

# RoboTwin 2.0: A Scalable Data Generator and Benchmark with Strong Domain Randomization for Robust Bimanual Robotic Manipulation

Tianxing Chen<sup>2,16\*</sup>, Zanxin Chen<sup>3,5\*</sup>, Baijun Chen<sup>15\*</sup>, Zijian Cai<sup>3,5\*</sup>, Yibin Liu<sup>13\*</sup>, Zixuan Li<sup>5\*</sup>, Qiwei Liang<sup>5</sup>, Xianliang Lin<sup>5</sup>, Yiheng Ge<sup>1</sup>, Zhenyu Gu<sup>7,8</sup>, Weiliang Deng<sup>3,11</sup>, Yubin Guo<sup>7,9</sup>, Tian Nian<sup>3,5</sup>, Xuanbing Xie<sup>12</sup>, Qiangyu Chen<sup>5</sup>, Kailun Su<sup>5</sup>, Tianling Xu<sup>10</sup>, Guodong Liu<sup>6,7</sup>, Mengkang Hu<sup>2</sup>, Huan-ang Gao<sup>6,16</sup>, Kaixuan Wang<sup>2,16</sup>, Zhixuan Liang<sup>2,3†</sup>, Yusen Qin<sup>4,6</sup>, Xiaokang Yang<sup>1</sup>, Ping Luo<sup>2,14✉</sup>, Yao Mu<sup>1,3✉†</sup>

<sup>1</sup> MoE key Lab of Artificial Intelligence, AI Institute, SJTU<sup>‡</sup>, <sup>2</sup> HKU MMLab<sup>‡</sup>, <sup>3</sup> Shanghai AI Lab, <sup>4</sup>D-Robotics, <sup>5</sup>SZU, <sup>6</sup>THU, <sup>7</sup>TeleAI, <sup>8</sup>FDU, <sup>9</sup>USTC, <sup>10</sup>SUSTech,

<sup>11</sup>SYSU, <sup>12</sup>CSU, <sup>13</sup>NEU, <sup>14</sup>HKU-SH ICRC, <sup>15</sup>NJU, <sup>16</sup>Lumina EAI

\* Equal contribution    ✉ Corresponding authors    † Co-project leads

‡ Equally leading organizations

Webpage: <https://robotwin-platform.github.io> Doc: <https://robotwin-platform.github.io/doc/>

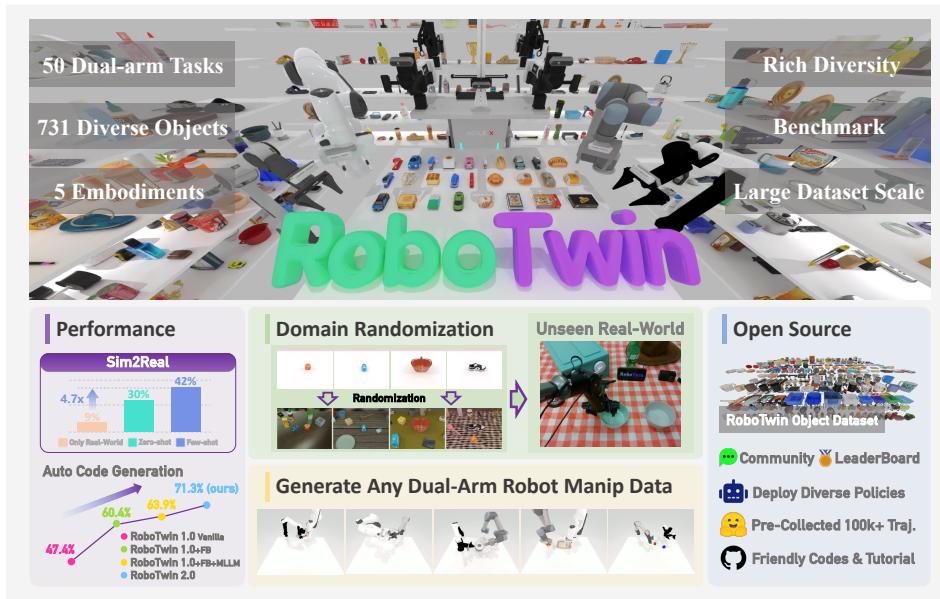


Figure 1: **Overview of RoboTwin 2.0.** RoboTwin 2.0 is a scalable framework for bimanual manipulation, integrating an expert data generation pipeline with a 50-task benchmark built on the RoboTwin Object Dataset (731 objects, 147 categories). A multimodal language agent automates task program synthesis, while flexible dual-arm configurations enable large-scale, diverse data collection. Policies trained on RoboTwin 2.0 exhibit improved robustness and generalization to unseen environments.

## Abstract

Simulation-based data synthesis has emerged as a powerful paradigm for enhancing real-world robotic manipulation. However, existing synthetic datasets remain insufficient for robust bimanual manipulation due to two challenges: (1) the lack of an efficient, scalable data generation method for novel tasks, and (2) oversimplified simulation environments that fail to capture real-world complexity. We present **RoboTwin 2.0**, a scalable simulation framework that enables automated,

large-scale generation of diverse and realistic data, along with unified evaluation protocols for dual-arm manipulation. We first construct RoboTwin-OD, a large-scale object library comprising 731 instances across 147 categories, each annotated with semantic and manipulation-relevant labels. Building on this foundation, we develop an expert data synthesis pipeline that combines multimodal large language models (MLLMs) with simulation-in-the-loop refinement to generate task-level execution code automatically. To improve sim-to-real transfer, RoboTwin 2.0 incorporates structured domain randomization along five axes: clutter, lighting, background, tabletop height and language instructions, thereby enhancing data diversity and policy robustness. We instantiate this framework across 50 dual-arm tasks spanning five robot embodiments. Empirical evaluation shows a 10.9% gain in code generation success rate. Building on this, we evaluate downstream policy learning. With a mix of large-scale synthetic data and only 10 real demonstrations, a vision–language–action (VLA) model achieves a 367% relative improvement over the 10-demo baseline. Even without real data, zero-shot models trained solely on synthetic data obtain a 228% relative gain, highlighting the effectiveness of our dataset in strengthening sim-to-real transfer and robustness to environmental variations. We release the data generator, benchmark, pre-collected dataset, and code to support scalable research in robust bimanual manipulation.

## 1 Introduction

Bimanual robotic manipulation is critical for enabling robots to perform complex real-world tasks such as collaborative assembly, tool use, and object handovers. Developing generalizable bimanual policies—particularly vision–language–action (VLA) foundation models—requires datasets that are simultaneously high-quality, diverse, and large-scale. In the absence of sufficient variability in object geometry, scene clutter, lighting conditions, instruction language, and robot embodiments, learned policies often overfit to narrow distributions and fail to generalize to novel environments or hardware platforms. Yet collecting real-world demonstrations at scale remains prohibitively expensive, time-consuming, and logistically challenging, especially when aiming to cover a broad range of tasks, objects, and embodiments.

Simulation-based data generation provides a scalable alternative for collecting large-scale multimodal datasets and has shown promise in enabling sim-to-real transfer [34, 11]. However, existing pipelines fall short in three critical aspects. First, they lack automated quality control: without an expert-level validation loop, many generated trajectories include execution failures or suboptimal grasps, which degrade policy learning. Second, their domain randomization is often superficial, yielding overly clean and homogeneous scenes that omit essential real-world factors such as clutter, lighting variation, and ambiguous language instructions—elements crucial for robust sim-to-real transfer. Third, they overlook cross-embodiment variation: different bimanual platforms can differ substantially in their kinematic capabilities and grasp strategies. For example, a low-degree-of-freedom (DoF) platform like the Piper often relies on lateral grasps due to its limited dexterity, whereas a high-DoF arm such as the Franka is capable of top-down precision grasps. Yet, current synthetic datasets rarely encode such embodiment-specific affordances or task constraints, limiting their generality.

To address these challenges, we introduce **RoboTwin 2.0**, a scalable simulation-based data generation framework designed to produce high-quality, diverse, realistic, and interaction-rich datasets for bimanual manipulation. RoboTwin 2.0 integrates three key components: (1) an automated expert data generation pipeline that leverages multimodal large language models (MLLMs) and simulation-in-the-loop feedback to iteratively validate and refine task execution code; (2) comprehensive domain randomization over language instructions, object clutter, background textures, lighting conditions, and tabletop configurations, aimed at closing the sim-to-real gap and enhancing policy generalization; and (3) embodiment-aware adaptation, in which object affordances are annotated and robot-specific action candidates are generated to account for heterogeneous dual-arm kinematics.

Building on these components, we introduce three new resources to support scalable research in bimanual manipulation: (1) the RoboTwin-OD asset library, comprising 731 annotated object instances across 147 categories; (2) an automated data generation pipeline with comprehensive domain randomization and a collection of over 100,000 expert trajectories spanning 50 tasks across five dual-arm robot platforms; and (3) a benchmark for evaluating policy generalization to cluttered

environments and open-ended language goals. Together, these resources enable the community to train and evaluate robust bimanual manipulation policies under conditions that closely reflect real-world complexity and diversity.

In summary, our main contributions are as follows: (1) We develop an automated expert data generation framework that integrates multimodal large language models with simulation-in-the-loop feedback to ensure high-quality, expert-level trajectories; (2) We propose a systematic domain randomization strategy that enhances policy robustness by increasing data diversity and sim-to-real generalization; (3) We introduce an embodiment-aware adaptation mechanism that generates robot-specific manipulation candidates based on object affordances; (4) We release the RoboTwin-OD asset library, a large-scale pre-collected multi-embodiment domain-randomized trajectory dataset, a scalable bimanual data generator, and a standardized evaluation benchmark to support scalable training and evaluation of generalizable policies across different robot embodiments, scene configurations, and language instructions.

## 2 Method

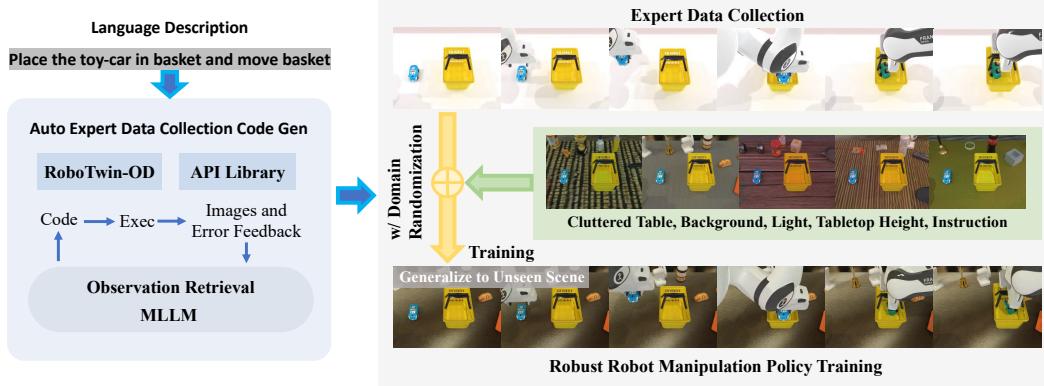


Figure 2: **RoboTwin 2.0 Pipeline.** Built on RoboTwin-OD and a skill API, the framework uses MLLM-based code generation with simulation feedback to produce expert task programs and domain-randomized trajectories for policy training and evaluation.

We illustrate the overall RoboTwin 2.0 pipeline in Fig. 2. The framework begins with a task code generation module that leverages multimodal large language models (MLLMs) and simulation-in-the-loop feedback to automatically synthesize executable task plans from natural language instructions. This module is grounded on a large-scale object asset library (RoboTwin-OD) and a predefined skill library, enabling scalable task instantiation across a broad range of object categories and manipulation scenarios. To ensure high-quality expert demonstrations, we integrate this automated generation pipeline with RoboTwin 2.0’s comprehensive domain randomization scheme, which diversifies observations along language, visual, and spatial axes. This pipeline supports the synthesis of diverse and realistic training data, facilitating the development of manipulation policies that are robust to real-world environmental variability.

### 2.1 Expert Code Generation via MLLMs and Simulation-in-the-Loop Feedback

Recent advances in language models show strong ability to generate intermediate task representations—such as textual plans [18], API calls, or executable code [33, 6, 19]—for complex robotic tasks. Multimodal large language models (MLLMs) further extend this capability by incorporating visual and proprioceptive inputs, enabling more grounded reasoning over real-world sensory data. However, prior systems often depend on strong manual priors or lack closed-loop feedback during program synthesis, which limits their robustness in diverse or dynamic environments.

To address these limitations, we propose an automated expert data generation pipeline that integrates programmatic code synthesis with multimodal execution feedback (Fig.3). The system adopts a closed-loop architecture with two agents: a code-generation agent and a vision–language model (VLM) observer. The code agent synthesizes task programs from instructions, while the observer monitors execution in simulation, detects failures, and suggests corrections. This iterative feedback

loop enables the code agent to refine programs automatically, producing robust, self-improving expert data with minimal human supervision. Unlike prior MLLM-based pipelines such as GenSim2[20] and RoboGen [43], our system supports zero-shot generation of complex dual-arm behaviors beyond primitive pick-and-place actions.

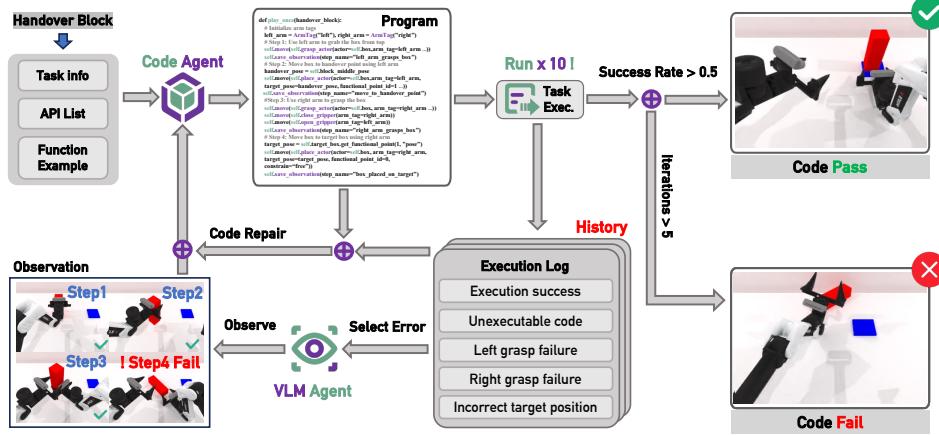


Figure 3: Expert Code Generation Pipeline.

**Input Specification.** Each task is defined by a task name (e.g., *Handover Block*) and a natural language description of the objective. The code-generation agent is conditioned on three key inputs: a general API list, a set of example function calls, and a hierarchical constraint specification. These components collectively guide the synthesis of Python code to execute the task. Additionally, each task may include task-specific function call examples to further ground code generation in context.

**Initial Code Generation.** The code-generation agent synthesizes an initial Python program conditioned on the provided task inputs. It models the program synthesis process as a structured prediction problem over the space of available API calls, leveraging natural language understanding and few-shot prompting from task-specific examples. The generated code specifies a stepwise sequence of robot actions designed to accomplish the target manipulation objective.

**Simulated Execution and Logging.** The generated program is executed ten times per iteration within a simulated robotic environment. Multiple trials are used to account for stochastic variations in simulation dynamics, robot controllers, and sensor noise. After each execution batch, the system generates a structured execution log that records the success or failure of each trial and annotates failure cases with their corresponding causes—such as unexecutable code, left/right grasp failure, or incorrect object placement.

**Multimodal Observation and Error Localization.** In parallel with execution, a vision-language model (VLM) agent observes the robot’s behavior across all ten trials. The VLM performs frame-by-frame inspection to evaluate the success of each program step and localize the point of failure when errors occur. Beyond temporal localization, the VLM also diagnoses failure modes by inferring whether the underlying cause stems from flawed logic, incorrect API usage, or other systemic issues. This diagnostic capability enables the system to address root causes rather than merely responding to superficial execution errors. The detail of VLM observation is shown in G.4.

**Code Repair and Iterative Refinement.** The code-generation agent receives two complementary feedback signals: (i) a quantitative execution log and (ii) a qualitative, localized diagnostic from the VLM. It integrates these inputs to revise the program by modifying or replacing instructions identified as failure-prone. The updated program is then re-evaluated in the next iteration, and the process continues until either the program achieves the setting success rate across ten simulated runs in one iteration or fails to do so after five consecutive refinements. This loop yields expert-level task code with minimal human supervision while avoiding indefinite refinement.

The outcome of this pipeline is a collection of robust, automatically synthesized programs that generate high-quality expert trajectories for downstream training and evaluation. By integrating multimodal reasoning with execution-level feedback, the system produces code that is not only syntactically correct but also semantically aligned with task objectives. This closed-loop generation

framework substantially reduces human supervision while enabling scalable and self-improving expert data creation for complex robotic manipulation tasks.

## 2.2 Domain Randomization for Robust Robotic Manipulation

To enhance policy robustness to real-world variability, we apply domain randomization along five dimensions: (1) cluttered distractor objects, (2) background textures, (3) lighting conditions, (4) tabletop heights, and (5) diverse language instructions. This systematic augmentation broadens the training distribution and markedly improves generalization to unseen scenarios. The effects of these randomizations are visualized in Fig. 4a.



Figure 4: **Visualization of domain randomization and our texture library.**

**Scene Clutter.** To enhance robustness to environmental variation, we augment tabletop scenes with task-irrelevant distractors drawn from RoboTwin-OD (731 objects across 147 categories; see Section 3.1). Each object includes placement annotations, enabling a generic API for semantically valid insertion. We ensure physical plausibility via collision-aware placement and precomputed volumes. To avoid policy confusion, distractors visually or semantically similar to task-relevant objects are excluded during sampling. This yields diverse yet unambiguous cluttered scenes for training.

**Diverse Background Textures.** We randomize tabletop surfaces and backgrounds using a large curated texture library. To build it, we first collected 1,000 diverse surface descriptions via LLM prompting and web crawling, then used Stable Diffusion v2 to generate 20 samples per description (20,000 total). After human-in-the-loop filtering, we obtained 11,000 high-quality textures. This library is applied in simulation to enrich visual diversity and reduce overfitting to clean synthetic environments (see Fig. 4b).

**Lighting Variation.** Real-world environments exhibit diverse illumination conditions, with variations in color temperature, source type, number, and placement. These factors alter object appearance and reflections, challenging vision-based manipulation. To enhance robustness, we randomize light color, type, intensity, and position within physically plausible bounds. As shown in Fig. 4a (second row), changes in color temperature can drastically shift object appearance (e.g., a shoe under warm vs. cool light). Training under such randomized conditions improves policy robustness to real-world illumination shifts.

**Tabletop Heights.** In practice, table heights vary across workspaces, affecting robot perception, kinematics, and interaction. To improve generalization, we uniformly randomize table height within a plausible range during simulation, introducing variability in viewpoints and spatial relations between robot and objects.

**Trajectory-Level Diverse Language Instructions.** To improve robustness to natural language variation, we use a multimodal LLM to generate diverse task templates and multiple object descriptions capturing geometry, appearance, and part-level attributes. Each task and object thus has several alternative phrasings, which can be flexibly combined. For every trajectory, we sample from these pools to compose instructions. For example, in *Move Can Pot*, the template “Use a to place A to the left of B” may yield diverse instructions such as “Use left arm to place sauce can to the left of

gray kitchenpot” or “Use left arm to place white plastic lid sauce can to the left of kitchenpot for boiling and cooking.” This combinatorial augmentation produces a large set of linguistically varied instructions and significantly improves generalization to unseen language and scene configurations (see Appendix H, I).

### 2.3 Embodiment-Aware Grasp Adaptation

Due to differences in DoF and kinematic structures, robotic arms exhibit varying reachable workspaces and preferred manipulation strategies for the same task. For example, when grasping a can, the Franka arm typically favors a top-down approach, while the lower-DoF Piper arm is better suited to side grasps. As a result, a task successfully completed by Franka using a top-down grasp may require a side approach when executed with Piper, as shown in Fig. 6.

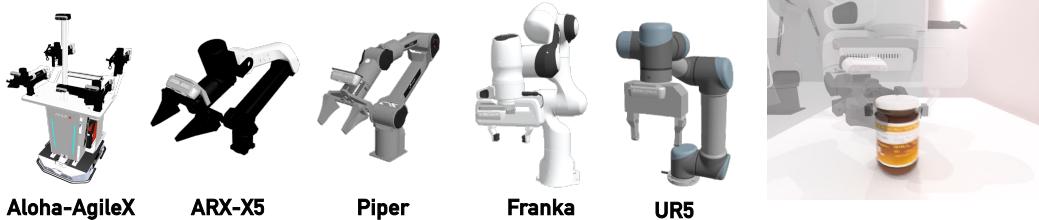


Figure 5: Five RoboTwin 2.0 Embodiments.



Figure 6: Different Grasping Behavior.

To address embodiment-specific variations, we annotate each object with a rich set of candidate manipulation poses that cover multiple grasp axes and approach directions. This design captures both manipulation diversity and robot-specific preferences. To further expand the feasible space, we apply angular perturbations biased toward directions with higher arm reachability. Concretely, for each object we generate candidate grasps by combining preferred operation directions, randomized pose perturbations, and parallelized motion planning attempts.

## 3 RoboTwin 2.0 Data Generator, Benchmark and Large Scale Dataset

### 3.1 RoboTwin-OD: RoboTwin Object Dataset

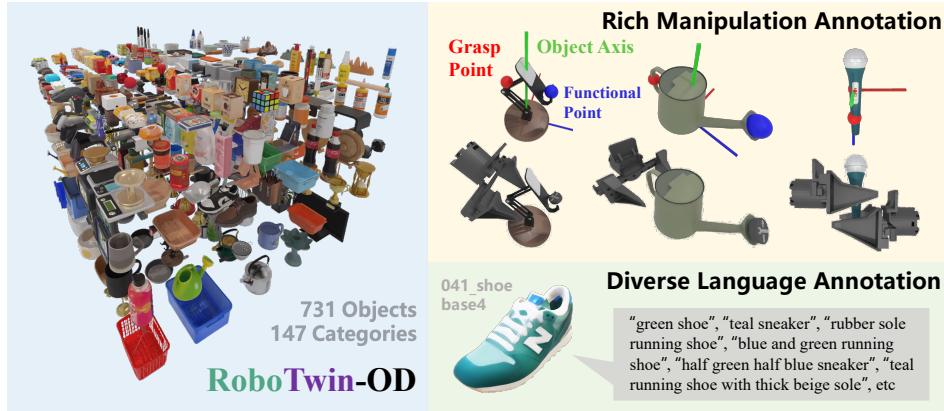


Figure 7: RoboTwin-OD. A large-scale object dataset for robotic manipulation with 147 categories and 731 objects, annotated with rich interaction labels and diverse language descriptions.

To enhance both manipulation capability and visual understanding, we construct a large-scale object dataset with rich semantic annotations, called RoboTwin-OD, covering 147 categories and 731 diverse objects. Specifically, this includes 534 instances across 111 categories that we generated in-house using RGB-to-3D reconstruction via the Rodin platform<sup>1</sup>, followed by convex decomposition and

<sup>1</sup><https://hyper3d.ai/>

mesh merging to ensure physically accurate collision models. In addition, RoboTwin-OD incorporates 153 objects from 27 categories in Objaverse [10], and 44 articulated object instances from 9 categories in SAPIEN PartNet-Mobility [48]. Objects from all sources, including Objaverse, are used to construct cluttered scenes, with Objaverse specifically enhancing the visual and semantic diversity of distractor objects. We also develop a comprehensive texture library for surfaces and backgrounds using generative AI and human-in-the-loop filtering to ensure visual realism and diversity.

For robust manipulation, policies must generalize across diverse objects, which requires datasets with broad category coverage and varied intra-class instances. To facilitate language grounding, we developed an automated object description generator with human verification, producing 15 annotations per object that vary in shape, texture, functionality, part structure, and granularity.

To further support object-centric interaction, we annotate each object with key point–axis information, including placement points, functional points, grasp points, and grasp axes, explicitly encoding affordances. Together with our manipulation API library, these annotations enable generalizable grasp execution in simulation. All object information is available at <http://robotwin-platform.github.io/doc/objects/>.

### 3.2 50 Tasks for Data Generation and Benchmarking

Building on our automated task generation framework, embodiment-adaptive behavior synthesis, and the large-scale RoboTwin-OD asset library, we construct a suite of 50+ dual-arm collaborative manipulation tasks. We further support data collection and evaluation on five distinct robot platforms, enabling comprehensive cross-embodiment benchmarking. Keyframes from representative tasks are shown in Fig. 8, and the complete task descriptions are available at <http://robotwin-platform.github.io/doc/tasks/>. We also pre-collected over 100,000 dual-arm manipulation trajectories across 50 tasks in RoboTwin 2.0, which are available at <https://huggingface.co/datasets/TianxingChen/RoboTwin2.0/tree/main/dataset>.

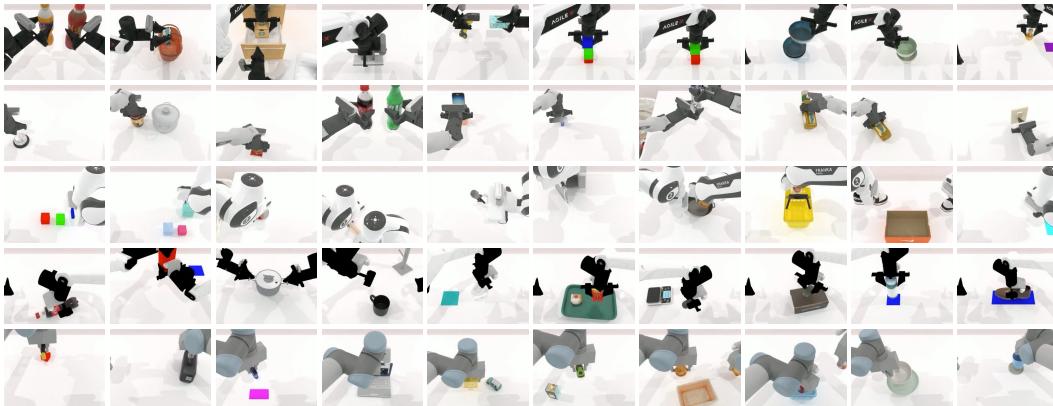


Figure 8: **50 RoboTwin 2.0 Bimanual Manipulation Tasks.**

## 4 Experiment

We design experiments to evaluate the effectiveness of RoboTwin 2.0 in three key aspects: (1) automating the generation of high-quality expert code for manipulation tasks; (2) improving policy robustness to environmental variation via diversified training data; and (3) demonstrating the utility and diversity of RoboTwin 2.0 as a standardized benchmark for evaluating policy generalization across tasks, scenes, and embodiments.

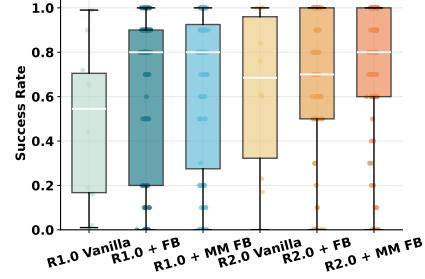
### 4.1 Evaluation of Automated Expert Code Generation

We evaluate our closed-loop expert data generation system on a suite of 10 robotic manipulation tasks, each specified with a natural language instruction. For each configuration, the code-generation agent produces multiple candidate programs, which are executed in simulation to account for stochasticity

in dynamics, control, and perception. Task-level success is defined as the average success rate across all executions, as described in Section 2.1.

**Table 1: Overall performance comparison across RoboTwin variants.** Evaluated on the subset of tasks supported by both RoboTwin 1.0 and RoboTwin 2.0. Per-task success rate comparison is provided in Appendix 8.

Method	ASR	Top5-ASR	CR-Iter	Token
R1.0 Vanilla	47.4%	57.6%	1.00	1236.6
R1.0 + FB	60.4%	71.4%	2.46	1190.4
R1.0 + MM FB	63.9%	74.2%	2.42	1465.0
R2.0 Vanilla	62.1%	68.0%	1.00	<b>569.4</b>
R2.0 + FB	66.7%	73.6%	1.89	581.6
R2.0 + MM FB	<b>71.3%</b>	<b>78.6%</b>	<b>1.76</b>	839.7



**Figure 9: RoboTwin Success Rate Distribution.**

We evaluate performance with four metrics: **ASR** (Average Success Rate), **Top5-ASR** (success over the top-5 candidates per task), **CR-Iter** (average refinement iterations before termination), and **Token** (average number of tokens in generated policy code). Results on RoboTwin 1.0 and 2.0 are reported in Table 1 under three configurations: *Vanilla* (one-shot code generation), *FB* (feedback-based repair via execution logs), and *MM FB* (multimodal feedback with vision–language diagnostics). Per-task success rates are provided in Appendix 8.

Across all settings, multimodal feedback yields consistent gains. In RoboTwin 1.0, ASR improves from 47.4% (Vanilla) to 63.9% (MM FB); in RoboTwin 2.0, it rises from 62.1% to 71.3%. Improvements are also evident in Top5-ASR, suggesting that perceptual feedback disproportionately benefits the best candidate programs. RoboTwin 2.0 converges faster than 1.0 (e.g., 1.76 vs. 2.42 CR-Iter in MM FB), indicating stronger priors and more efficient refinement. Token cost is also substantially reduced, especially in Vanilla (569.4 vs. 1236.6), reflecting more concise initial code.

Figure 9 further shows that feedback narrows the success-rate distribution and raises the median. RoboTwin 2.0 with multimodal feedback achieves compact distributions centered above 80%, highlighting robustness and reliability.

Overall, three findings emerge: (1) vision–language feedback not only detects failures but also guides precise repairs; (2) architectural improvements in RoboTwin 2.0 accelerate convergence and reduce token usage; and (3) combining symbolic execution logs with perceptual diagnostics yields more reliable, semantically aligned expert data. Together, these results validate the effectiveness of our closed-loop, self-improving code generation architecture. Detailed setups, metric definitions, and additional analyses are provided in Appendix G.

## 4.2 Evaluating Efficiency with and without Adaptive Grasping

**Table 2: Overall Performance Comparison between RoboTwin 1.0 and RoboTwin 2.0.**

Method	Aloha-AgileX	Piper	Franka	UR5	ARX-X5	Average
RoboTwin 1.0	65.1%	2.4%	<b>67.3%</b>	<b>57.6%</b>	68.6%	52.2%
RoboTwin 2.0	<b>78.8%</b>	<b>25.1%</b>	67.2%	57.1%	<b>74.2%</b>	<b>60.5%</b>
Difference	<b>+13.7%</b>	<b>+22.7%</b>	<b>-0.1%</b>	<b>-0.5%</b>	<b>+5.6%</b>	<b>+8.3%</b>

To evaluate the effectiveness of our embodiment-aware grasp augmentation strategy, we measure the task success rate of automated data collection across 50 RoboTwin 2.0 tasks on five different robot embodiments. As shown in Table 2, we compare our RoboTwin 2.0 pipeline against the RoboTwin 1.0 baseline, which lacks diverse grasping and candidate augmentation. Results show that our method improves success rates, particularly for robots with constrained planning spaces, achieving an average improvement of 8.3% across all embodiments. Specifically, for high-DoF arms with large reachable workspaces, such as Franka and UR5 (7-DoF), success rates remain largely unchanged, indicating limited benefit when the robot already has sufficient kinematic flexibility. However, for lower-DoF platforms such as Aloha-AgileX, Piper, and ARX-X5 (6-DoF), our method

leads to substantial gains of 13.5%, 22.7%, and 5.6%, respectively. These results demonstrate that our approach provides additional feasible grasp options that effectively mitigate the planning limitations of low-DoF manipulators. Success rates for all tasks can be found in Appendix L.

### 4.3 Assessing the Impact of RoboTwin 2.0 on Policy Robustness

Our goal is to evaluate whether the domain-randomized data in RoboTwin 2.0 can endow models with robustness to environmental perturbations. To this end, we first pre-train RDT and Pi0 on 9,600 expert trajectories collected from 32 tasks (300 per task) under two settings: clean (non-randomized) and domain-randomized.

For comparison, we also evaluate the released pretrained weights of RDT and Pi0 without additional fine-tuning. To further study generalization, we select five unseen tasks and collect 50 clean demonstrations per task for single-task training and fine-tuning. Finally, all policies—including ACT, DP, RDT, and Pi0—are evaluated under domain-randomized conditions to measure robustness in previously unseen environments. Detailed configurations are provided in Appendix C and D.

Table 3: Evaluating the Impact of RoboTwin 2.0 on Policy Robustness.

Simulation Tasks	ACT	DP	RDT	Pi0	RDT +Clean	Pi0 +Clean	RDT +Rand.	Pi0 +Rand.
Stack Bowls Two	0.0%	0.0%	30.0%	41.0%	8.0%	55.0%	49.0%	62.0%
Pick Dual Bottles	0.0%	0.0%	13.0%	12.0%	12.0%	15.0%	17.0%	7.0%
Move Can Pot	4.0%	0.0%	12.0%	21.0%	13.0%	35.0%	18.0%	22.0%
Place Object Basket	0.0%	0.0%	17.0%	2.0%	9.0%	8.0%	6.0%	22.0%
Place Shoe	0.0%	0.0%	7.0%	6.0%	9.0%	6.0%	30.0%	18.0%
Open Laptop	0.0%	0.0%	32.0%	46.0%	21.0%	33.0%	35.0%	50.0%
Press Stapler	6.0%	0.0%	24.0%	29.0%	21.0%	26.0%	27.0%	31.0%
Turn Switch	2.0%	1.0%	15.0%	23.0%	24.0%	21.0%	16.0%	21.0%
<b>Average</b>	2.0%	0.0%	18.8%	22.5%	14.6%	24.9%	24.8%	29.1%

As shown in Table 3, we observe that models fine-tuned with clean data show negligible improvements in average success rate compared to their pretrained counterparts, indicating that data without domain randomization does not help the model handle environmental variations. This also suggests that the low success rate of pretrained VLAs in simulation is not due to a Real-to-Sim gap, since we provide clean simulation data yet observe no clear improvement. In contrast, models pretrained with RoboTwin 2.0 data exhibit significantly improved generalization. Specifically, RDT and Pi0 achieve relative improvements of 31.9% and 29.3%. Notably, this performance gain persists even though the downstream tasks were trained using only clean, non-randomized data. This demonstrates that domain-randomized pretraining with RoboTwin 2.0 effectively equips models with robustness to visual and spatial variations. As a result, models pretrained with RoboTwin 2.0 can adapt to new tasks without requiring additional data augmentation or complex scene variations.

### 4.4 Evaluation on Sim-to-Real Performance

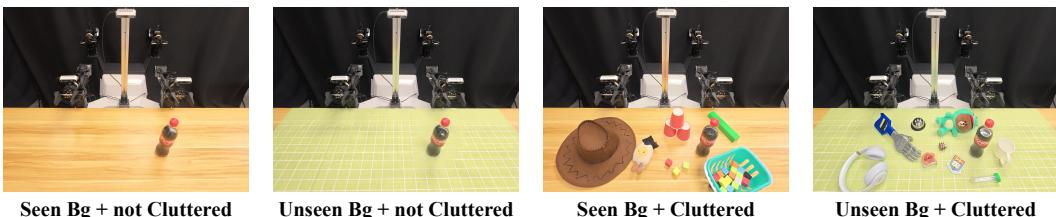


Figure 10: Real-World Evaluation across Four Configurations.

To evaluate RoboTwin 2.0’s effectiveness in enhancing real-world policy robustness, we conduct experiments on four bimanual tasks: *Stack Bowls*, *Handover Block*, *Pick Bottle*, and *Click Bell*. All experiments use RDT as the policy backbone and are executed on the COBOT-Magic dual-arm platform. We compare three training settings: (1) 10 real-world demonstrations in clean tabletop environments; (2) the same demonstrations augmented with 1,000 domain-randomized synthetic trajectories generated under cluttered scenes with varied lighting and backgrounds; and (3) a synthetic-only setting trained solely on the 1,000 domain-randomized trajectories. To further improve robustness to camera jitter and calibration errors, we apply random 3D perturbations to simulated camera poses (position and orientation), with the displacement magnitude bounded by 1 cm.

Evaluation is conducted under four test configurations: (i) clean tabletop with seen backgrounds, (ii) clean tabletop with unseen backgrounds, (iii) cluttered tabletop with seen backgrounds, and (iv) cluttered tabletop with unseen backgrounds (Fig.10). Since the synthetic-only setting excludes seen backgrounds during training, the corresponding entries in Table 4 are omitted. This setup directly tests whether RoboTwin 2.0 enables robust policy generalization without additional real-world data from visually complex environments.

**Table 4: Real-World Experiment Results.** We conduct controlled experiments on 4 dual-arm tasks: *Stack Bowls*, *Handover Block*, *Pick Bottle*, and *Click Bell*, each evaluated under 4 different settings.

Real World Task	Background Type	Cluttered or Not	10 Clean Real	10 Clean Real + 1k RoboTwin 2.0	1k RoboTwin 2.0 (Zero-shot)
<i>Stack Bowls</i>	Seen	False	22.0%	<b>64.0%</b>	/
		True	12.0%	<b>58.0%</b>	/
	Unseen	False	10.0%	50.0%	<b>60.0%</b>
		True	12.0%	<b>56.0%</b>	52.0%
<i>Handover Block</i>	Seen	False	40.0%	<b>48.0%</b>	/
		True	<b>16.0%</b>	12.0%	/
	Unseen	False	36.0%	<b>56.0%</b>	<b>56.0%</b>
		True	0.0%	<b>36.0%</b>	20.0%
<i>Pick Bottle</i>	Seen	False	20.0%	<b>36.0%</b>	/
		True	8.0%	<b>40.0%</b>	/
	Unseen	False	4.0%	<b>26.0%</b>	10.0%
		True	8.0%	28.0%	<b>32.0%</b>
<i>Click Bell</i>	Seen	False	<b>36.0%</b>	24.0%	/
		True	20.0%	<b>56.0%</b>	/
	Unseen	False	12.0%	<b>24.0%</b>	20.0%
		True	16.0%	<b>48.0%</b>	14.0%
<i>Average</i>	Seen	False	29.5%	<b>43.0%<sub>+13.5%</sub></b>	/
		True	14.0%	<b>41.5%<sub>+27.5%</sub></b>	/
	Unseen	False	15.5%	<b>39.0%<sub>+23.5%</sub></b>	<b>36.5%<sub>+21.0%</sub></b>
		True	9.0%	<b>42.0%<sub>+33.0%</sub></b>	<b>29.5%<sub>+20.5%</sub></b>

The experimental results show that real-world bimanual policies augmented with RoboTwin 2.0 achieve clear gains in robustness. In the few-shot setting—where 1,000 domain-randomized synthetic trajectories are combined with just 10 real-world demonstrations—the average success rate across all evaluation settings improves by 24.4%, with per-configuration gains of 13.5%, 27.5%, 23.5%, and 33.0%, respectively. In the zero-shot setting trained solely on synthetic data, we still observe notable improvements of 21.0% and 20.5% on the two unseen-background scenarios. Notably, performance gains become larger in visually complex scenes, indicating that RoboTwin 2.0 is especially effective under challenging conditions.

These improvements stem from two factors: (1) the high visual and physical fidelity of RoboTwin 2.0, enabling direct sim-to-real transfer, and (2) the ability of domain-randomized synthetic data to prepare policies for environmental variations absent from clean real-world demonstrations. Importantly, the

strong performance of the few-shot setting suggests that only minimal real-world data is needed to effectively bridge the sim-to-real gap.

#### 4.5 RoboTwin 2.0 Benchmark

To evaluate the benchmarking utility and generalization challenges of RoboTwin 2.0, we assess five policy models: ACT, DP, RDT, Pi0, and DP3. **All VLAs are fine-tuned from their released pretrained weights in the single-task setting.** Evaluations are conducted on all 50 benchmark tasks using the Aloha AgileX dual-arm embodiment. For each task, **50 clean expert demonstrations** are used for training, and policies are tested with 100 rollouts under two conditions: *Easy* (clean) and *Hard* (domain-randomized with clutter, lighting, textures, and height variations). We provide a visualization of the benchmark setting in Appendix J. We report success rates as indicators of few-shot adaptation and robustness. Detailed setups are provided in Appendix C and D, and full results are available in Appendix K and on the [Leaderboard](#).

Table 5: **Subset of RoboTwin 2.0 benchmark.** Full results in Appendix K and [Leaderboard](#).

Simulation Task	RDT		Pi0		ACT		DP		DP3	
	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
Adjust Bottle	81%	<b>75%</b>	90%	56%	97%	23%	97%	0%	<b>99%</b>	3%
Beat Block Hammer	<b>77%</b>	<b>37%</b>	43%	21%	56%	3%	42%	0%	72%	8%
Blocks Ranking RGB	3%	0%	<b>19%</b>	<b>5%</b>	1%	0%	0%	0%	3%	0%
Blocks Ranking Size	0%	0%	<b>7%</b>	<b>1%</b>	0%	0%	1%	0%	2%	0%
Click Alarmclock	61%	12%	63%	11%	32%	4%	61%	5%	<b>77%</b>	<b>14%</b>
Click Bell	80%	<b>9%</b>	44%	3%	58%	3%	54%	0%	<b>90%</b>	0%
Dump Bin Bigbin	64%	32%	83%	24%	68%	1%	49%	0%	<b>85%</b>	<b>53%</b>
Grab Roller	74%	43%	96%	<b>80%</b>	94%	25%	<b>98%</b>	0%	<b>98%</b>	2%
Handover Block	45%	<b>14%</b>	45%	8%	42%	0%	10%	0%	<b>70%</b>	0%
Handover Mic	90%	<b>31%</b>	98%	13%	85%	0%	53%	0%	<b>100%</b>	3%
Hanging Mug	<b>23%</b>	<b>16%</b>	11%	3%	7%	0%	8%	0%	17%	1%
Lift Pot	72%	9%	84%	<b>36%</b>	88%	0%	39%	0%	<b>97%</b>	0%
<b>...</b>										
Move Pillbottle Pad	8%	0%	21%	<b>1%</b>	0%	0%	1%	0%	<b>41%</b>	0%
<b>Average (in %)</b>	34.5	13.7	46.4	<b>16.3</b>	29.7	1.7	28.0	0.6	<b>55.2</b>	5.0

Fig.5 and Appendix K report results on 50 tasks. Non-pretrained models (ACT, DP, DP3) perform poorly under Hard conditions, while pretrained models (RDT, Pi0) show stronger resilience, suggesting that vision–language–action pretraining provides useful priors for generalization. Still, success rates drop by 20.8% (RDT) and 30.1% (Pi0) from clean to randomized settings, underscoring robustness under domain shifts as a key challenge, likely due to limited diversity in pretraining data. DP3 achieves the best few-shot performance with limited samples, highlighting the role of 3D information, though its strong results partly stem from perfect point clouds and clean background segmentation in simulation. Together with Sections 4.3 and 4.4, these findings show RoboTwin 2.0’s value in complementing existing datasets with diverse, domain-randomized trajectories for improved generalization and robustness.

## 5 Related Work

### 5.1 Datasets and Benchmarks for Robotic Manipulation

Physics-based simulators underpin modern manipulation research. Existing platforms provide complementary strengths: SAPIEN [48] enables dynamic interaction with 2,300+ articulated objects; ManiSkill2 [16] supplies millions of demonstrations; Meta-World [50], CALVIN [32], LIBERO [30], and RoboVerse [15] target multi-task, language-conditioned, lifelong, and domain-randomized settings; RoboCasa [35] offers large-scale human demonstrations but lacks automation and dual-arm focus.

Large-scale real-world datasets further bridge sim-to-real: AgiBot World [4], RoboMIND [47], Open X-Embodiment [36], and Bridge [12] contribute millions of trajectories across diverse tasks, robots, and environments.

RoboTwin-1.0 [34] mirrored real demonstrations with simulated replicas for dual-arm benchmarking. In this work, RoboTwin 2.0 integrates LLM-driven feedback and systematic domain randomization across visual, physical, and task dimensions, producing richer corpora that improve policy robustness and generalization. A detailed comparison with prior benchmarks is provided in Appendix B.

## 5.2 Robot Learning in Manipulation

Many task-specific policy architectures [42, 21, 51, 9, 14, 7, 29, 42, 27, 28, 45, 44, 8] achieve strong single-task performance but struggle to transfer across embodiments. In contrast, foundation models trained on million-scale, multi-robot corpora have enabled robust zero-shot generalization: RT-1 [3] unifies vision, language and actions in a single transformer for real-time kitchen tasks; RT-2 [2] co-fine-tunes large vision–language models on web and robot data to unlock semantic planning and object reasoning; diffusion-based RDT-1B [31] and the  $\pi_0$ [1] capture diverse bimanual dynamics from over a million episodes. Vision–language–action (VLA) frameworks like OpenVLA [23] and CogACT [26], together with adaptations like Octo [40], LAPA [49], and OpenVLA-OFT [22] demonstrate efficient fine-tuning to novel robots and sensor modalities.

To further advance this direction, our work introduces digital-twin data collection paired with extensive domain randomization, yielding datasets that closely mirror real robot dynamics and support the training of robust and generalizable bi-manual manipulation policies.

## 5.3 Domain Randomization in Imitation Learning

Prior works have shown that randomizing visual and physical parameters, including but not limited to textures, lighting, camera pose, mass, friction and control latency combined with noise injection in expert demonstrations, enables sim-to-real transfer and robust visuomotor policies [41, 37, 5, 7, 28], and optimizing over worst-case ensembles further improves resilience to extreme domain shifts [38, 25, 27]. However, these approaches apply randomization in isolation and lack bidirectional digital-twin feedback; our method integrates interactive simulation feedback with systematic domain randomization to generate higher-fidelity imitation data.

## 6 Conclusion

This paper presented RoboTwin 2.0, a scalable simulation framework for generating diverse, high-fidelity expert data to support robust bimanual manipulation. Our system integrates MLLM-based task generation, embodiment-adaptive behavior synthesis, and comprehensive domain randomization to address key limitations in prior synthetic data generator.

By leveraging an annotated object library and automating trajectory generation, RoboTwin 2.0 produces data with rich visual, linguistic, and physical diversity while minimizing manual engineering effort. Experiments demonstrate its effectiveness in improving policy robustness to cluttered environments, generalization to unseen tasks, and cross-embodiment manipulation.

These findings highlight the importance of scalable, automated generation of semantically rich, domain-randomized data for learning robust manipulation policies. RoboTwin 2.0 provides a foundation for unified benchmarks and scalable sim-to-real pipelines, with future work focusing on real-world deployment and multi-object task complexity.

## 7 Acknowledgments

This paper is partially supported by AgileX Robotics, D-Robotics, and the Jockey Club STEM Lab of Autonomous Intelligent Systems funded by The Hong Kong Jockey Club Charities Trust.

## References

- [1] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. *pi\_0*: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.
- [2] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.
- [3] 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. *arXiv preprint arXiv:2212.06817*, 2022.
- [4] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xu Huang, Shu Jiang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.
- [5] Yevgen Chebotar, Ankur Handa, Viktor Makoviychuk, Miles Macklin, Jan Issac, Nathan Ratliff, and Dieter Fox. Closing the sim-to-real loop: Adapting simulation randomization with real world experience. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8973–8979. IEEE, 2019.
- [6] Junting Chen, Yao Mu, Qiaojun Yu, Tianming Wei, Silang Wu, Zhecheng Yuan, Zhixuan Liang, Chao Yang, Kaipeng Zhang, Wenqi Shao, Yu Qiao, Huazhe Xu, Mingyu Ding, and Ping Luo. Roboscript: Code generation for free-form manipulation tasks across real and simulation, 2024.
- [7] Tianxing Chen, Yao Mu, Zhixuan Liang, Zanxin Chen, Shijia Peng, Qiangyu Chen, Mingkun Xu, Ruizhen Hu, Hongyuan Zhang, Xuelong Li, et al. G3flow: Generative 3d semantic flow for pose-aware and generalizable object manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1735–1744, 2025.
- [8] Tianxing Chen, Kaixuan Wang, Zhaojun Yang, Yuhao Zhang, Zanxin Chen, Baijun Chen, Wanxi Dong, Ziyuan Liu, Dong Chen, Tianshuo Yang, et al. Benchmarking generalizable bimanual manipulation: Robotwin dual-arm collaboration challenge at cvpr 2025 meis workshop. *arXiv preprint arXiv:2506.23351*, 2025.
- [9] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.
- [10] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13142–13153, 2023.
- [11] Shengliang Deng, Mi Yan, Songlin Wei, Haixin Ma, Yuxin Yang, Jiayi Chen, Zhiqi Zhang, Taoyu Yang, Xuheng Zhang, Heming Cui, et al. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. *arXiv preprint arXiv:2505.03233*, 2025.
- [12] 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. *arXiv preprint arXiv:2109.13396*, 2021.
- [13] Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, et al. Codebert: A pre-trained model for programming and natural languages. *arXiv preprint arXiv:2002.08155*, 2020.
- [14] Zipeng Fu, Tony Z Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024.
- [15] Haoran Geng, Feishi Wang, Songlin Wei, Yuyang Li, Bangjun Wang, Boshi An, Charlie Tianyue Cheng, Haozhe Lou, Peihao Li, Yen-Jen Wang, et al. Roboverse: Towards a unified platform, dataset and benchmark for scalable and generalizable robot learning. *arXiv preprint arXiv:2504.18904*, 2025.
- [16] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, et al. Maniskill2: A unified benchmark for generalizable manipulation skills. In *The Eleventh International Conference on Learning Representations*, 2023.

- [17] Daya Guo, Shuai Lu, Nan Duan, Yanlin Wang, Ming Zhou, and Jian Yin. Unixcoder: Unified cross-modal pre-training for code representation. *arXiv preprint arXiv:2203.03850*, 2022.
- [18] Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, and Ping Luo. Hiagent: Hierarchical working memory management for solving long-horizon agent tasks with large language model. *arXiv preprint arXiv:2408.09559*, 2024.
- [19] Mengkang Hu, Tianxing Chen, Yude Zou, Yuheng Lei, Qiguang Chen, Ming Li, Yao Mu, Hongyuan Zhang, Wenqi Shao, and Ping Luo. Text2world: Benchmarking large language models for symbolic world model generation. *arXiv preprint arXiv:2502.13092*, 2025.
- [20] Pu Hua, Minghuan Liu, Annabella Macaluso, Yunfeng Lin, Weinan Zhang, Huazhe Xu, and Lirui Wang. Gensim2: Scaling robot data generation with multi-modal and reasoning llms. In *8th Annual Conference on Robot Learning*.
- [21] Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. *arXiv preprint arXiv:2402.10885*, 2024.
- [22] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025.
- [23] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. Openvla: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*.
- [24] Zhiqian Lan, Yuxuan Jiang, Ruiqi Wang, Xuanbing Xie, Rongkui Zhang, Yicheng Zhu, Peihang Li, Tianshuo Yang, Tianxing Chen, Haoyu Gao, et al. Autobio: A simulation and benchmark for robotic automation in digital biology laboratory. *arXiv preprint arXiv:2505.14030*, 2025.
- [25] Michael Laskey, Jonathan Lee, Roy Fox, Anca Dragan, and Ken Goldberg. Dart: Noise injection for robust imitation learning. In *Conference on robot learning*, pages 143–156. PMLR, 2017.
- [26] 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.
- [27] Zhixuan Liang, Yao Mu, Mingyu Ding, Fei Ni, Masayoshi Tomizuka, and Ping Luo. Adaptddiffuser: Diffusion models as adaptive self-evolving planners. In *International Conference on Machine Learning*, pages 20725–20745. PMLR, 2023.
- [28] Zhixuan Liang, Yao Mu, Hengbo Ma, Masayoshi Tomizuka, Mingyu Ding, and Ping Luo. Skillddiffuser: Interpretable hierarchical planning via skill abstractions in diffusion-based task execution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16467–16476, 2024.
- [29] Zhixuan Liang, Yao Mu, Yixiao Wang, Tianxing Chen, Wenqi Shao, Wei Zhan, Masayoshi Tomizuka, Ping Luo, and Mingyu Ding. Dexhanddiff: Interaction-aware diffusion planning for adaptive dexterous manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1745–1755, 2025.
- [30] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.
- [31] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.
- [32] 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.
- [33] Yao Mu, Junting Chen, Qing-Long Zhang, Shoufa Chen, Qiaojun Yu, Chongjian Ge, Runjian Chen, Zhixuan Liang, Mengkang Hu, Chaofan Tao, et al. Robocodex: Multimodal code generation for robotic behavior synthesis. In *International Conference on Machine Learning*, pages 36434–36454. PMLR, 2024.
- [34] Yao Mu, Tianxing Chen, Zanxin Chen, Shijia Peng, Zhiqian Lan, Zeyu Gao, Zhixuan Liang, Qiaojun Yu, Yude Zou, Mingkun Xu, et al. Robotwin: Dual-arm robot benchmark with generative digital twins. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025.

- [35] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In *Robotics: Science and Systems (RSS)*, 2024.
- [36] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.
- [37] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3803–3810. IEEE, 2018.
- [38] Aravind Rajeswaran, Sarvjeet Ghotra, Balaraman Ravindran, and Sergey Levine. Epopt: Learning robust neural network policies using model ensembles. In *International Conference on Learning Representations*, 2017.
- [39] Shuo Ren, Daya Guo, Shuai Lu, Long Zhou, Shujie Liu, Duyu Tang, Neel Sundaresan, Ming Zhou, Ambrosio Blanco, and Shuai Ma. Codebleu: a method for automatic evaluation of code synthesis. *arXiv preprint arXiv:2009.10297*, 2020.
- [40] 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. *arXiv preprint arXiv:2405.12213*, 2024.
- [41] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.
- [42] Chenxi Wang, Hongjie Fang, Hao-Shu Fang, and Cewu Lu. Rise: 3d perception makes real-world robot imitation simple and effective. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2870–2877. IEEE, 2024.
- [43] Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation, 2023.
- [44] Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. Dexvla: Vision-language model with plug-in diffusion expert for general robot control. *arXiv preprint arXiv:2502.05855*, 2025.
- [45] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Zhibin Tang, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, Yixin Peng, Feifei Feng, and Jian Tang. Tinyvla: Toward fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters*, 10(4):3988–3995, 2025.
- [46] Wu Wen, Xiaobo Xue, Ya Li, Peng Gu, and Jianfeng Xu. Code similarity detection using ast and textual information. *International Journal of Performability Engineering*, 15(10):2683, 2019.
- [47] Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhiqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *arXiv preprint arXiv:2412.13877*, 2024.
- [48] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11097–11107, 2020.
- [49] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. In *CoRL 2024 Workshop on Whole-body Control and Bimanual Manipulation: Applications in Humanoids and Beyond*.
- [50] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR, 2020.
- [51] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy. *arXiv e-prints*, pages arXiv–2403, 2024.
- [52] Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.

## A Contributions

### *Project Leaders*

Tianxing Chen, Yao Mu, Zhixuan Liang

### *Roadmap & Methodology*

Yao Mu, Tianxing Chen, Ping Luo, Yusen Qin, Xiaokang Yang, Kaixuan Wang

### *Data Generator & Benchmark*

Tianxing Chen, Zanxin Chen, Baijun Chen, Qiwei Liang, Zixuan Li, Xianliang Lin

### *CodeGen Agent*

Yibin Liu, Zanxin Chen, Yiheng Ge, Tianxing Chen, Mengkang Hu

### *RoboTwin-OD*

Baijun Chen, Qiangyu Chen, Kailun Su, Xuanbing Xie, Zanxin Chen

### *Policies Training & Evaluation*

Tianxing Chen, Zijian Cai, Tian Nian, Huan-ang Gao, Tianling Xu

### *Real-World Deployment*

Tianxing Chen, Tian Nian, Weiliang Deng

### *Domain Randomization*

Baijun Chen, Yubin Guo, Qiwei Liang, Zhenyu Gu, Guodong Liu, Zanxin Chen, Tianxing Chen

## B Benchmarking RoboTwin 2.0 Against Existing Datasets

We compare RoboTwin 2.0 against existing benchmarks and datasets across several key dimensions, including the number of supported tasks, the presence of domain randomization, support for automatic data generation, and compatibility with vision-language-action (VLA) model training and evaluation. The comparison is summarized in Table 6.

Table 6: **Comparison of RoboTwin 2.0 with previous manipulation benchmarks and datasets.**

Benchmark & Dataset	#Tasks	Domain Randomization	Auto Data Generation	VLA Model Train & Eval
Meta-world [50]	50	✗	✓	✗
Robosuite [52]	9	✗	✗	✗
RoboCasa [50]	25	✓	✗	✗
Maniskill2 [16]	20	✗	✓	✗
AutoBio [24]	16	✗	✓	✓
RoboTwin 1.0 [34]	14	✗	✓	✓
RoboTwin 2.0 (ours)	50	✓	✓	✓

## C Domain Randomization Setting

Domain randomization in all experiments includes cluttered scenes, random lighting, table height variation (up to 3 cm), unseen language instructions and randomized background textures.

## D Policies Training Details

**RDT** in experiment 4.3 was pretrained for 100,000 steps with a batch size of 16 per GPU on 8 GPUs, and all single-task fine-tuning was conducted for 10,000 steps with a batch size of 16 per GPU on 4 GPUs.

**Pi0** in experiment 4.3 was pretrained for 100,000 steps with a batch size of 32, and all fine-tuning was performed for 30,000 steps using the same batch size.

**ACT** was trained under a unified setup with a chunk size of 50, batch size of 8, and single-GPU training for 6,000 epochs. During deployment, we applied `temporal_agg` for temporal aggregation to improve execution stability.

**DP** was trained for 600 epochs with a batch size of 128 and a planning horizon of 8.

**DP3** was trained for 3,000 epochs with a batch size of 256, using a planning horizon of 8 and a point cloud resolution of 1,024, with precise segmentation of the background and tabletop.

## E Support for Flexible Embodiment Combinations

Our object-centric, embodiment-agnostic data generation framework enables seamless deployment across a wide range of dual-arm robotic systems. The pipeline supports flexible embodiment configurations, allowing arbitrary combinations of heterogeneous manipulators and relative arm placements. This design ensures compatibility with diverse hardware setups and facilitates extensibility to future robotic platforms.



Figure 11: **Heterogeneous Dual-Arm Control via Object-Centric Manipulation.**

To execute high-success-rate manipulation trajectories across different embodiments (see Section 2.3), we integrate Curobo, a high-performance, GPU-accelerated motion planner that enables efficient and reliable planning under varied kinematic constraints.

Currently, our framework supports five robotic arms—Franka, Piper, UR5, ARX-X5, and Aloha-AgileX—along with multiple gripper types, including the Panda gripper and WSG gripper. As shown in Fig. 11, we demonstrate successful task executions across a variety of dual-arm pairings, highlighting RoboTwin 2.0’s ability to scale to heterogeneous robot configurations and its readiness for future real-world deployment.

## F Improvements of RoboTwin 2.0 over RoboTwin 1.0 Policy Codebase

Metric	RoboTwin 1.0	RoboTwin 2.0
Prompt Token Length ↓	5901.0	<b>4719.1</b>
Code Token Length ↓	1236.6	<b>569.4</b>
Parallelism Control ↑	✗	✓
AST Similarity [46] ↑	23.72%	<b>44.78%</b>
CodeBLEU Similarity [39] ↑	17.18%	<b>18.53%</b>
CodeBERT Similarity [13] ↑	97.72%	<b>98.80%</b>
Unixcoder Similarity [17] ↑	76.24%	<b>82.21%</b>
Avg. VLM Token Cost (per observation)	—	<b>6894</b>

Table 7: **Code Generation Efficiency and Quality Comparison.** Evaluation of prompt and generated code characteristics, along with code similarity metrics (AST Structural Similarity, CodeBERT, Unixcoder cosine similarity) against expert-written code, for RoboTwin 1.0 and RoboTwin 2.0 in zero-shot generation. The VLM observer cost is also reported for RoboTwin 2.0.

We first quantify the architectural impact of RoboTwin 2.0 in a one-shot generation without code repair and iterative refinement. Table 7 shows that RoboTwin 2.0 yields significantly shorter programs (569.4 vs. 1236.6 tokens), with reduced prompt length and higher structural similarity to human-written code. Crucially, it enables dual-arm parallelism via a unified API abstraction, which is absent in RoboTwin 1.0.

These improvements stem from the structured prompting and geometric API modularization designed into RoboTwin 2.0. Higher AST similarity (+21.06%), CodeBERT similarity (+1.08%), and Unixcoder alignment (+5.97%) indicate that RoboTwin 2.0 not only reduces code size but also improves semantic clarity and functional alignment.

In addition, RoboTwin 2.0 integrates a **VLM observer**, a plug-and-play module triggered only when execution fails. To quantify its overhead, we estimated VLM usage via the Kimi API (assuming each image = 1,024 tokens) over three representative tasks: the average cost was 6,295 input tokens and 599 output tokens, totaling **6,894 tokens**. While this introduces moderate overhead, the VLM enables

RoboTwin 2.0 to catch and correct errors invisible to execution logging, significantly enhancing robustness and overall task success. Importantly, the observer remains optional and can be disabled when prioritizing token efficiency.

## G Experimental Details and Metric Definitions for Code Generation

We use the *DeepSeek-V3* model for program synthesis and the *moonshot-v1-32k-vision-preview* model for multimodal error localization and verification. These models were selected for their strong performance in language reasoning and visual understanding while maintaining efficiency suitable for large-scale iterative refinement. The success rate of the  $i$ -th program is computed as  $R_i = \frac{1}{M} \sum_{j=1}^M s_{i,j}$ , and the final success rate for a given task under a specific system variant is then defined as  $R_{\text{task}} = \frac{1}{N} \sum_{i=1}^N R_i$ . For detailed usage instructions, refer to <https://robotwin-platform.github.io/doc/usage/expert-code-gen.html>.

### G.1 Metric Definitions

We report the following metrics across all tasks:

**ASR (Average Success Rate)** is the average of  $R_{\text{task}}$  across all 10 tasks. It reflects overall task performance across all generated programs.

**Top5-ASR** is the mean success rate computed using only the top 5 highest-performing programs per task. This metric estimates system potential under a best-of-selection strategy.

**CR-Iter** indicates the average number of feedback iterations required per task before reaching a success rate above 50% or exhausting the iteration budget.

**Token** denotes the average number of tokens of policy code generated by the language model per task. It serves as a proxy for computational cost and LLM inference budget.

These metrics jointly evaluate both the reliability and efficiency of the expert data generation pipeline under varying conditions of feedback, model capability, and refinement strategy.

### G.2 Task-Specific Performance Comparison on Code Generation

We compare the code generation success rates of RoboTwin 2.0 and RoboTwin 1.0 across all tasks. As shown, RoboTwin 2.0 consistently matches or outperforms the baseline on the majority of tasks, demonstrating the effectiveness of our multimodal feedback and refinement pipeline.

Task	R1.0 Vanilla	R1.0 + FB	R1.0 + MM FB	R2.0 Vanilla	R2.0 + FB	R2.0 + MMFB
beat_block_hammer	16%	48%	<b>56%</b>	23%	34%	53%
handover_block	2%	41%	<b>45%</b>	17%	<b>50%</b>	27%
pick_diverse_bottles	<b>65%</b>	<b>65%</b>	64%	60%	60%	62%
pick_dual_bottles	99%	99%	<b>100%</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>
place_container_plate	66%	79%	<b>91%</b>	84%	84%	82%
place_dual_shoes	19%	22%	<b>25%</b>	0%	2%	22%
place_empty_cup	90%	90%	<b>100%</b>	61%	61%	85%
place_shoe	72%	90%	90%	<b>100%</b>	<b>100%</b>	<b>100%</b>
stack_blocks_three	1%	2%	4%	76%	76%	<b>82%</b>
stack_blocks_two	44%	68%	64%	<b>100%</b>	<b>100%</b>	<b>100%</b>

Table 8: Task-Specific Performance Comparison between RoboTwin 2.0 and RoboTwin 1.0. R1.0/R2.0: RoboTwin 1.0 / 2.0. Bold numbers indicate the best result for each task.

### G.3 Per-task Success Rates of Code Generation

We report the success rates of all tasks in Tab. 9.

Table 9: Per-task success rates of our proposed R2.0 + MM FB algorithm on all RoboTwin 2.0-supported tasks.

Task	Rate	Task	Rate	Task	Rate	Task	Rate
Adjust Bottle	100%	Beat Block Hammer	53%	Blocks Ranking Rgb	80%	Blocks Ranking Size	80%
Click Alarmclock	0%	Click Bell	10%	Dump Bin Bigbin	0%	Grab Roller	74%
Handover Block	27%	Handover Mic	0%	Hanging Mug	0%	Lift Pot	40%
Move Can Pot	30%	Move Pillbottle Pad	50%	Move Playingcard Away	90%	Move Stapler Pad	100%
Open Laptop	0%	Open Microwave	0%	Pick Diverse Bottles	62%	Pick Dual Bottles	100%
Place A2B Left	50%	Place A2B Right	60%	Place Bread Basket	0%	Place Bread Skillet	0%
Place Can Basket	0%	Place Cans Plasticbox	100%	Place Container Plate	82%	Place Dual Shoes	22%
Place Empty Cup	85%	Place Fan	70%	Place Burger Fries	100%	Place Mouse Pad	100%
Place Object Basket	0%	Place Object Scale	80%	Place Object Stand	90%	Place Phone Stand	0%
Place Shoe	100%	Press Stapler	0%	Put Bottles Dustbin	0%	Put Object Cabinet	0%
Rotate Qrcode	80%	Scan Object	0%	Shake Bottle	0%	Shake Bottle Horizontally	0%
Stack Blocks Three	82%	Stack Blocks Two	100%	Stack Bowls Three	20%	Stack Bowls Two	30%
Stamp Seal	20%	Turn Switch	0%	Avg Success Rate		43.34%	

#### G.4 Multimodal Observation and Error Localization

To further investigate the capability of the VLM observer, we manually curated a dataset of 130 execution sequences, including 101 failed trials and 29 successful trials. Each sequence consists of the natural language task instruction, a series of visual observations, and policy code. This dataset enables us to evaluate both binary error detection and fine-grained error localization.

**Error Detection.** The VLM observer was first tasked with evaluating whether a robotic execution successfully completed the instructed task. The confusion matrix is as follows: TP = 16, FP = 61, TN = 40, FN = 13. The derived performance metrics are: Accuracy = 0.431, Precision = 0.208, Recall = 0.552, and F1-score = 0.302. These results suggest that while the VLM observer is reasonably sensitive in identifying execution failures (recall), its tendency to over-predict errors leads to a high false positive rate.

**Error Localization.** For failed executions, we further tested whether the VLM observer could correctly localize the program step or code snippet that led to failure. Among the 40 true negative samples (correctly identified failed trials), the observer successfully pinpointed the failure step in 12 cases, yielding a localization accuracy of 30%. This indicates that, while the observer shows promising diagnostic ability, there remains significant room for improvement in reliable error attribution.

**Case Study.** We illustrate one successful case of VLM observer reasoning. Given a task instruction and five visual observations, the observer produced a step-by-step analysis:

- Step 1: The left arm successfully picks up the mug from the left side.
- Step 2: The left arm successfully places the mug at the middle position.
- Step 3: The right arm successfully picks up the mug from the middle position.
- Step 4: The right arm attempts to hang the mug on the rack but fails.
- Step 5: The right arm is in the process of moving away after attempting to hang the mug.

The overall task was not successfully completed. The failure occurred in Step 4, where the right arm attempted to hang the mug on the rack but did not succeed. The mug is not properly positioned on the rack. The error message in the code indicates an `AttributeError`: 'list' object has no attribute 'p'. This error occurs in the line where the `target_point` is being set to `middle_target_pose.p`. The issue is likely due to the fact that `middle_target_pose` is a list, and lists do not have an attribute 'p'. The code should be modified to correctly access the position data from the `middle_target_pose` list.

To fix the error, the code should be updated to access the position data correctly. For example, if `middle_target_pose` is a list containing the position and orientation, the position data should be accessed using `middle_target_pose[0]` instead of `middle_target_pose.p`. Additionally, the failure in Step 4 could be due to the incorrect position or orientation of the mug when attempting to hang it on the rack. The code should be reviewed to ensure that the mug is positioned and oriented correctly before executing the hanging motion.

**Failure Modes.** Despite these successes, we also observed cases where the VLM observer fails. Misclassifications primarily occur in failed executions: the observer may hallucinate correct outcomes or overlook subtle orientation errors. For instance, in a stacking task where boxes are misaligned, the observer sometimes marks the execution as successful if all boxes are placed on top of each other. Similarly, in tasks requiring fine-grained pose accuracy (e.g., placing shoes in the correct orientation), small but critical errors may be ignored. Furthermore, failures stemming from invisible factors—such as incorrect grasp axis parameters—remain challenging for purely vision-based observers to diagnose.

## G.5 LLM-Generated Code and Human-Written Code Case Study

### LLM-Generated Code (gpt\_place\_shoe)

```
class gpt_place_shoe(place_shoe):
    def play_once(self):
        # Initial observation
        self.save_camera_images(task_name="place_shoe", step_name="step1_initial_scene_state", generate_num_id="generate_num_0")

        # Get the shoe's position to determine which arm to use
        shoe_pose = self.shoe.get_pose()
        shoe_position = shoe_pose.p
        arm_tag = ArmTag("left" if shoe_position[0] < 0 else "right")

        # Grasp the shoe
        self.move(self.grasp_actor(actor=self.shoe, arm_tag=arm_tag, pre_grasp_dis=0.1, grasp_dis=0))
        self.save_camera_images(task_name="place_shoe", step_name="step2_shoe_grasped", generate_num_id="generate_num_0")

        # Lift the shoe up to avoid collision
        self.move(self.move_by_displacement(arm_tag=arm_tag, z=0.07, move_axis='world'))
        self.save_camera_images(task_name="place_shoe", step_name="step3_shoe_lifted", generate_num_id="generate_num_0")

        # Get target pose from the top functional point of the target block
        target_pose = self.target_block.get_functional_point(1, "pose")

        # Place the shoe on the target block with head towards left side
        self.move(self.place_actor(actor=self.shoe, arm_tag=arm_tag, target_pose=target_pose, functional_point_id=0, pre_dis=0.1, dis=0.02, is_open=True, constrain="align", pre_dis_axis='fp'))
        self.save_camera_images(task_name="place_shoe", step_name="step4_shoe_placed", generate_num_id="generate_num_0")

        # Lift the gripper up after placing
        self.move(self.move_by_displacement(arm_tag=arm_tag, z=0.07, move_axis='world'))

        # Return arm to origin
        self.move(self.back_to_origin(arm_tag=arm_tag))

        # Final observation
        self.move(self.save_camera_images(task_name="place_shoe", step_name="step5_final_scene_state", generate_num_id="generate_num_0"))
```

Listing 1: LLM-Generated Code for gpt\_place\_shoe Task

### Human-Written Code (place\_shoe)

```
class place_shoe(base_task):
    def play_once(self):
        # Get the shoe's position to determine which arm to use
        shoe_pose = self.shoe.get_pose().p
```

```

    arm_tag = ArmTag("left" if shoe_pose[0] < 0 else "right")

    # Grasp the shoe with specified pre-grasp distance and gripper position
    self.move(self.grasp_actor(self.shoe, arm_tag=arm_tag, pre_grasp_dis=0.1,
                                gripper_pos=0))

    # Lift the shoe up by 0.07 meters in z-direction
    self.move(self.move_by_displacement(arm_tag=arm_tag, z=0.07))

    # Get target_block's functional point as target pose
    target_pose = self.target_block.get_functional_point(0)

    # Place the shoe on the target_block with alignment constraint and specified
    # pre-placement distance
    self.move(self.place_actor(self.shoe, arm_tag=arm_tag, target_pose=
                                target_pose, functional_point_id=0, pre_dis=0.12, constrain="align"))

    # Open the gripper to release the shoe
    self.move(self.open_gripper(arm_tag=arm_tag))

```

Listing 2: Human-Written Code for place\_shoe Task

The LLM generated code tends to be more verbose, explicitly logging intermediate visual states and detailing parameters (e.g., `pre_dis_axis='fp'`, `is_open=True`), while human-written scripts are more minimal, omitting intermediate steps and favoring compact execution. Despite functional similarity, the structural differences illustrate that **MLLM-generated programs are not only executable but emphasize step-by-step clarity**, contributing to more robust feedback and repair.

## H Task Instruction and Object Description Example

### Instruction Templates (task: ‘Pick Dual Bottles’)

```
"Use {a} to place {A} left of {B}.", "Set {A} to the left of {B}.", "Move {A} beside {B} using
{a}.", "Place {A} on {B}'s left side.", "Using {a}, position {A} next to {B}.", "Stick {A} on the
left of {B}.", "Use {a} and place {A} on {B}'s left.", etc
```

### Object Description

```
# object id - '001_bottle/0':
"red bottle", "red soda bottle", "plastic red bottle", "red bottle with yellow label", "red
plastic bottle with smooth surface", "yellow text printed on red bottle surface", "red bottle with
white label design and markings", "red bottle with white sealing and brown top screw cap", etc
# object id - '039_mug/0':
"black mug", "dark coffee mug", "sleek black mug", "black ceramic mug", "single-handle mug",
"smooth black surface mug", "medium-sized drinking mug", "round mug with curved side", "dark mug
with sturdy handle", "solid black mug with smooth finish", etc
```

## I Prompts for Generating Task Instructions and Object Descriptions

```

# Task Instruction Template
- Goal: Generate task instruction template
- Requirements:
  - Generate 60 items. Vary in sentence length and structure
  - Use natural action verbs (grab, slide, place)
- split
  - 50 items for training
  - 10 items for evaluation

## Schema Requirements

```

```

- Goal: Use placeholders for objects in instructions
- Requirements:
  - Format: {X} for objects defined in schema
  - Include all object placeholders ({A-Z}) in every instruction
  - Omit arm references and placeholders ({a-z}) in 50% of instructions
  - Ensure natural flow when placeholders are replaced with actual values

# Object Description
- Goal: Generate natural object descriptions for robotic manipulation
- Requirements:
  - Generate 15 items. Vary in sentence length and structure
  - Use natural oral language
  - Include essential physical properties (color, shape, size, texture)
  - Use noun-focused phrases
  - For multi-part objects, use structures like 'X with Y'
- split
  - 12 items for training
  - 3 items for evaluation

# Episode
An episode is a specified task, in which each task may have different objects to be manipulated,
resulting in the same task template being reused by replacing the placeholders with specific objects.
For example:
{A} -> 'medium-sized yellow bottle'
{A} -> 'green drink bottle with bold labels'

General Task -> Specific Episode:
{A} -> bottle/0.glb
{A} -> bottle/1.glb

The number of task instructions for an episode can be calculated by:
Episode_num = TaskInstruction_num * ObjectDescription_num

```

Listing 3: Prompts for Generating Task Instructions and Object Descriptions

## J RoboTwin 2.0 Benchmark Setting Visualization

We visualize the simulation settings of the RoboTwin 2.0 benchmark in Fig. 12. All models are trained on 50 clean (non-randomized) demonstrations per task (blue). For evaluation, the *Easy* setting also uses clean environments, while the *Hard* setting employs domain-randomized environments (green).



Figure 12: **Heterogeneous Dual-Arm Control via Object-Centric Manipulation.**

## K Full RoboTwin 2.0 Benchmark

We report the evaluation results of five policies on the RoboTwin 2.0 benchmark under the *Easy* and *Hard* settings. Note that these two settings differ only in evaluation conditions, while the training setup remains identical. A continuously maintained online leaderboard is available at <https://robotwin-platform.github.io/leaderboard>.

Table 10: RoboTwin 2.0 Simulation Benchmark (clean vs randomized, 50+ tasks).

Simulation Task	RDT		Pi0		ACT		DP		DP3	
	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
<i>Adjust Bottle</i>	81%	<b>75%</b>	90%	56%	97%	23%	97%	0%	<b>99%</b>	3%
<i>Beat Block Hammer</i>	<b>77%</b>	<b>37%</b>	43%	21%	56%	3%	42%	0%	72%	8%
<i>Blocks Ranking RGB</i>	3%	0%	<b>19%</b>	<b>5%</b>	1%	0%	0%	0%	3%	0%
<i>Blocks Ranking Size</i>	0%	0%	<b>7%</b>	<b>1%</b>	0%	0%	1%	0%	2%	0%
<i>Click Alarmclock</i>	61%	12%	63%	11%	32%	4%	61%	5%	<b>77%</b>	<b>14%</b>
<i>Click Bell</i>	80%	<b>9%</b>	44%	3%	58%	3%	54%	0%	<b>90%</b>	0%
<i>Dump Bin Bigbin</i>	64%	32%	83%	24%	68%	1%	49%	0%	<b>85%</b>	<b>53%</b>
<i>Grab Roller</i>	74%	43%	96%	<b>80%</b>	94%	25%	<b>98%</b>	0%	<b>98%</b>	2%
<i>Handover Block</i>	45%	<b>14%</b>	45%	8%	42%	0%	10%	0%	<b>70%</b>	0%
<i>Handover Mic</i>	90%	<b>31%</b>	98%	13%	85%	0%	53%	0%	<b>100%</b>	3%
<i>Hanging Mug</i>	<b>23%</b>	<b>16%</b>	11%	3%	7%	0%	8%	0%	17%	1%
<i>Lift Pot</i>	72%	9%	84%	<b>36%</b>	88%	0%	39%	0%	<b>97%</b>	0%
<i>Move Can Pot</i>	25%	12%	58%	<b>21%</b>	22%	4%	39%	0%	<b>70%</b>	6%
<i>Move Pillbottle Pad</i>	8%	0%	21%	<b>1%</b>	0%	0%	1%	0%	<b>41%</b>	0%
<i>Move Playingcard Away</i>	43%	11%	53%	<b>22%</b>	36%	0%	47%	0%	<b>68%</b>	3%
<i>Move Stapler Pad</i>	2%	0%	0%	<b>2%</b>	0%	0%	1%	0%	<b>12%</b>	0%
<i>Open Laptop</i>	59%	32%	<b>85%</b>	<b>46%</b>	56%	0%	49%	0%	82%	7%
<i>Open Microwave</i>	37%	20%	80%	<b>50%</b>	<b>86%</b>	0%	5%	0%	61%	22%
<i>Pick Diverse Bottles</i>	2%	0%	27%	<b>6%</b>	7%	0%	6%	0%	<b>52%</b>	1%
<i>Pick Dual Bottles</i>	42%	<b>13%</b>	57%	12%	31%	0%	24%	0%	<b>60%</b>	1%
<i>Place A2B Left</i>	3%	1%	31%	1%	1%	0%	2%	0%	<b>46%</b>	<b>2%</b>
<i>Place A2B Right</i>	1%	1%	27%	<b>6%</b>	0%	0%	13%	0%	<b>49%</b>	0%
<i>Place Bread Basket</i>	10%	2%	17%	<b>4%</b>	6%	0%	14%	0%	<b>26%</b>	1%
<i>Place Bread Skillet</i>	5%	<b>1%</b>	<b>23%</b>	<b>1%</b>	7%	0%	11%	0%	19%	0%
<i>Place Burger Fries</i>	50%	<b>27%</b>	<b>80%</b>	4%	49%	0%	72%	0%	72%	18%
<i>Place Can Basket</i>	19%	<b>6%</b>	41%	5%	1%	0%	18%	0%	<b>67%</b>	2%
<i>Place Cans Plasticbox</i>	6%	<b>5%</b>	34%	2%	16%	0%	40%	0%	<b>48%</b>	3%
<i>Place Container Plate</i>	78%	17%	<b>88%</b>	<b>45%</b>	72%	1%	41%	0%	86%	1%
<i>Place Dual Shoes</i>	4%	<b>4%</b>	<b>15%</b>	0%	9%	0%	8%	0%	13%	0%
<i>Place Empty Cup</i>	56%	7%	37%	<b>11%</b>	61%	0%	37%	0%	<b>65%</b>	1%
<i>Place Fan</i>	12%	2%	20%	<b>10%</b>	1%	0%	3%	0%	<b>36%</b>	1%
<i>Place Mouse Pad</i>	1%	0%	<b>7%</b>	<b>1%</b>	0%	0%	0%	0%	4%	<b>1%</b>
<i>Place Object Basket</i>	33%	<b>17%</b>	16%	2%	15%	0%	15%	0%	<b>65%</b>	0%
<i>Place Object Scale</i>	1%	<b>0%</b>	<b>10%</b>	<b>0%</b>	0%	<b>0%</b>	1%	<b>0%</b>	15%	<b>0%</b>
<i>Place Object Stand</i>	15%	5%	36%	<b>11%</b>	1%	0%	22%	0%	<b>60%</b>	0%
<i>Place Phone Stand</i>	15%	6%	35%	<b>7%</b>	2%	0%	13%	0%	<b>44%</b>	2%
<i>Place Shoe</i>	35%	<b>7%</b>	28%	6%	5%	0%	23%	0%	<b>58%</b>	2%
<i>Press Stapler</i>	41%	24%	62%	<b>29%</b>	31%	6%	6%	0%	<b>69%</b>	3%
<i>Put Bottles Dustbin</i>	21%	4%	54%	13%	27%	1%	22%	0%	<b>60%</b>	<b>21%</b>
<i>Put Object Cabinet</i>	33%	<b>18%</b>	68%	<b>18%</b>	15%	0%	42%	0%	72%	1%
<i>Rotate QRcode</i>	50%	5%	68%	<b>15%</b>	1%	0%	13%	0%	<b>74%</b>	1%
<i>Scan Object</i>	4%	<b>1%</b>	18%	<b>1%</b>	2%	0%	9%	0%	<b>31%</b>	<b>1%</b>
<i>Shake Bottle Horizontally</i>	84%	<b>51%</b>	99%	<b>51%</b>	63%	4%	59%	18%	<b>100%</b>	25%
<i>Shake Bottle</i>	74%	45%	97%	<b>60%</b>	74%	10%	65%	8%	<b>98%</b>	19%
<i>Stack Blocks Three</i>	2%	<b>0%</b>	<b>17%</b>	<b>0%</b>	0%	<b>0%</b>	0%	<b>0%</b>	1%	<b>0%</b>
<i>Stack Blocks Two</i>	21%	<b>2%</b>	<b>42%</b>	1%	25%	0%	7%	0%	24%	0%
<i>Stack Bowls Three</i>	51%	17%	<b>66%</b>	<b>24%</b>	48%	0%	63%	0%	57%	5%
<i>Stack Bowls Two</i>	76%	30%	<b>91%</b>	<b>41%</b>	82%	0%	61%	0%	83%	6%
<i>Stamp Seal</i>	1%	0%	3%	<b>4%</b>	2%	0%	2%	0%	<b>18%</b>	0%
<i>Turn Switch</i>	35%	15%	27%	<b>23%</b>	5%	2%	36%	1%	<b>46%</b>	8%
<i>Average (%)</i>	34.5	13.7	46.4	<b>16.3</b>	29.7	1.7	28.0	0.6	<b>55.2</b>	5.0

## L Success Rates of Different Embodiments on RoboTwin 2.0 Tasks

Table 11 reports the success rates of five robot embodiments across the 50 RoboTwin 2.0 tasks, using the same set of expert programs for data generation.

Table 11: Success Rates of Different Embodiments on RoboTwin 2.0 Tasks.

Task Name	RoboTwin1.0					RoboTwin2.0				
	Aloha	ARX	Franka	Piper	UR5	Aloha	ARX	Franka	Piper	UR5
Adjust Bottle	92%	88%	39%	0%	7%	93%	94%	34%	0%	12%
Beat Block Hammer	68%	86%	95%	0%	86%	64%	93%	98%	15%	90%
Blocks Ranking Rgb	92%	98%	96%	0%	82%	96%	97%	99%	13%	53%
Blocks Ranking Size	90%	95%	92%	0%	60%	96%	97%	89%	7%	38%
Click Alarmclock	89%	99%	100%	0%	95%	92%	99%	100%	0%	95%
Click Bell	100%	100%	100%	9%	100%	100%	100%	100%	91%	100%
Dump Bin Bigbin	85%	98%	90%	0%	82%	84%	100%	84%	9%	80%
Grab Roller	95%	69%	99%	0%	80%	95%	69%	99%	7%	81%
Handover Block	1%	3%	0%	0%	4%	83%	81%	0%	44%	0%
Handover Mic	62%	80%	92%	28%	0%	87%	98%	84%	65%	14%
Hanging Mug	68%	76%	5%	0%	12%	63%	73%	11%	0%	11%
Lift Pot	27%	50%	24%	5%	40%	27%	50%	36%	31%	40%
Move Can Pot	18%	0%	37%	2%	4%	93%	65%	92%	96%	99%
Move Pillbottle Pad	30%	52%	15%	0%	35%	67%	90%	69%	47%	86%
Move Playingcard Away	93%	100%	100%	0%	87%	99%	100%	100%	63%	66%
Move Stapler Pad	94%	92%	88%	0%	95%	92%	96%	89%	13%	75%
Open Laptop	76%	91%	78%	14%	55%	82%	92%	77%	23%	51%
Open Microwave	65%	85%	75%	5%	33%	96%	80%	59%	2%	23%
Pick Diverse Bottles	11%	1%	0%	0%	0%	51%	2%	0%	27%	4%
Pick Dual Bottles	8%	3%	0%	0%	0%	92%	6%	0%	81%	7%
Place A2B Left	65%	75%	70%	0%	72%	80%	88%	64%	29%	76%
Place A2B Right	70%	68%	68%	0%	69%	81%	82%	64%	31%	66%
Place Bread Basket	91%	91%	69%	0%	78%	89%	88%	62%	1%	67%
Place Bread Skillet	31%	28%	42%	0%	42%	34%	26%	42%	0%	37%
Place Can Basket	47%	1%	38%	0%	11%	70%	28%	61%	0%	3%
Place Cans Plasticbox	96%	93%	98%	0%	11%	100%	96%	85%	0%	82%
Place Container Plate	86%	85%	83%	0%	82%	89%	86%	86%	37%	81%
Place Dual Shoes	73%	28%	36%	0%	40%	77%	31%	41%	1%	32%
Place Empty Cup	92%	100%	100%	0%	100%	92%	100%	100%	4%	100%
Place Fan	93%	96%	75%	0%	85%	95%	93%	83%	0%	65%
Place Burger Fries	96%	95%	85%	0%	78%	97%	98%	80%	36%	74%
Place Mouse Pad	100%	80%	99%	2%	96%	99%	89%	100%	23%	73%
Place Object Basket	68%	13%	68%	0%	30%	74%	14%	61%	0%	7%
Place Object Scale	77%	93%	94%	0%	87%	78%	92%	82%	2%	76%
Place Object Stand	90%	92%	81%	0%	90%	97%	99%	81%	9%	92%
Place Phone Stand	66%	78%	52%	22%	44%	66%	78%	45%	53%	49%
Place Shoe	87%	85%	70%	0%	97%	84%	85%	74%	7%	91%
Press Stapler	87%	96%	99%	0%	77%	98%	96%	100%	59%	72%
Put Bottles Dustbin	0%	0%	0%	0%	0%	71%	1%	0%	56%	0%
Put Object Cabinet	13%	56%	43%	0%	0%	14%	24%	55%	0%	0%
Rotate Qrcode	78%	83%	98%	0%	81%	75%	74%	94%	0%	67%
Scan Object	8%	13%	21%	0%	8%	4%	45%	26%	0%	19%
Shake Bottle	62%	95%	82%	1%	98%	89%	94%	85%	74%	97%
Shake Bottle Horizontally	64%	93%	81%	1%	97%	90%	94%	85%	74%	98%
Stack Blocks Three	98%	97%	95%	0%	83%	94%	96%	80%	0%	51%
Stack Blocks Two	99%	99%	100%	0%	94%	98%	99%	96%	2%	68%
Stack Bowls Three	27%	64%	76%	0%	76%	43%	58%	82%	0%	81%
Stack Bowls Two	63%	84%	88%	0%	94%	78%	82%	88%	4%	94%
Stamp Seal	46%	91%	95%	0%	100%	56%	91%	4%	37%	100%
Turn Switch	27%	3%	51%	28%	10%	74%	3%	36%	81%	10%
<b>Average</b>	<b>65.3%</b>	<b>68.8%</b>	<b>67.6%</b>	<b>2.3%</b>	<b>57.7%</b>	<b>78.8%</b>	<b>74.2%</b>	<b>67.2%</b>	<b>25.1%</b>	<b>57.1%</b>
<b>Difference</b>	/	/	/	/	/	+13.5%	+5.4%	-0.4%	+22.8%	-0.6%