38	Single+Dual	✓	✓	3	✓	✓	✓	Human Teleoperation

trajectories are collected through a teleoperation system that captures natural human motion patterns and maps them onto robots to drive the same motion trajectories. These trajectories encompass RGB-D data from distinct viewpoints, detailed proprioceptive state information of the robot body, specific information regarding the robot's end effector, and a linguistic description of the task at hand. Containing such comprehensive and detailed information, these data are valuable for training robotic models to perform complex manipulation tasks.

At the same time, we not only publish the 107k successful trajectories but also document the 5k trajectories of real-world failure cases. The robot model can explore the causes of failures by learning from these failure case trajectories, thereby improving its performance through such learning experiences. This technique is representative of Reinforcement Learning from Human Feedback (RLHF) [18, 81], where human oversight and feedback direct the learning process of models, leading the models to produce more desirable and accurate outcomes. In addition, we annotate a total of 10k robot trajectories in RoboMIND with frame-level fine-grained language descriptions. These annotated trajectories encompass a wide range of robot tasks. To ensure accuracy and reliability, each annotation undergoes verification and correction by multiple reviewers. We believe that these additional failure cases and fine-grained linguistic annotations will further advance research in robot learning, particularly in areas such as failure recovery [66], task planning [60], visual question answering [25], among others.

Beyond establishing such a large-scale and diverse dataset, we conduct extensive experiments to not only validate the dataset's effectiveness but also evaluate various algorithms' performance, providing a comprehensive benchmark analysis. Specifically, we evaluate the task success rates using single-task imitation learning methods, including ACT [116], Diffusion Policy [17], and BAKU [39]. Additionally, we assess the generalization capabilities and task success rates of Vision-Language-Action (VLA) large models such as OpenVLA [77], RDT-1B [65], and CrossFormer [24]. The experimental results demonstrate that RoboMIND can be effectively utilized by various single-task imitation learning algorithms and successfully adapted to VLA large models. The high-quality information provided by our dataset enables successful task execution across different approaches in real-world scenarios. Furthermore, pre-training the entire VLA models using the full RoboMIND dataset results in significant improvements in task performance across multiple robot types. To simplify the use of RoboMIND, we provide the code scripts that adapt RoboMIND files with the open-source LeRobot framework [11] at https://github.com/x-humanoid-robomind/x-humanoid-training-toolchain/.

II. RELATED WORK

Robotic Manipulation. Traditional manipulation policies typically rely on state-based reinforcement learning [3, 46, 111]. In contrast, recent works [73, 27, 28] incorporate visual observations as input to predict action poses. Imitation learning policies, in particular, enable robots to acquire stable manipulation skills by imitating an expert through demonstration [23, 104, 112]. Driven by advancements in diffusion-based generative models [41, 95, 89], diffusion policy [17] and subsequent works [82, 86, 105] focus on transforming random Gaussian noise into coherent action sequences, with methods such as DP3 [113] and 3D Diffuser Actor [49] further enhancing this process in 3D space. On the other hand, some Multimodal Large Language Models (MLLMs) [2, 25, 43] enable robots to comprehend natural language and visual scenes, automatically generating task plans. Meanwhile, Vision-Language-Action (VLA) models [117, 58, 57, 64, 51] empower MLLMs to predict low-level $SE(3)$ poses, demonstrating interpretability and generalization in diverse scenarios. Given the critical role of 3D spatial information in complex manipulation tasks, several works [116, 35, 94, 33] explore the encoding of point cloud data or multi-view images for 3D imitation learning. However, most existing methods are trained on simulation datasets or self-collected real-world datasets, and the robotics community still lacks a unified large-scale dataset.

Robotic Learning Datasets. Interacting with spatial configurations in real-world environments is vital for robots. However, collecting data with a real robotic arm often incurs substantial costs [77, 50]. General-purpose simulators [19, 52, 67, 76] replicate the physical world and provide virtual environments for training policy models, significantly reducing the costs and time associated with data collection. To meet the training demands of complex and long-horizon tasks, simulators based on real-world environments are developed [53, 13, 109, 90], featuring photorealistic 3D assets and scenes built with game engines. However, the sim-to-real gap significantly impacts the manipulation accuracy of imitation learning policies. As a result, some research shifts towards directly collecting real-world data, including datasets gathered through automated scripts or expert agents [83, 38, 54, 10, 22, 47], as well as those obtained via human teleoperation [70, 93, 26, 9, 45, 99, 8, 29]. As shown in Table I, we compare RoboMIND with representative publicly available real-world datasets for robot manipulation. RoboSet [8] and BridgeData V2 [99] include over 50k trajectories, but are limited to 6 and 13 skill types, respectively. In contrast, RH20T [29] covers 33 tasks, while its data scale is relatively small compared to the others. Recently, Open X-Embodiment [77] has made a large effort to unify existing robot datasets into a standardized format, incorporating data from diverse robots collected through collaboration among 21 institutions. Following this, ARIO [102] further integrates real-world and simulated data into a standard format, aiming to bridge the gaps in existing data resources. DROID [50] collects 76k demonstration trajectories via human teleoperation. Although previous large-scale datasets offer diverse scenarios, most focus on a single embodiment type—the two-finger gripper—and lack dexterous hands, limiting task variety. In contrast, our proposed RoboMIND features four distinct embodiments, including both grippers and dexterous hands, and expands the number of task types to 479 with long-horizon dual-arm tasks for complex skill training. Most importantly, RoboMIND is collected in a standardized setting, ensuring consistency and minimizing variability.

Large-scale Policy Learning. Learning robotic policies from large and diverse datasets has become a major research focus in the field of robotics. One series of works leverages egocentric human videos [36, 20, 21, 37] to assist in robot action learning. Leveraging large-scale human videos, previous works investigate learning robotic representations [75, 5], manipulation priors [69, 48], and dexterous hand control [68, 107]. Another prominent approach, VLA models, leverages multimodal instruction datasets [71, 62, 42] and robot data [9, 72, 91, 103] for co-training or pretraining, enhancing the model's reasoning and generalization abilities. Specifically, RT-2 [117] innovatively incorporates large-scale internet data and low-level action data for co-finetuning; RoboFlamingo [58] directly loads the pretrained parameters from OpenFlamingo [4] for visual instruction tuning; Robo-Mamba [64] utilizes high-level common sense and robotics-related reasoning data for co-training. Finally, a series of works [65, 51, 56] leverage large assembler datasets, such

as Open X-Embodiment and ARIO, for pre-training. The large-scale pre-training significantly enhances the fine-tuning efficiency and generalization capability of policy models. Our proposed real-world dataset and digital twin simulator provide a large-scale pretraining dataset and a high-quality fine-tuning dataset for policy learning in real-world applications, whose efficacy is demonstrated via abundant experiments.

III. DATASET COLLECTION AND PROCESSING

In this work, we primarily introduce how the RoboMIND dataset is collected on the robots and detail the process of cleaning the RoboMIND dataset. Our dataset is collected from four different robotic embodiments (Franka Emika Panda [34], Tien Kung [12], AgileX Cobot Magic V2.0 [87], and UR5e [88]), totaling **107k** trajectories on **479** tasks, **96** different object classes, and **38** operational skills. To support the development of such a large-scale dataset, we develop an intelligent data platform designed to collect, filter, and process the dataset efficiently. This platform uses a cloud-native architecture and distributed computing to handle large-scale data processing, offering five main functionalities and their corresponding modules:

1) **Data Collection:** Collect data from four types of robots using teleoperation equipments and then automatically transmit the collected data to the data platform;
2) **Data Storage:** Package and store the collected dataset in a standardized H5 format, including both visual data of the robot's executed actions and robotic proprioceptive data of its movements;
3) **Data Preprocessing:** Filter the dataset based on predefined standards, evaluating task execution accuracy, motion trajectory smoothness, and the presence of occlusion or motion blur in the visual data;
4) **Data Classification:** Categorize the collected dataset by robot type and specific tasks performed;
5) **Data Annotation:** Perform detailed linguistic annotations on the collected dataset.

A. Data Collection and Storage

Teleoperation is widely applied in the data collection processes for various types of robots [85, 114, 108, 40, 63, 106, 84, 16, 100]. Different types of robots also have specific teleoperation devices for data collection. For example, researchers typically use VR headsets and motion capture suits to collect humanoid robot motion data. They capture the state of human movements and map this motion onto the humanoid robot platform, enabling the robot to replicate these movements while simultaneously collecting a comprehensive dataset [16, 100]. RoboMIND contains teleoperation data in real-world and simulation environments from various types of robots, such as single-arm robots (Franka Emika Panda [34], UR5e [88]), dual-arm robots (AgileX Cobot Magic V2.0 [87]), and humanoid robots (X-Humanoid Tien Kung [12]).

For the single-arm robots, following the Gello [106], we construct the 3D-printed components and the servo motors that match the Degrees of Freedom (DoF) of the robotic arm (see
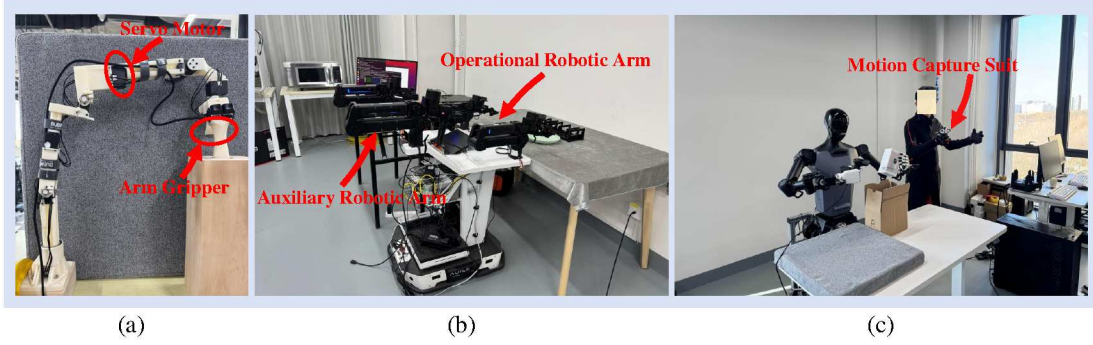
Fig. 2: Visualization of teleoperation methods for different robots. (a) Using 3D-printed components to control the single-arm robots. (b) Regulating the main robotic arm from the auxiliary arm for dual-arm operation. (c) Adopting a motion capture suit to map onto the humanoid robot for operation.

Figure 2(a)). The motion of these 3D-printed components is mapped to the robotic arm's movements, thereby driving the arm. Additionally, we use depth cameras to record the RGB-D information of the robotic arm movement and simultaneously receive the robot state of the robotic arm.

For the dual-arm robots, we directly utilize a bilateral tele-operation device similar to the Mobile ALOHA system [31] on the robot to collect the dataset. Figure 2(b) shows that we employ a teleoperation structure using an auxiliary robotic arm to control the main robotic arm.

For the humanoid robots, Figure 3 illustrates the structural design of the Tien Kung humanoid robot utilized in Robo-MIND. In terms of configuration, it is highly modeled after humans. The robotic arm is flexible and has a strong load-carrying capacity, making it suitable for performing operational tasks to collect datasets. The dexterous hand is integrated with multiple sensors for precise operation. With 42 degrees of freedom throughout the whole body, it can perform a wide variety of movements. In terms of visual perception, depth cameras are installed on its head, chest, waist, and back. The head

Fig. 3: The Tien Kung humanoid robot configuration.

is equipped with the Orbbec Gemini 335 [79], and the other parts are equipped with the Orbbec Gemini 335L [80]. These cameras use active and passive stereo vision technology to provide multiple data streams, accurately record visual perception information. Besides Gello-style teleoperation devices, we use motion capture suits Xsens [74] to collect motion data from various joints of the human body and then map the human joint movements to the corresponding joint movements of a humanoid robot. This allows the humanoid robot to perform the same actions as the human body, enabling remote operation for data collection. Using motion capture suits provides a more accurate and direct method for capturing human movement, compared to relying on VR headsets [16] and cameras [30] for human pose recognition. Figure 2(c) visualizes how we use a motion capture suit to collect data for humanoid robot operation.
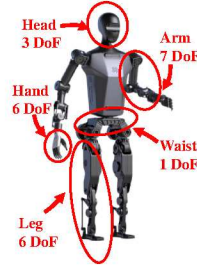
TABLE II: Examples of the task definitions for Franka, AgileX, and Tien Kung robots.

Task Name	Task Description
FR-PlaceBreadPlate	The Franka single-arm robot grasps a piece of bread and places it on a plate.
AX-PackBowl	The AgileX robot packs the bowls.
HR-OpenDrawer LowerCabinet	The Tien Kung robot opens the bottom drawer of the cabinet.

To optimize storage efficiency and facilitate dataset organization, we consolidate each collected trajectory, encompassing multi-view RGB-D data, robot proprioceptive state information, specific end-effector state information, and teleoperation body state information, into a single H5 format file.

B. Data Preprocessing and Classification

All data is collected from operators controlling the teleoperation system in real-time, and errors can arise due to physical limitations such as fatigue, habits, distractions, or external disruptions. To mitigate these issues, we employ a rotation rest system for operators and strive to provide a comfortable working environment to help them stay focused. Additionally, we perform comprehensive quality checks on collected data to ensure its reliability. We define quality assurance criteria, such as avoiding unnecessary contacts and repeated grabbing (see Figure 4). The quality assurance consists of three steps:

- **Initial Inspection:** Quickly review videos to ensure there is no obvious technical issue, such as frame loss or freezing.
- **Detailed Inspection:** Review the video frame-by-frame or in slow motion to carefully check if the conditions described in Figure 4 are present.
- **Data Filtering and Issue Logging:** Document specific timestamps and descriptions for non-compliant data and categorize it for further processing or improvement.

For data classification, we adopt a task-centric data collection protocol, where each task serves as the fundamental unit of the dataset. We classify the collected datasets according to the task names, and each task name is comprehensively
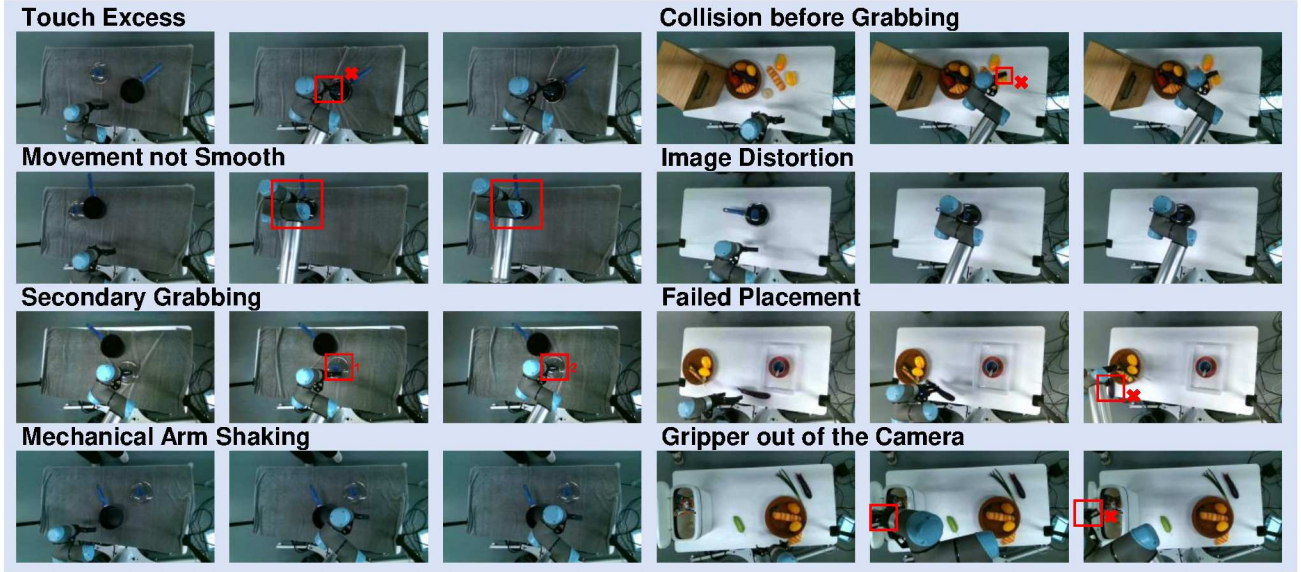
Fig. 4: We define 8 quality assurance criteria in the data collection process. **Touch Excess:** Unnecessary contact with objects by the robotic arm; **Movement not Smooth:** Noticeable jerking or interruptions in robotic arm movements; **Secondary Grabbing:** Repeated grasping attempts after failures in robotic arm operations; **Mechanical Arm Shaking:** Abnormal vibrations in the robotic arm; **Collision before Grabbing:** Collision of the gripper with surrounding objects before grasping; **Image Distortion:** Data collection quality issues; **Failed Placement:** Incorrect placement of objects; **Gripper out of the Camera:** Frames in which the gripper exceeds video frame boundaries. During the data inspection process, all failures were annotated from videos. We show 8 trajectory examples that failed to pass the quality assurance due to different reasons. Each example includes three images that depict the dynamic process of the trajectory. We use red boxes and markers to highlight the reasons for failure.
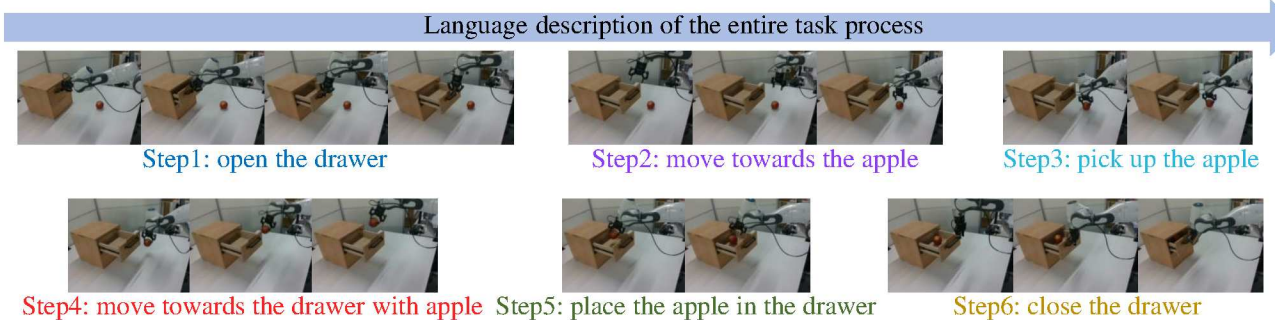

Fig. 5: Example of language description annotation. The video of the robotic arm placing the apple in the drawer is divided into six segments using Gemini. The language descriptions provided for each segment were initially generated by Gemini and subsequently refined through manual revision.

defined by four key components: (1) the specific robotic embodiment utilized; (2) the manipulation skill being executed; (3) the objects involved in the task; and (4) detailed scene descriptions, including object positions, spatial relationships, and environmental constraints or interfering elements. Table II shows examples of the task definition.

This structured task-based framework ensures systematic data collection and enables fine-grained analysis of robotic manipulation capabilities across different scenarios and tasks.

C. Data Annotation

While the visual and robot proprioceptive information can be extracted directly from the collected videos and trajectories, we need to provide better semantic information from the data

to aid model training. For each collection task, its detailed and accurate linguistic descriptions are provided. These linguistic annotations can be utilized for training currently popular VLA models. In addition, RoboMIND collection tasks encompass numerous long horizon tasks, where a uniform linguistic description may be insufficient to capture the full complexity and nuances of the entire task. Thus, we offer detailed fine-grained linguistic annotations for each movement occurring within a trajectory, as illustrated in Figure 5. We annotate 10k successful robot motion trajectories, which are contained in long horizon manipulation tasks. The annotation process involves two primary steps. First, we use Gemini [96] to segment each video based on the sequence of operations and generate detailed text descriptions for each segment. These descriptions

accurately capture the operational steps and relevant context. Second, we manually refine Gemini's annotations regarding the following key aspects:

- Identifying key manipulated objects;
- Detecting and describing all critical actions in the video;
- Ensuring accurate description of operational details;
- Applying reasonable granularity in temporal segmentation;
- Maintaining consistent temporal logic.

This thorough process enhances the precision and reliability of the language annotations for the collected trajectories. We show the annotation of a video of a Franka Emika Panda arm picking the apple and placing it in the drawer using the above standard procedure in Figure 5. The results show that our annotation scheme can accurately segment the key actions in the video and provide precise language descriptions of these key actions. These detailed descriptions can be used for training models like RT-H [7].

IV. DATASET ANALYSIS

Based on a standardized procedure, we collected a large-scale, multi-embodiment dataset named RoboMIND. This dataset consists of 107k high-quality trajectories across 4 robotic embodiments, 479 tasks, 96 object classes, and 38 skills. Robotic data diversity plays a crucial role in model generalization, encompassing various dimensions across hardware and environmental settings. In this section, we perform a thorough quantitative analysis of key diversity dimensions, including robot variety, task length variation, task diversity, and object diversity. We analyze RoboMIND across these dimensions, showing that it offers comprehensive training data to learn generalizable manipulation policies. Furthermore, unlike previous works [77, 50], RoboMIND offers unique data types, such as language descriptions and failure case demonstrations, which enhance the policy model's ability to perform fine-grained task planning and reflect on failure actions.

A. Quantitative Analysis

Heterogeneous Embodiments. A manipulation dataset with different robotic embodiment types improves generalization to various actions and joint DoFs in downstream tasks. We select four mainstream hardware platforms, each paired with different actuators: the single-arm robots, Franka Emika Panda and UR5e with grippers; the dual-arm robot AgileX Cobot Magic V2.0 with grippers; and the humanoid robot Tien Kung equipped with dexterous hands. Figure 1(a) shows the distribution of trajectories across different embodiments in our dataset. Franka accounts for 49.2% of the total trajectories, with over 26,070 simulation-based trajectories from our digital twin environment and 26,866 real-world trajectories collected via human teleoperation. The remaining three embodiments consist solely of real-world demonstrations. Specifically, the dual-arm data enhances the dataset's diversity and complexity, supporting the training of coordination skills and more long-horizon tasks. Additionally, the humanoid robot with dexterous hands, which constitutes 17.8% of the trajectories, can perform
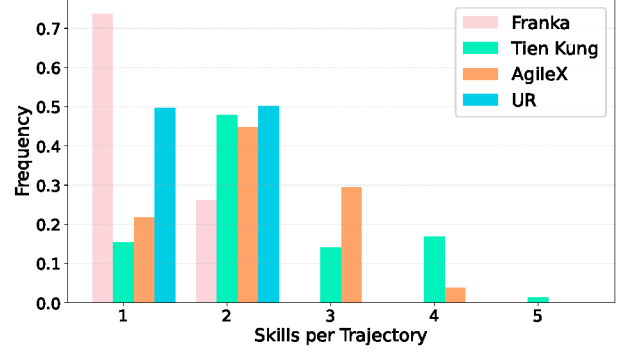
(a) Skill number distribution histogram for each embodiment. We observe that over 70% of the Franka tasks involve only a single skill, while over 75% of the Tien Kung and AgileX tasks involve two or more skills, indicating that these dual-arm tasks are mostly long-horizon tasks.

● Pick   ● Place   ● Hand-over
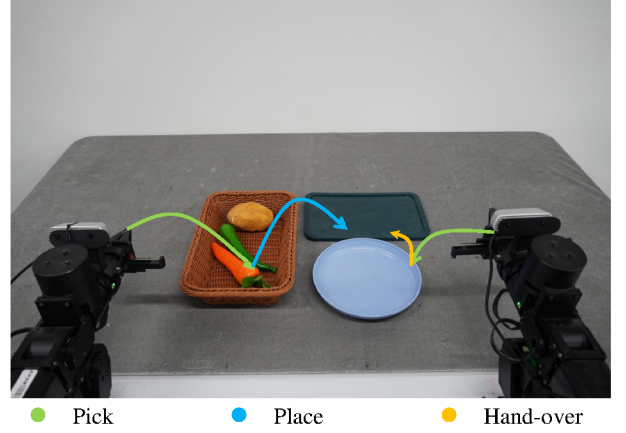
(b) The `AX-PutCarrot` task with the AgileX robot is visualized, involving a sequence of three different skills: pick, hand over, and place.

Fig. 6: Analysis and visualization of skill distribution across different robotic embodiments.

a series of complex, human-like manipulation skills. The heterogeneous set of embodiment data collected under a unified standard can provide pretraining data for policy models with different action spaces [65, 51], as well as experimental data for the cross-embodiment transfer research [110, 15].

Tasks with Various Horizon Lengths. In addition to the diversity across robot, the varied task horizons in the dataset directly impact the temporal generalization capabilities of policies in real-world scenarios. We calculate the average task horizon (the number of time steps in one trajectory) for each embodiment, as shown in Figure 1(b). Tasks collected by Franka and UR have shorter trajectories (fewer than 200 time steps), making them ideal for training primitive skills. In contrast, tasks from Tien Kung and AgileX have longer trajectories (over 500 time steps), better suited for long-horizon task training and skill composition. Since each task involves a varying number of skills, we computed the skill number distribution for each embodiment in Figure 6(a), offering a clearer view of task horizons. AgileX tasks typically involve two or more combined skills, while Tien Kung tasks
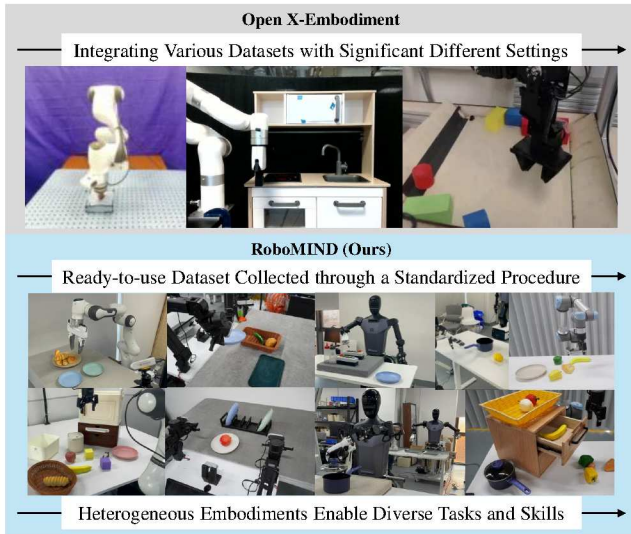
Fig. 7: Distribution of objects in RoboMIND, categorized as domestic, industrial, kitchen, office, and retail. The y-axis uses a logarithmic scale for counts above 500, with exact numbers shown for values exceeding it.

vary in length, with some incorporating up to five skills per task. To provide a clearer explanation of long-horizon task construction, we show an AgileX task involving three skills and visualize its dual-arm trajectory in Figure 6(b). First, the left and right arms perform the pick skill on the carrot and blue plate, respectively. Next, the left arm hands the carrot to the right arm's plate. Finally, the right arm places the blue plate onto the black plate. The entire process involves complex coordination and long-horizon manipulation.

Task Classification. Unlike the previous dataset categorizing tasks based on de-duplicated verbs [50], we categorize tasks by summarizing the manipulation skills from task language descriptions, considering various axes such as actions, objects, and trajectory horizons. Each trajectory may belong to multiple task types, with only the primary type being counted for each trajectory. As shown in Figure 1(c), tasks are categorized into six types:

1) **Articulated Manipulations (Artic. M.):** Opening, closing, and turning on or off objects with articulated joints;
2) **Coordination Manipulations (Coord. M.):** Dual-arm coordination between the robot's arms;
3) **Basic Manipulations (Basic M.):** Fundamental skills like grasping, holding, lifting, and placing;
4) **Multiple Object Interactions (Obj. Int.):** Interaction with multiple objects, e.g., pushing one cube across another;
5) **Precision Manipulations (Precision M.):** Complex manipulation and control skills, such as pouring liquid into a cup or inserting a battery;
6) **Scene Understanding (Scene U.):** Actions with major challenges related to the semantic understanding of the scene, like closing the upper drawer from the right side or placing four large blocks of different colors into corresponding colored boxes.

By breaking down the language descriptions into fine-grained tasks based on verb-noun combinations, RoboMIND includes 479 distinct tasks. In summary, RoboMIND encompasses a range of skills (i.e., verbs from descriptions) beyond basic manipulations, significantly enhancing the policy

model's manipulation robustness in handling complex and long-horizon tasks.

Diverse Objects. A generalized policy needs to learn not only a variety of task skills but also how to execute each skill consistently when interacting with different objects. Robo-MIND includes over 96 object categories from five usage scenarios, as shown in Figure 1(d), covering most daily life settings: domestic, industrial, kitchen, office, and retail. To provide a detailed overview, we summarize trajectories for all objects categorized by usage scenario in Figure 7. In each scene, we design multiple tasks involving a variety of objects. Specifically, in the kitchen, the dataset includes common food items such as strawberries, eggs, bananas, and pears, along with articulated objects like oven doors and bread machines; Domestic scenarios feature both rigid objects like tennis balls and deformable objects like toys; Office and industrial scenarios include small objects that require precise control, such as batteries and gears. This wide variety of objects increases the dataset's complexity and supports better generalization to unseen objects in downstream tasks.

B. Qualitative Analysis

Standardized Settings. RoboMIND features standardized settings to form a large-scale real-world manipulation dataset. As shown in Figure 8, we compare our dataset with Open X-Embodiment, another large-scale robotic learning dataset. Although Open X-Embodiment contains a vast amount of data, the significantly different settings make it difficult to learn efficient manipulation policies across the entire dataset. In contrast, RoboMIND is collected through a carefully designed standardized procedure, making it ready-to-use for other roboticists. Meanwhile, its heterogeneous embodiments, diverse tasks, and various skills are suitable for training generalizable policies, whether for primitive skills or long-horizon manipulations.

Failure Case Demonstrations. We also release 5k trajectories of the robot task failure cases. The failure cases documented include scenarios where different types of humane operators failed to complete their assigned tasks, as well as in-

Fig. 8: Comparison between Open X-Embodiment and Robo-MIND. RoboMIND features heterogeneous embodiments with diverse tasks and skills while providing ease of use due to standardized settings.
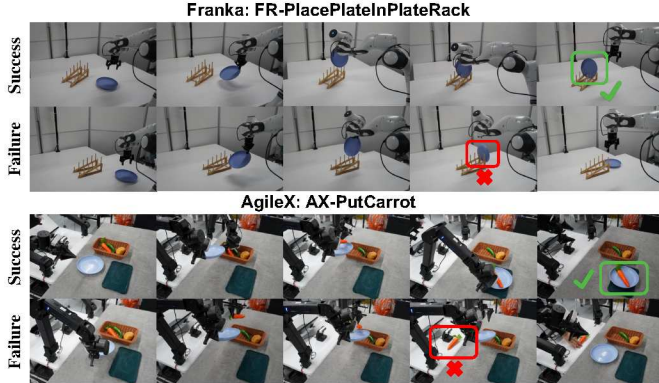

Fig. 9: Visualization of failed data collection cases. We present two examples of failure from Franka and AgileX. In the FR-PlacePlateInPlateRack task (the second row), the Franka arm fails to align with the slot, causing the plate to slip due to operator interference. In the AX-PutCarrot task (the fourth row), the AgileX gripper unexpectedly opens, dropping the carrot. These failure cases were filtered out during quality inspection to maintain the dataset quality.

stances where robots encountered failures during the execution of operational tasks. We present the visualization examples from the Franka and AgileX robots of these failure cases in Figure 9. For the FR-PlacePlateInPlateRack task performed by Franka, a successful execution shows the robotic arm accurately placing a plate into the plate rack. In the failure case, the arm fails to locate the correct slot position, causing the plate to slip out of the rack, likely due to visual occlusion or interference from the operator. For the AX-PutCarrot task performed by AgileX, successful execution demonstrates the robot's collaborative manipulation to place a carrot onto the plate. In the failure case, the robot's gripper unexpectedly opens, causing the carrot to drop prematurely and resulting in
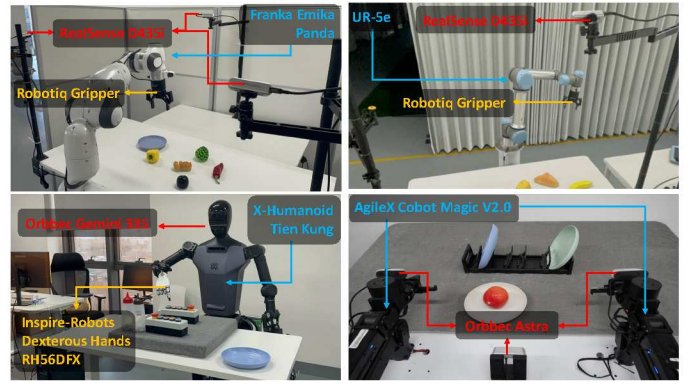
Fig. 10: Robotic real-world setup. For the Franka robot, we use cameras positioned at the top, left, and right viewpoints to record the visual information of the task trajectories. For the Tien Kung and AgileX robots, we use their built-in cameras to record visual information. For the UR robot, we use an external top camera.

task failure-presumably due to accidental gripper activation by the operator. During the data quality inspection process, these failed trajectories are identified, categorized, and documented, further enhancing the overall quality of the dataset.

The failure data is intended to advance research in areas like failure detection and recovery, data augmentation, and reward generation for reinforcement learning. For instance, training a binary classifier on success/failure data can aid failure detection, SSDF [104] filters high-quality data segments from failure data, and the RLHF approach [6] uses failure data as negative examples to learn accurate patterns. Our failure data can be used seamlessly in these works.

V. ANALYZING ROBOT LEARNING WITH ROBOMIND

Following the detailed description of RoboMIND's collection process and an in-depth analysis of its characteristics, we conducted a series of comprehensive experiments employing various robot manipulation learning methods. RoboMIND serves as a benchmark to evaluate the performance and limitations of these methods. In the subsequent experiments, we assessed the performance of single-task imitation learning models (ACT [116], Diffusion Policy [17], and BAKU [39]), as well as VLA large models (RDT-1B [65], OpenVLA [51], and CrossFormer [24]), which can perform multiple tasks with RoboMIND. Subsequently, we validated the ability of the VLA models to generalize across various scenarios and manipulate different types of objects. Additionally, we applied RoboMIND to pre-train the aforementioned VLA large models, demonstrating that RoboMIND also facilitates cross-embodiment task execution for the VLA large models. Finally, we provided some failure case analyses and validated the effectiveness of our digital twin simulation data via co-training.

A. Experiment Setup

Real-world Robotic Setup. Our real-world robotic setup is shown in Figure 10. The robotic platforms used in this study are equipped as follows: **(1) Franka Emika Panda** [34]
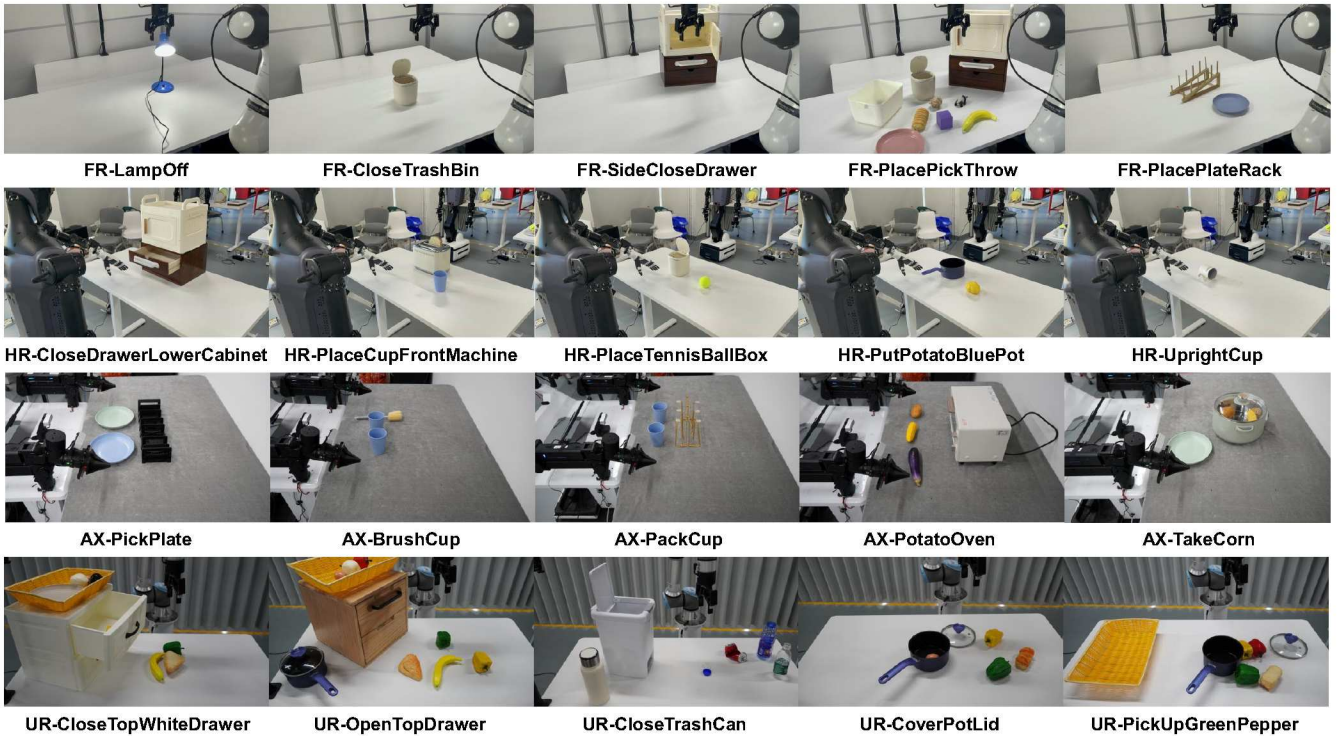
Fig. 11: Diverse task examples across 4 robotic embodiments in RoboMIND. The dataset features tasks performed by four distinct robotic embodiments: Franka (the first row), Tien Kung (the second row), AgileX (the third row), and UR (the fourth row). For each robotic embodiment, we have selected 5 representative task scenarios.

features three Intel RealSense D435i cameras [44] (left, top, and right) with resolutions of 480 × 640, 720 × 1280, and 480 × 640 pixels, respectively, and a Robotiq gripper. **(2) Tien Kung** [12] utilizes two Inspire-Robots RH56DFX dexterous hands and Orbbec Gemini 335 cameras [79] on the head and chest, both at 480 × 640 resolution. **(3) AgileX Cobot Magic V2.0** [87] is fitted with two hand-eye Orbbec Astra cameras [78] and one front-facing camera, all at 480 × 640 resolution. **(4) UR5e** [88] is paired with a top-mounted Intel RealSense D435i camera at 480 × 640 resolution and employs a Robotiq gripper.

Representative Tasks. RoboMIND encompasses a diverse collection of 479 distinct manipulation tasks collected across four different robot embodiments. Representative examples of these tasks are illustrated in Figure 11. Below, we provide a representative task for each robot to elucidate the nomenclature and functionality associated with these operations.

- `FR-SideCloseDrawer`. This task requires the Franka robotic arm to locate the outer edge of a cabinet door accurately. The robot needs to make contact with the door edge and push it along a curved path. The goal is to completely close the cabinet door.
- `HR-UprightCup`. In this task, the Tien Kung humanoid robot needs to grasp a cup that is lying on its side. The robot must then execute a 90-degree rotation movement to bring the cup to an upright position. Finally, it needs to place the upright cup on the table surface gently.
- `AX-TakeCorn`. For this task, the AgileX robot must first

use its left hand to locate and open the pot lid. The robot then extends its right hand into the pot to grip the corn. Finally, it needs to carefully lift the corn out of the pot and place it onto a plate.

- `UR-CloseTopWhiteDrawer`. This task is performed by the UR5e robot, wherein the robot is required to close the uppermost drawer of a set of stacked white drawers.

B. Single-task Imitation Learning Models

Experimental Task Design. We carried out our single-task experiments on a large set of single tasks. We used a total of 45 tasks which were grouped based on the robots performing them. Franka, Tien Kung, AgileX, and UR5e carried out 15, 10, 15, and 5 tasks respectively. We carefully chose these tasks to include a wide variety of actions collected in RoboMIND. These actions ranged from simpler tasks like picking up different objects and placing them in specified spots, to more complex tasks like pulling and pushing articulated objects. Additional tasks involved dual-arm coordination and precise operations, posing further challenges to the learning capabilities of the models.

Training and Evaluation Setup. In terms of the imitation learning algorithms, we used three well-known and commonly used methods: ACT [116], Diffusion Policy [17], and BAKU [39]. For ACT and BAKU, we followed the default model settings as recommended in their original papers. For Diffusion Policy, we followed the implementation in DROID [50]. Using the three algorithms, we trained the single-task model from scratch for each dataset. After training, we
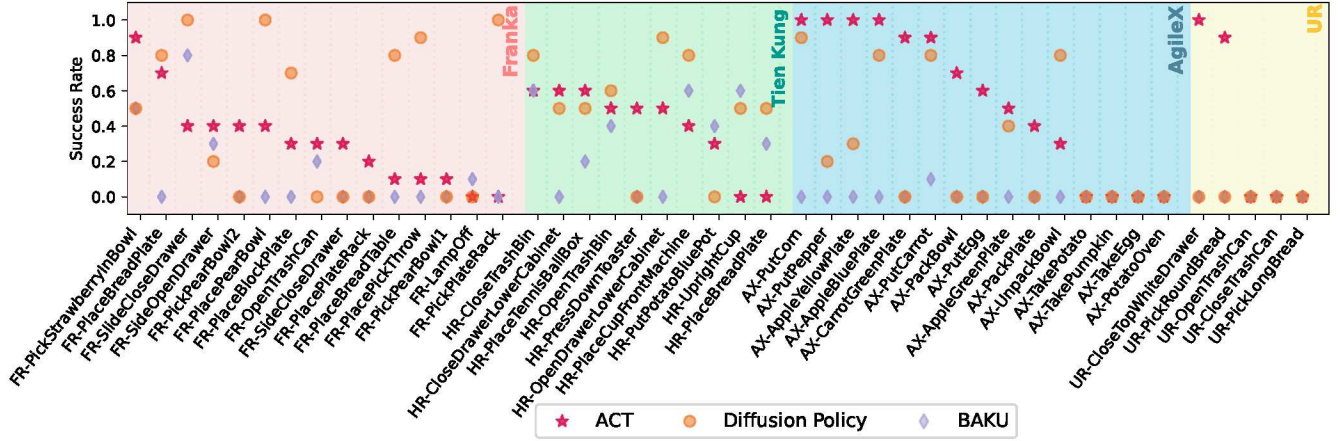
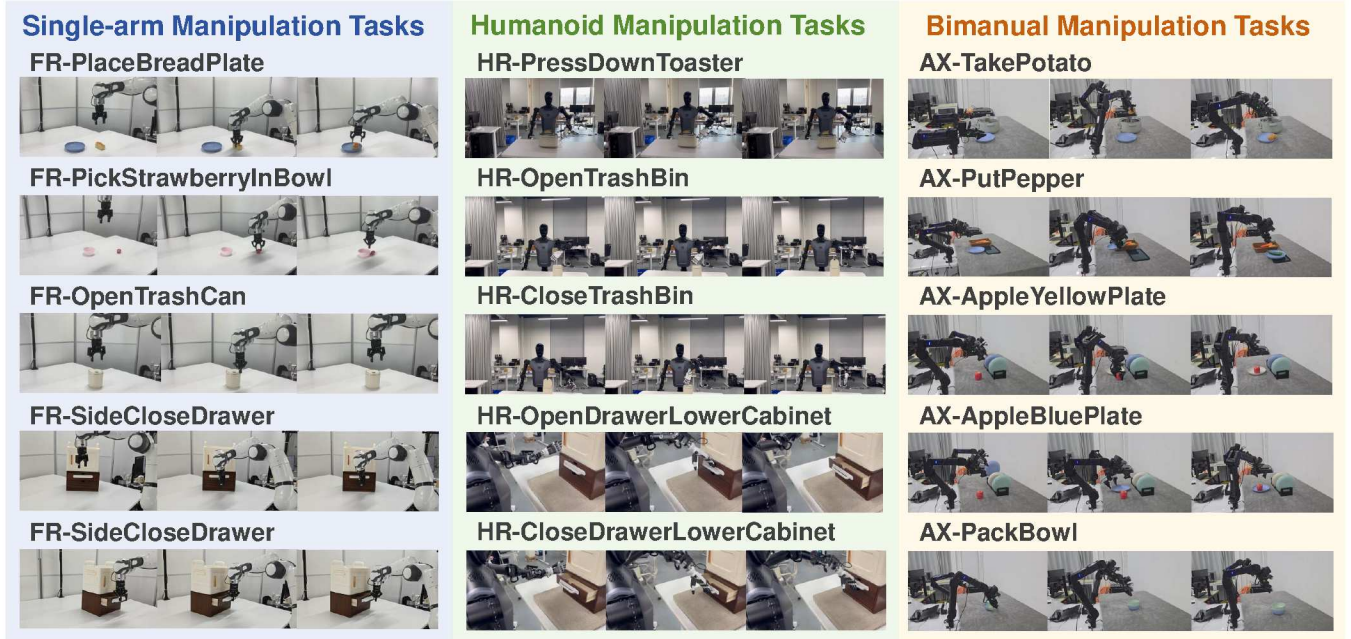Fig. 12: Success rates of ACT, Diffusion Policy, and BAKU on RoboMIND.

Fig. 13: Visualization of the selected tasks on single-arm, dual-arm, and humanoid robots used in experiments of the vision-language-action models.

directly deployed the models in real-world environments for evaluation. We assessed the performance of each model using its success rate in the tasks. Each model was tested ten times, and the testers recorded the success or failure of each test and the reasons if there were any failures. This thorough process gave us valuable insights for further developments.

Experimental Results. Figure 12 presents the performance of ACT [116], Diffusion Policy [17], and BAKU [39] across 45 tasks using four types of robots, evaluated in terms of the success rate. In Figure 12, we found that ACT achieves an average success rate of 55.3% across 15 tasks on AgileX, outperforming Franka (30.7%), UR5e (38.0%), and Tien Kung (34.0%). Additionally, ACT also showed promising results on several humanoid robot tasks, including a 60% success rate on `HR-CloseDrawerLowerCabinet`. These results not only illustrated that ACT shows robust performance in complex

dexterous hand manipulation tasks but also underscored the high quality of data gathered in RoboMIND. Similarly, Diffusion Policy also demonstrated its capacity to learn complex tasks, outperforming ACT in several tasks on Franka and Tien Kung. Therefore, we believe that the single-arm, dual-arm, and dexterous hand datasets in RoboMIND can serve as high-quality training sets to improve the performance of single-task imitation learning, thereby advancing the development of the entire imitation learning field. On the other hand, BAKU exhibits lower success rates across most tasks. This discrepancy could be attributed to the hyper-parameter settings from the original BAKU paper, which is primarily optimized for simulation environments rather than real-world robotic platforms tested in our experiments. The significant performance gap underscores the challenges in directly transferring models from simulated settings to physical robots.

C. Vision-Language-Action Large Models

Experimental Task Design. This section seeks to examine the performance of VLA large-parameter robot model when applied to RoboMIND. We picked fifteen tasks performed by different types of robots from the single-task imitation learning experiments. Figure 13 illustrates the tasks we chose for Franka single-arm robot, the Tien Kung humanoid robot, and the AgileX dual-arm robot. For **the Franka single-arm robot**, these selected tasks encompass common robotic arm operations, such as picking and placing, pushing and pulling, along with more nuanced tasks that require precise manipulation, including picking objects of varying sizes and accurately positioning the robotic arm to open a trash bin lid. For **the Tien Kung humanoid robot**, the tasks are divided into two main categories. The first category consists of tasks similar to those performed by the single-arm Franka robot, which are intended to evaluate the model's performance across different robot types. The second category involves using the humanoid robot's dexterous hands to perform precise operations, such as flipping a toaster switch to toast bread, to assess the model's accuracy in positioning and manipulation. For **the AgileX dual-arm robot**, we chose dual-arm tasks that involve coordinated actions, such as the left arm retrieving a plate from a rack and the right arm placing an apple on the plate. This selection emphasizes the unique capabilities and coordination required in dual-arm operations.

Training and Evaluation Setup. We evaluated the performance of three models (OpenVLA [51], RDT-1B [65], and CrossFormer [24]) fine-tuned by the demonstrations from RoboMIND in completing various real-world tasks. Given that the VLA large model exhibits excellent generalization performance, we employed an aggregated dataset sourced from multitask demonstrations for fine-tuning the VLA models. Specifically, we took the official pre-trained VLA models and fine-tuned them on the multitask datasets for each type of robot, and evaluated their performance on each individual task to determine the extent of generalization achieved, by conducting ten trials for each task. We tested ten trials for each experiment. For OpenVLA [51], which involves fine-tuning the Llama 2 model [98] using a large robotic dataset and adapting it to be a 7-DoF VLA model, we only tested it on the Franka single-arm robot, since the output of OpenVLA is the condition of one end effector and only supports single-arm manipulations.

Experimental Results. Table III presents the success rates for various robot tasks performed using the three different VLA models. The experimental results show that the VLA large models fine-tuned on expert demonstrations from RoboMIND performed well across various different robot tasks. The fine-tuned RDT-1B, compared to CrossFormer and OpenVLA, demonstrated significantly enhanced performance in executing tasks across a range of robot models. This improvement is especially notable for dual-arm manipulation tasks, where RDT-1B excelled. Although the performance of OpenVLA being inferior to that of RDT-1B, it nonetheless achieved a

TABLE III: Success rates of the VLA models in the fine-tuning settings using RoboMIND. **Color boxes** represent the first best performance in all tables of this paper.

Single-arm Manipulation Task	OpenVLA [51]	RDT-1B [65]	CrossFormer [24]
FR-PlaceBreadPlate	4/10	7/10	0/10
FR-PickStrawberryInBowl	0/10	4/10	0/10
FR-OpenTrashCan	3/10	3/10	0/10
FR-SideCloseDrawer	7/10	6/10	2/10
FR-SideOpenDrawer	0/10	2/10	4/10

Humanoid Manipulation Tasks	OpenVLA [51]	RDT-1B [65]	CrossFormer [24]
HR-OpenDrawerLowerCabinet	–	5/10	5/10
HR-CloseDrawerLowerCabinet	–	5/10	3/10
HR-OpenTrashBin	–	2/10	0/10
HR-CloseTrashBin	–	3/10	3/10
HR-PressDownToaster	–	3/10	7/10

Bimanual Manipulation Task	OpenVLA [51]	RDT-1B [65]	CrossFormer [24]
AX-TakePotato	–	6/10	0/10
AX-PutPepper	–	9/10	0/10
AX-AppleYellowPlate	–	10/10	0/10
AX-AppleBluePlate	–	6/10	0/10
AX-PackBowl	–	8/10	0/10

comparable task success rate for straightforward tasks like `FR-PlaceBreadPlate` and `FR-SlideCloseDrawer`. CrossFormer, after being fine-tuned with RoboMIND, demonstrated performance improvements in tasks executed by single-arm and humanoid robots, compared with no success in all tasks without fine-tuning.

D. Leveraging RoboMIND to Enhance VLA Large Models

Training and Evaluation Setup. Currently, most VLA large models are trained with datasets from robots with arms and grippers and can only be applied to the same types of robots. It is noting that RoboMIND contains valuable data from diverse robots including the Tien Kung humanoid robots with dexterous hands, and we applied this dataset in the pre-training of the RDT-1B and CrossFormer models to enhance their ability to handle real-world dexterous manipulation tasks. After that, we fine-tuned the VLA large models using the expert multitask datasets, which is a small subset (about 1%) of RoboMIND. We conducted ten tests for each model on each task, and we compared the results with those from the official pre-trained and expert data fine-tined models.

Experimental Results. Table IV presents the experimental results of RDT-1B and CrossFormer that were first trained on the entire RoboMIND dataset and then fine-tuned on the expert multitask dataset, compared to those fine-tuned directly on the expert multitask dataset. The results show that training different VLA models using the full RoboMIND dataset led to significant improvements in task success rate across a variety of robot tasks. Especially for dual-arm tasks on the CrossFormer, training with the full RoboMIND dataset significantly enhanced its performance. The training effect improved from being unable to complete each dual-arm task to achieving nearly every test success on `AX-TakePotato`, `AX-PutPepper`, and `AX-AppleBluePlate`. For `HR-PressDownToaster` from humanoid manipulation tasks, training CrossFormer using RoboMIND also achieved a 100% task success rate. This improvement underscores the robustness and versatility of RoboMIND in facilitating more effective and reliable robotic

TABLE IV: Success rates of the VLA models before and after training with RoboMIND. The notation '(origin)' indicates models fine-tuned directly on the expert multitask dataset without training on RoboMIND, while '(RoboMIND)' denotes models first trained on the entire RoboMIND dataset and subsequently fine-tuned on the expert multitask dataset.

Single-arm Manipulation Task	RDT-1B (origin)	RDT-1B (RoboMIND)	CrossFormer (origin)	CrossFormer (RoboMIND)
FR-PlaceBreadPlate	7/10	9/10	0/10	10/10
FR-PickStrawberryInBowl	4/10	6/10	0/10	8/10
FR-OpenTrashCan	3/10	6/10	0/10	0/10
FR-SideCloseDrawer	6/10	8/10	2/10	8/10
FR-SideOpenDrawer	2/10	5/10	4/10	3/10
Humanoid Manipulation Tasks	RDT-1B (origin)	RDT-1B (RoboMIND)	CrossFormer (origin)	CrossFormer (RoboMIND)
HR-OpenDrawerLowerCabinet	5/10	6/10	5/10	4/10
HR-CloseDrawerLowerCabinet	5/10	7/10	3/10	7/10
HR-OpenTrashBin	2/10	4/10	0/10	4/10
HR-CloseTrashBin	3/10	4/10	3/10	3/10
HR-PressDownToaster	3/10	4/10	7/10	10/10
Bimanual Manipulation Task	RDT-1B (origin)	RDT-1B (RoboMIND)	CrossFormer (origin)	CrossFormer (RoboMIND)
AX-TakePotato	6/10	8/10	0/10	10/10
AX-PutPepper	9/10	10/10	0/10	10/10
AX-AppleYellowPlate	10/10	10/10	0/10	5/10
AX-AppleBluePlate	6/10	9/10	0/10	9/10
AX-PackBowl	8/10	10/10	0/10	4/10

TABLE V: Success rates of RDT-1B pre-trained with and without the humanoid data.

Success Rate	RDT-1B (w/o Humanoid)	RDT-1B (w/ Humanoid)
FR-PlaceBreadPlate	9/10	9/10
FR-PickStrawberryInBowl	5/10	6/10 (1/10 ↑)
FR-OpenTrashCan	5/10	6/10 (1/10 ↑)
FR-SideCloseDrawer	8/10	8/10
FR-SideOpenDrawer	3/10	5/10 (2/10 ↑)

TABLE VI: Generalization results of VLA large models on the FR-PlaceBreadPlate-related tasks.

Generalization of Backgrounds and Objects	OpenVLA	RDT-1B	CrossFormer
FR-PlaceBreadPlate	4/10	9/10	10/10
FR-PlaceCornPlate	1/10	5/10	6/10
FR-PlaceBananaPlate	1/10	6/10	9/10
FR-PlaceApplePlate	0/10	3/10	2/10
FR-PlaceBreadPlate (Unseen Background 1)	0/10	1/10	2/10
FR-PlaceBreadPlate (Unseen Background 2)	0/10	1/10	0/10
FR-PlaceBreadPlate (Unseen Background 3)	0/10	0/10	0/10

manipulations.

Ablation Studies on Humanoid Data. The standardized data across diverse combinations of robotic platforms eases investigations on cross-embodiment generalization and improvements. For example, it is worth exploring whether the humanoid data is helpful for policy learning for other robots, such as the single-arm Franka. We conducted an ablation study by pre-training the RDT-1B model using an incomplete RoboMIND dataset excluding the humanoids data (19k of 107k in RoboMIND) and fine-tuning the model on the expert multitask datasets. Table V shows that the humanoid data enhances the performance of RDT-1B on the unimanual tasks, especially the difficult tasks. Overall, training with the full RoboMIND dataset obtained a 13.3% relative improvement (0.68 v.s. 0.6) of success rates against the incomplete dataset without humanoid data.

E. Generalization of VLA Large Models

Evaluation Setup. We conducted tests to validate the generalization of using RoboMIND to fine-tune the VLA large models, assessing their ability to generalize across real task scenarios with varying backgrounds and different objects of manipulation. Specifically, we evaluated the generalization performance on the FR-PlaceBreadPlate task of Open-VLA [51], RDT-1B [65], and CrossFormer [24] funetuned on the Franka multitask dataset in Section V-C. Both RDT-1B and CrossFormer are trained on the entire RoboMIND dataset and subsequently fine-tuned using the Franka expert multitask dataset. OpenVLA, in contrast, is directly fine-tuned using the Franka expert multitask dataset. As shown in Figure 14, we executed the FR-PlaceBreadPlate task on three tablecloths with different unseen background patterns and replaced the grasped bread object with an apple, a banana, and a corn. We tested ten trials for each experiment.

Experimental Results. As presented in Table VI, both RDT-1B and CrossFormer exhibited good generalizations for manipulating objects, especially for objects like bananas that are similar in shape to the bread-like objects in the training data. However, when it comes to generalizing across unseen backgrounds, RDT-1B, OpenVLA, and CrossFormer performed relatively poorly in the FR-PlaceBreadPlate task.

F. Failure Case Analysis on Real-world Experiments

During the testing phase, we recorded not only whether the model's task execution was successful but also the reasons for any failures. We predefined nine failure categories: (1) Inaccurate Positioning; (2) Cannot Close Gripper; (3) Cannot Approach Object; (4) Early Gripper Release; (5) Object Drop; (6) Cannot Release Gripper; (7) Cannot Return to Home Pose; (8) Collision Recovery Failure; (9) Excessive Speed.

In Figure 15, we showed the distribution of failure reasons for the ACT across 45 single tasks performed on the four robotic embodiments, as described in Section . We presented
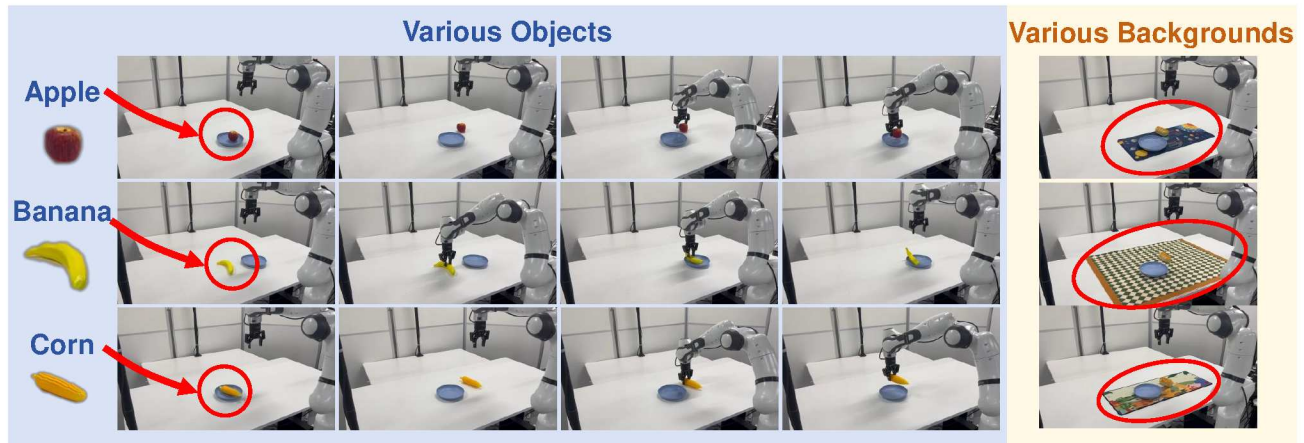
Fig. 14: Unseen objects and backgrounds used to evaluate the generalization ability of the VLA large models.
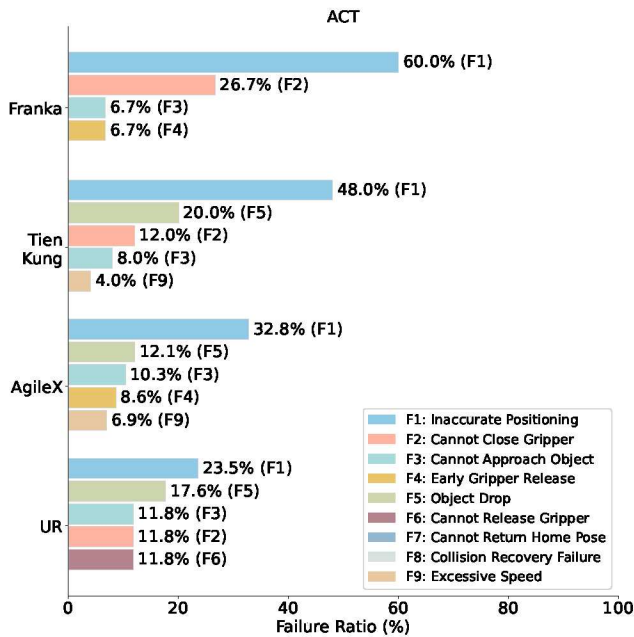

Fig. 15: Top five failure reasons for each embodiment of the ACT algorithm. The x-axis denotes the proportion for each failure among all unsuccessful test cases. The y-axis denotes different embodiments.

the top five most frequent failure reasons for each robotic embodiment. Firstly, we observe that, for ACT, "Inaccurate Positioning" is the most common failure reason across all rollouts. For instance, in the humanoid robot tasks, failures due to "Inaccurate Positioning" accounted for as much as 48%. This highlights the critical importance of accurately positioning the robotic arm in 3D space to execute skills successfully, representing the first step toward achieving task success. It can be noted that the improper gripper actions, such as "Cannot Close Gripper" and "Object Drop", were significant contributors to overall task failures. This issue arises because the number of frames used for gripper actions is typically limited, thereby complicating the learning process.

From a data perspective, which is often overlooked by researchers and developers, the reasons for failure provide insights into improving data quality. The collected data frequently fall short of the task designer's expectations due to various factors such as hardware limitations, physical state, external interference, and communication issues. For instance, inaccurate localization may stem from non-random placement of objects in the dataset, despite instructions for random placement. To address this, we can collect additional data from previously neglected locations to better represent the task environment and improve the success rate. Similarly, gripper non-closure is likely due to the data collector moving too quickly when closing the jaws, resulting in insufficient frames being captured. This makes training more challenging. To mitigate this, we can instruct collectors to slow down during jaw closure to ensure adequate data capture. By refining data collection practices, we can enhance the robustness and reliability of the imitation learning algorithms, ultimately leading to better performance in real-world applications.

G. Real and Simulation Data

To validate the effectiveness of simulation data in Robo-MIND, we conducted two sets of experiments.

Co-training with Real and Simulation Data. Firstly, we combined both real-world and simulation data for training. We selected a complex Franka robotic arm task, `FR-UprightBlueCup`, which requires the robotic arm to rotate nearly 90 degrees, insert its gripper horizontally into the cup's opening, and restore an overturned cup to its upright position. As shown in Figure 16, we constructed a digital twin simulation environment that closely mirrors the real-world setup, including the robotic arm, table surface, objects, and cameras. We collected 100 real-world trajectories and 500 trajectories in the simulation. We then trained and evaluated the ACT model using different ratios of real-world to simulation data, including real-world data only, simulation data only, and mixed ratios of 100:100, 100:200, 100:300, 100:400, and 100:500. Notably, we did not employ any sim2real transfer techniques but instead directly combined both types of data for co-training. Figure 17 shows the success rates of ACT in both
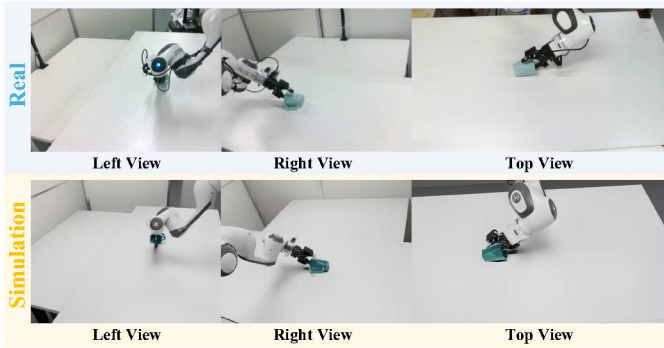
Fig. 16: Experimental setup in real-world and simulation environments. The top and bottom rows show observations from the left view, right view, and top view in the real-world and simulation environments, respectively. We can see that the two environments are very similar, as the simulation environment was constructed to mirror the real environment.
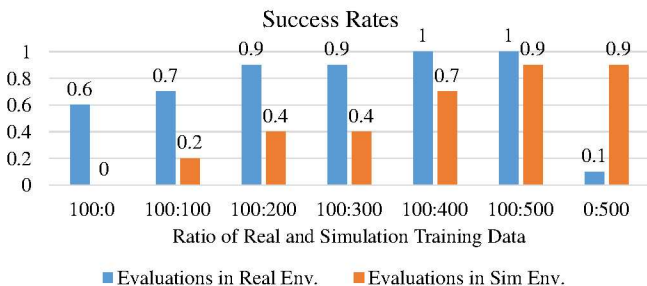

Fig. 17: Success rates of models trained with different ratios of real-world and simulation data.

real-world and simulation environments under different experimental settings. Our observations revealed that increasing the proportion of simulation data improved success rates in both real-world and simulation environments, thanks to our highly accurate simulation environment that closely resembles real-world conditions. However, we also discovered that simulation data alone is insufficient for real-world performance, with real-world data playing a crucial role. For instance, while the combination of 100 real-world trajectories and 500 simulation trajectories achieved a 90% success rate in the simulated environments, using simulation data alone resulted in a dramatic decrease to a 10% success rate in the real world. The primary cause of failure was the cup slipping from the gripper during rotation due to insufficient grip closure. This suggests that significant disparities exist between simulated and real-world physics, particularly for contact-rich tasks, indicating room for improvement in simulation fidelity.

Performance Correlations between Real and Simulation Environments. We further experimented to validate the correlations between performance in the real and simulation environments of the same model We selected five tasks, trained single-task ACT and Diffusion Policy with real-world data, and evaluated them in both the real world and the digital twin simulation environment. Table VII shows positive correlations of the sim-real test results. This aligns with findings in prior research [59] and confirms the positive correlation between

TABLE VII: Success rates of ACT and Diffusion Policy tested in real and simulation environments.

Tests in Different Environments	ACT		Diffusion Policy	
	Sim	Real	Sim	Real
FR-PickStrawberryInBowl	3/10	9/10	0/10	5/10
FR-PlaceBreadPlate	2/10	7/10	3/10	8/10
FR-SlideCloseDrawer	2/10	4/10	7/10	10/10
FR-PlacePearBowl	0/10	4/10	4/10	10/10
FR-PlaceBlockPlate	0/10	3/10	2/10	7/10
Pearson Correlation Coefficient	0.83		0.91	

simulation and real-world results.

VI. DISCUSSION AND FUTURE WORK

In this work, we introduce RoboMIND, a large-scale, multi-embodiment dataset for robot manipulation. RoboMIND includes four distinct embodiments, 107k high-quality demonstrations across 479 tasks, 96 objects, and 38 unique skills, collected through an intelligent data platform with a carefully designed quality assurance process.

We present quantitative analyses of RoboMIND, highlighting its heterogeneous embodiments, diverse episode lengths, broad task coverage, and a wide range of objects drawn from five common scenarios: domestic, industrial, kitchen, office, and retail. We also compare RoboMIND qualitatively with the Open X-Embodiments dataset, considering factors such as uniform settings, multiple viewpoints, and embodiment diversity. These analyses underscore the richness of RoboMIND and its potential to advance research in robot manipulation.

We conduct experiments on several popular imitation learning robot models, assessing their pre-training performance and generalization capabilities on RoboMIND. Our results indicate an urgent need to enhance accurate positioning and precise control in current algorithms, especially for long-horizon tasks. For potential investigations, we suggest that the high-quality, diverse data of RoboMIND is especially suited—but not limited—to fostering cross-embodiment generalization, adapting imitation learning models to downstream tasks, and exploring data augmentation strategies for improved visual- and task-level generalization.

As an ongoing research project, we continue to expand RoboMIND using standardized collection and quality assurance procedures. Therefore, we believe RoboMIND can serve as a ready-to-use dataset and consistently boost progress in embodied AI research.

VII. LIMITATIONS

One limitation of RoboMIND is the relatively simple background environments. While our primary focus has been on constructing a large-scale, high-quality set of robotic trajectories, we plan to investigate in the future whether incorporating more complex backgrounds can enhance the model's manipulation performance, either through data generation or further collection efforts. Additionally, although RoboMIND covers a broad range of robot tasks and environments, it currently lacks data from mobile manipulation scenarios. However, since two of our robotic embodiments are mobile, we plan to

expand RoboMIND to include mobile manipulation tasks in the future. Additionally, RoboMIND can be further enriched by adding more informative annotations such as high-level planning instructions and annotations of trajectory quality.

ACKNOWLEDGMENTS

AUTHOR CONTRIBUTIONS

- **Project Leaders:** Zhengping Che and Xiaozhu Ju
- **Project Coordinators:** Kun Wu, Zhuqin Yang, Chengkai Hou, and Jiaming Liu
- **Data Collection and Processing:** Zhiyuan Xu, Guang Yang, Fei Liao, Zhen Zhao, Guangyu Li, Zhao Jin, Lecheng Wang, Kun Wu, Meng Li, and Pei Ren
- **Dataset Annotation:** Yulin Luo, Zeyu Gao, Zhenyu Wang, and Sixiang Qian
- **Algorithm Development:**
  - ACT: Kun Wu
  - Diffusion Policy: Kun Wu, Jilei Mao, and Xinhua Wang
  - BAKU: Meng Li
  - OpenVLA: Yaoxu Lyu, Xingyu Wang, Chenxuan Li, Chenyang Gu, and Yankai Fu
  - RDT-1B: Di Wu, Jingyang He, Sixiang Chen, and Zeyu Gao
  - CrossFormer: Shichao Fan and Xinhua Wang
- **Paper Writing:** Chengkai Hou, Jiaming Liu, Kun Wu, Meng Li, Yinuo Zhao, Mengzhen Liu, Xinhua Wang, Zhengping Che, and Shanghang Zhang
- **Project Support:** Zhuqin Yang, Kun Wu, Chengkai Hou, Jiaming Liu, Yinuo Zhao, Ning Liu, Xinhua Wang, Shichao Fan, Pei Ren, Qiang Zhang, Mengzhen Liu, and Pengju An
- **Project Advisors:** Jian Tang and Shanghang Zhang

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 287–318. PMLR, 14–18 Dec 2023.

[3] OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

[4] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

[5] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023.

[6] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[7] Suneel Belkhale, Tianli Ding, Ted Xiao, Pierre Sermanet, Quon Vuong, Jonathan Tompson, Yevgen Chebotar, Debidatta Dwibedi, and Dorsa Sadigh. Rt-h: Action hierarchies using language. In *https://arxiv.org/abs/2403.01823*, 2024.

[8] Homanga Bharadhwaj, Jay Vakil, Mohit Sharma, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4788–4795. IEEE, 2024.

[9] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. RT-1: Robotics Transformer for Real-World

Control at Scale. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023.

[10] Serkan Cabi, Sergio Gómez Colmenarejo, Alexander Novikov, Ksenia Konyushkova, Scott Reed, Rae Jeong, Konrad Zolna, Yusuf Aytar, David Budden, Mel Vecerik, et al. Scaling data-driven robotics with reward sketching and batch reinforcement learning. In *Proceedings of Robotics: Science and Systems*, July 2020.

[11] Remi Cadene, Simon Alibert, Alexander Soare, Quentin Gallouedec, Adil Zouitine, and Thomas Wolf. Lerobot: State-of-the-art machine learning for real-world robotics in pytorch. https://github.com/huggingface/lerobot, 2024.

[12] The Beijing Humanoid Robot Innovation Center. X-humanoid tien kung, 2024. URL https://x-humanoid.com/.

[13] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *2017 International Conference on 3D Vision (3DV)*, pages 667–676, 2017.

[14] Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, et al. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

[15] Lawrence Yunliang Chen, Kush Hari, Karthik Dharmarajan, Chenfeng Xu, Quan Vuong, and Ken Goldberg. Mirage: Cross-embodiment zero-shot policy transfer with cross-painting. In *Proceedings of Robotics: Science and Systems*, 2024.

[16] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. In *8th Annual Conference on Robot Learning*, 2024.

[17] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *RSS*, 2023.

[18] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[19] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning, 2016.

[20] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018.

[21] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022.

[22] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. In *Conference on Robot Learning*, pages 885–897. PMLR, 2020.

[23] Mingdi Deng, Zhijun Li, Yu Kang, CL Philip Chen, and Xiaoli Chu. A learning-based hierarchical control scheme for an exoskeleton robot in human–robot cooperative manipulation. *IEEE transactions on cybernetics*, 50(1):112–125, 2018.

[24] Ria Doshi, Homer Rich Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation. In *8th Annual Conference on Robot Learning*, 2024.

[25] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[26] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge Data: Boosting Generalization of Robotic Skills with Cross-Domain Datasets. In *Proceedings of Robotics: Science and Systems*, New York City, NY, USA, June 2022. doi: 10.15607/RSS.2022.XVIII.063.

[27] Ben Eisner, Harry Zhang, and David Held. FlowBot3D: Learning 3D Articulation Flow to Manipulate Articulated Objects. In *Proceedings of Robotics: Science and Systems*, June 2022. doi: 10.15607/RSS.2022.XVIII.018.

[28] Hao-Shu Fang, Chenxi Wang, Hongjie Fang, Minghao Gou, Jirong Liu, Hengxu Yan, Wenhai Liu, Yichen Xie, and Cewu Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 2023.

[29] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 653–660. IEEE, 2024.

[30] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. In *8th Annual Conference on Robot Learning*, 2024.

[31] Zipeng Fu, Tony Z. Zhao, and Chelsea Finn. Mobile ALOHA: Learning bimanual mobile manipulation using

low-cost whole-body teleoperation. In *8th Annual Conference on Robot Learning*, 2024.

[32] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

[33] Theophile Gervet, Zhou Xian, Nikolaos Gkanatsios, and Katerina Fragkiadaki. Act3d: 3d feature field transformers for multi-task robotic manipulation. In *7th Annual Conference on Robot Learning*, 2023.

[34] Franka Robotics GmbH. Franka robotics, 2024. URL https://franka.de/.

[35] Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. Rvt: Robotic view transformer for 3d object manipulation. In *Conference on Robot Learning*, pages 694–710. PMLR, 2023.

[36] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

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

[38] Abhinav Gupta, Adithyavairavan Murali, Dhiraj Prakashchand Gandhi, and Lerrel Pinto. Robot learning in homes: Improving generalization and reducing dataset bias. *Advances in neural information processing systems*, 31, 2018.

[39] Siddhant Haldar, Zhuoran Peng, and Lerrel Pinto. BAKU: An efficient transformer for multi-task policy learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[40] Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan Ratliff, and Dieter Fox. Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9164–9170. IEEE, 2020.

[41] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

[42] https://sharegpt.com/. Sharegpt, 2023. URL https://sharegpt.com/.

[43] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. In *7th Annual Conference on Robot Learning*, 2023.

[44] Intel. Depth camera d435i. https://www.intelrealsense.com/depth-camera-d435i/, 2019.

[45] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pages 991–1002. PMLR, 2022.

[46] Shirin Joshi, Sulabh Kumra, and Ferat Sahin. Robotic grasping using deep reinforcement learning. In *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*, pages 1461–1466. IEEE, 2020.

[47] Dmitry Kalashnikov, Jacob Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. Mt-opt: Continuous multi-task robotic reinforcement learning at scale. *arXiv preprint arXiv:2104.08212*, 2021.

[48] Aditya Kannan, Kenneth Shaw, Shikhar Bahl, Pragna Mannam, and Deepak Pathak. DEFT: Dexterous finetuning for hand policies. In *7th Annual Conference on Robot Learning*, 2023.

[49] Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. In *8th Annual Conference on Robot Learning*, 2024.

[50] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. DROID: A large-scale in-the-wild robot manipulation dataset. In *RSS 2024 Workshop: Data Generation for Robotics*, 2024.

[51] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. OpenVLA: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*, 2024.

[52] Nathan Koenig and Andrew Howard. Design and use paradigms for gazebo, an open-source multi-robot simulator. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 3, pages 2149–2154, 2004.

[53] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv*, 2017.

[54] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International journal of robotics research*, 37(4-5):421–436, 2018.

[55] Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, et al. Datacomp-LM:

In search of the next generation of training sets for language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[56] Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024.

[57] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18061–18070, 2024.

[58] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, Hang Li, and Tao Kong. Vision-language foundation models as effective robot imitators. In *The Twelfth International Conference on Learning Representations*, 2024.

[59] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024.

[60] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE, 2023.

[61] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, Miami, Florida, USA, November 2024.

[62] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[63] Huihan Liu, Soroush Nasiriany, Lance Zhang, Zhiyao Bao, and Yuke Zhu. Robot learning on the job: Human-in-the-loop autonomy and learning during deployment. *The International Journal of Robotics Research*, page 02783649241273901, 2022.

[64] Jiaming Liu, Mengzhen Liu, Zhenyu Wang, Pengju An, Xiaoqi Li, Kaichen Zhou, Senqiao Yang, Renrui Zhang, Yandong Guo, and Shanghang Zhang. Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[65] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. RDT-1b: a diffusion foundation model for bimanual manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[66] Zeyi Liu, Arpit Bahety, and Shuran Song. REFLECT: Summarizing robot experiences for failure explanation and correction. In *7th Annual Conference on Robot Learning*, 2023.

[67] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance GPU based physics simulation for robot learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[68] Priyanka Mandikal and Kristen Grauman. Learning dexterous grasping with object-centric visual affordances. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 6169–6176. IEEE, 2021.

[69] Priyanka Mandikal and Kristen Grauman. Dexvip: Learning dexterous grasping with human hand pose priors from video. In *Conference on Robot Learning*, pages 651–661. PMLR, 2022.

[70] Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In *Conference on Robot Learning*, pages 879–893. PMLR, 2018.

[71] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

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

[73] Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021.

[74] Movella. Xsens. https://www.movella.com/products/xsens, 2025. Accessed: 2025-01-15.

[75] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *6th Annual Conference on Robot Learning*, 2022.

[76] NVIDIA. Nvidia isaac sim: Robotics simulation and synthetic data, 2023. URL https://developer.nvidia.com/isaac-sim.

[77] Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhi-

ram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open x-embodiment: Robotic learning datasets and rt-x models : Open x-embodiment collaboration0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903, 2024.

[78] ORBBEC. Astra series - structured light camera. https://www.orbbec.com/products/structured-light-camera/astra-series/, 2022.

[79] ORBBEC. Gemini 335 - 3d vision for a 3d world. https://www.orbbec.com/products/stereo-vision-camera/gemini-335/, 2024.

[80] ORBBEC. Gemini 335l - 3d vision for a 3d world. https://www.orbbec.com/products/stereo-vision-camera/gemini-335l/, 2024.

[81] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[82] Tim Pearce, Tabish Rashid, Anssi Kanervisto, Dave Bignell, Mingfei Sun, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Ida Momennejad, Katja Hofmann, et al. Imitating human behaviour with diffusion models. *ICLR*, 2023.

[83] Lerrel Pinto and Abhinav Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 3406–3413. IEEE, 2016.

[84] Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system. In *Robotics: Science and Systems*, 2023.

[85] Daniel Rakita, Bilge Mutlu, and Michael Gleicher. A motion retargeting method for effective mimicry-based teleoperation of robot arms. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 361–370, 2017.

[86] Moritz Reuss, Maximilian Li, Xiaogang Jia, and Rudolf Lioutikov. Goal conditioned imitation learning using score-based diffusion policies. In *Robotics: Science and Systems*, 2023.

[87] AgileX Robotics. Agilex cobot magic, 2024. URL https://global.agilex.ai/products/cobot-magic.

[88] Universal Robots. Universal robots ur5e, 2024. URL https://www.universal-robots.com/products/ur5e/.

[89] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[90] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied ai research. In *IEEE/CVF International Conference on Computer Vision*, Oct 2019.

[91] Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 645–652, 2024.

[92] Nur Muhammad Mahi Shafiullah, Anant Rai, Haritheja Etukuru, Yiqian Liu, Ishan Misra, Soumith Chintala, and Lerrel Pinto. On bringing robots home. *arXiv preprint arXiv:2311.16098*, 2023.

[93] Pratyusha Sharma, Lekha Mohan, Lerrel Pinto, and Abhinav Gupta. Multiple interactions made easy (mime): Large scale demonstrations data for imitation. In *Conference on robot learning*, pages 906–915. PMLR, 2018.

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

[95] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ICLR*, 2021.

[96] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[97] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, 2024.

[98] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[99] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.

[100] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. In *2nd Workshop on Dexterous Manipulation: Design, Perception and Control (RSS)*, 2024.

[101] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei

Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 121475–121499, 2024.

[102] Zhiqiang Wang, Hao Zheng, Yunshuang Nie, Wenjun Xu, Qingwei Wang, Hua Ye, Zhe Li, Kaidong Zhang, Xuewen Cheng, Wanxi Dong, et al. All robots in one: A new standard and unified dataset for versatile, general-purpose embodied agents. *arXiv preprint arXiv:2408.10899*, 2024.

[103] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, Yaxin Peng, et al. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters*, 2025.

[104] Kun Wu, Ning Liu, Zhen Zhao, Di Qiu, Jinming Li, Zhengping Che, Zhiyuan Xu, and Jian Tang. Learning from imperfect demonstrations with self-supervision for robotic manipulation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[105] Kun Wu, Yichen Zhu, Jinming Li, Junjie Wen, Ning Liu, Zhiyuan Xu, and Jian Tang. Discrete policy: Learning disentangled action space for multi-task robotic manipulation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[106] Philipp Wu, Yide Shentu, Zhongke Yi, Xingyu Lin, and Pieter Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12156–12163, 2024.

[107] Yueh-Hua Wu, Jiashun Wang, and Xiaolong Wang. Learning generalizable dexterous manipulation from human grasp affordance. In *Conference on Robot Learning*, pages 618–629. PMLR, 2023.

[108] Yuqiang Wu, Pietro Balatti, Marta Lorenzini, Fei Zhao, Wansoo Kim, and Arash Ajoudani. A teleoperation interface for loco-manipulation control of mobile collaborative robotic assistant. *IEEE Robotics and Automation Letters*, 4(4):3593–3600, 2019.

[109] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. Sapien: A simulated part-based interactive environment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2020.

[110] Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela Veloso, and Shuran Song. Xskill: Cross embodiment skill discovery. In *Conference on Robot Learning*, pages 3536–3555. PMLR, 2023.

[111] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. In *International Conference on Learning Representations*, 2022.

[112] Maryam Zare, Parham M Kebria, Abbas Khosravi, and Saeid Nahavandi. A survey of imitation learning: Algorithms, recent developments, and challenges. *IEEE Transactions on Cybernetics*, 2024.

[113] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[114] Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Xi Chen, Ken Goldberg, and Pieter Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5628–5635. IEEE, 2018.

[115] Tianle Zhang, Dongjiang Li, Yihang Li, Zecui Zeng, Lin Zhao, Lei Sun, Yue Chen, Xuelong Wei, Yibing Zhan, Lusong Li, et al. Empowering embodied manipulation: A bimanual-mobile robot manipulation dataset for household tasks. *arXiv preprint arXiv:2405.18860*, 2024.

[116] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. In *Proceedings of Robotics: Science and Systems*, July 2023.

[117] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *7th Annual Conference on Robot Learning*, 2023.