# RealMirror: A Comprehensive, Open-Source Vision-Language-Action Platform for Embodied AI

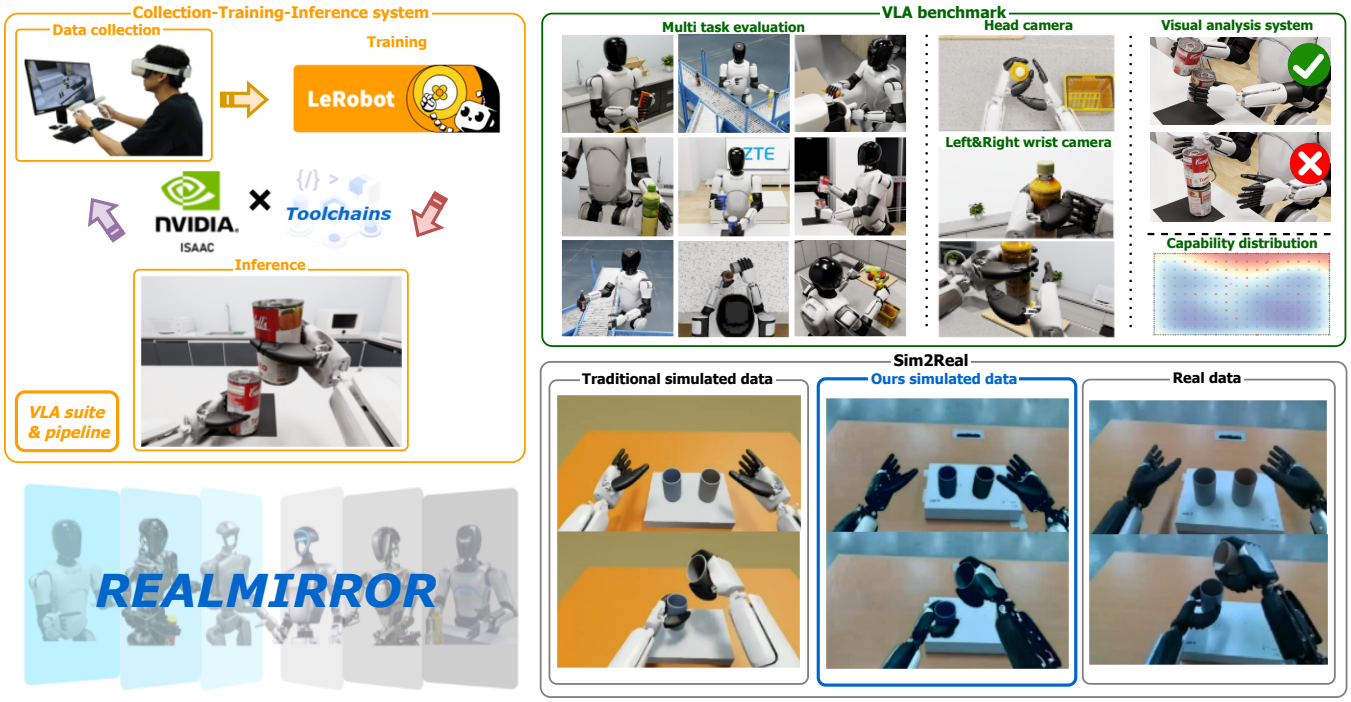


Fig. 1: The RealMirror framework for accelerating VLA research in humanoid robots. (a) Top-left: Our data collection, training, and inference system is developed based on VR teleoperation, Lerobot, and Isaac Sim. (b) Top-right: This platform populates a benchmark for humanoid robots with multiple scenarios and various VLA models. (c) Bottom-right: The striking photorealism of the simulation bridges the reality gap, enabling zero-shot Sim2Real transfer without any fine-tuning.

***Abstract*— The emerging field of Vision-Language-Action (VLA) for humanoid robots faces several fundamental challenges, including the high cost of data acquisition, the lack of a standardized benchmark, and the significant gap between simulation and the real world. To overcome these obstacles, we propose RealMirror, a comprehensive, open-source embodied AI VLA platform. RealMirror builds an efficient, low-cost data collection, model training, and inference system that enables end-to-end VLA research without requiring a real robot. To facilitate model evolution and fair comparison, we also introduce a dedicated VLA benchmark for humanoid robots, featuring multiple scenarios, extensive trajectories, and various VLA models. Furthermore, by integrating generative models and 3D Gaussian Splatting to reconstruct realistic environments and robot models, we successfully demonstrate zero-shot Sim2Real transfer, where models trained exclusively on simulation data can perform tasks on a real robot seamlessly, without any fine-tuning. In conclusion, with the unification of these critical components, RealMirror provides a robust framework that significantly accelerates the development of VLA models for humanoid robots. Project page: https://terminators2025.github.io/RealMirror.github.io**

## I. INTRODUCTION

The rapid evolution of Large Language Models (LLMs) like GPT [1], Qwen [2], and Deepseek [3] has significantly advanced the development of Artificial General Intelligence (AGI). While exhibiting remarkable model performance, they lack the ability to perform tasks in the real world. The vision of embodied AI can overcome this limitation by creating intelligent agents capable of perceiving, understanding, and physically interacting with the real world. The latest developments in humanoid robots and Vision-Language-Action (VLA) models are making this vision possible [4], [5], [6].

However, there are still a series of profound challenges that need to be addressed. First and foremost, the acquisition of high-quality interactive data remains an immense and costly bottleneck. Unlike large language models that can leverage vast internet datasets, embodied AI requires data generated from real robot interactions. This process is inherently time-consuming, expensive, and sometimes dangerous. Despite the availability of open-source robot datasets [7], [8], [9], their offline nature prevents them from supporting the interactive iteration and validation necessary for embodied AI. Simulation platforms offer a promising alternative to address this data bottleneck, but existing platforms [10], [11], [12], [13] are primarily designed for robotic arms and grippers, lacking support for complex systems such as humanoid

robots and dexterous hands. Furthermore, even platforms that support humanoid robots [14] often suffer from insufficient environmental realism, which prevents a seamless Sim2Real transfer. This "reality gap" often results in suboptimal performance when models trained in simulation are deployed on real robots. Finally, the absence of a unified open-source humanoid robot benchmark for objectively evaluating and comparing model performance presents a significant obstacle in VLA, hindering systematic research.

To address these issues, we propose RealMirror, a comprehensive, open-source embodied AI VLA platform, as shown in Fig. 1. Firstly, to tackle the data acquisition and interactive validation bottleneck, we build an efficient, low-cost data collection, model training, and model inference system. We optimized the teleoperation and communication frameworks, which, compared to the general communication framework [15], significantly enhanced the real-time performance and efficiency of data collection. When integrated into RealMirror, this enables end-to-end VLA research without the need for a real robot. Secondly, we propose a dedicated VLA benchmark for humanoid robots to accelerate algorithm research and fair comparison. This benchmark includes five distinct scenarios and over 1,000 robot trajectories, designed to evaluate a suite of core competencies, from fundamental manipulations to dual-arm collaboration. Additionally, we conduct extensive automated evaluations on a variety of representative VLA models [4], [5], [6]. Finally, to bridge the Sim2Real gap, we employ generative models [16] to create high-fidelity 3D assets from real-world images, and integrate 3D Gaussian Splatting (3DGS) [17] to reconstruct realistic environments and controllable robot models from video. Without any fine-tuning on real-world data, the model trained solely on simulation data can complete tasks seamlessly on a real robot.

In summary, our platform RealMirror enables researchers to perform data collection, model training, model inference, and performance evaluation in a unified system, thereby accelerating the development of VLA for humanoid robots. Our contributions are as follows:

1) **We build an efficient, low-cost data collection, model training, and model inference system** that enables end-to-end VLA research without requiring a real robot.

2) **We propose a dedicated VLA benchmark for humanoid robots** that facilitates model evolution and fair comparison through extensive experiments and automated evaluation across multiple scenarios and various VLA models.

3) **We demonstrate the feasibility of zero-shot Sim2Real** by integrating generative models and 3DGS to reconstruct realistic environments and robot models, thus enabling models trained solely on simulation data to perform tasks seamlessly on a real robot without any fine-tuning.

## II. RELATED WORK

### A. Vision-Language-Action

Foundational works in this domain include RT-1 [18], a pioneering work that applied the Transformer architecture to robot control, and ACT [4], which employs a Variational Autoencoder for imitation learning. The development of RT-2 [19] and OpenVLA [20] further advanced this research by fine-tuning large-scale Vision-Language Models (VLMs), enabling robots to leverage web-scale knowledge. Notably, Diffusion Policy [5] introduced a policy learning method based on diffusion models, which is effective at handling high-dimensional and complex actions.

Recent advancements have spurred the adoption of hierarchical [21], [22] and dual-system architectures [23], [24], [6]. The Hi Robot framework [21], for instance, leverages a pre-trained VLM as a high-level planner. Concurrently, a new wave of models—including PI0 [23], GROOT N1 [24] and SmolVLA [6]—employs a dual-system architecture that combines a VLM for interpreting environments and instructions with a diffusion-based Transformer action expert for real-time action generation, representing the current frontier in embodied AI research.

### B. Sim2Real

Simulation has become an indispensable tool for training embodied agents like VLA models, as it circumvents the cost, inefficiency, and safety risks inherent to real-world data collection [25], [26]. This has spurred the development of numerous simulators to support large-scale robot learning [27], [11], [12]. However, the utility of simulation data is fundamentally limited by the Sim2Real gap [25], [28]. Consequently, bridging this gap has been a central theme in robotics research, with significant progress in domains such as legged locomotion [29], [30], autonomous driving [31], and dexterous manipulation [32], [33].

The frontier of this research is zero-shot Sim2Real transfer, where policies trained entirely in simulation are deployed directly to hardware without fine-tuning. While recent works have demonstrated its feasibility for tasks like locomotion and mobile manipulation [34], [35], a critical gap persists for general-purpose, high-dimensional visuomotor policies. Humanoid robots equipped with dexterous hands represent a pinnacle of this challenge, demanding nuanced, whole-body control for contact-rich tasks. Our work directly addresses this challenge by introducing RealMirror, the first platform and benchmark designed to facilitate and evaluate zero-shot Sim2Real transfer for dexterous humanoid VLA policies, thereby catalyzing reproducible research in this ambitious domain of embodied AI.

### C. Simulation Platform

Contemporary embodied robots are in urgent need of an interactive, high-fidelity simulation platform. Rcare world [36] is a high-fidelity, human-centric robotic caregiving simulation environment built with Unity. The ManiSkill series of works [12], [11] focuses on manipulation skills over diverse objects in a full-physics simulator. In a similar vein, a benchmark for robotic learning is presented in Rlbench [13], which features a set of pre-defined tasks. Furthermore, Behavior-1k [37] and AgentWorld [14] increase scene complexity by simulating human-like activities. Gaussian Splat-

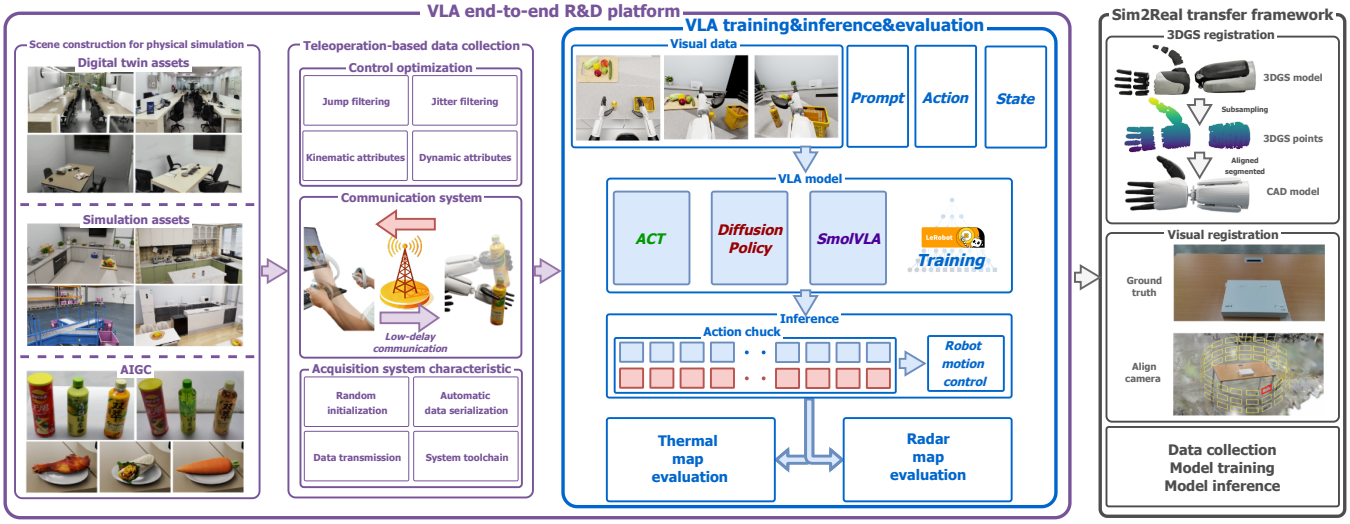


Fig. 2: An overview of the integrated RealMirror pipeline. The platform offers an end-to-end solution for VLA research, encompassing asset acquisition and scene construction, as well as optimized teleoperation-based data collection and the training and evaluation of multiple VLA models. To bridge the critical reality gap, RealMirror incorporates a zero-shot Sim2Real module that uses 3D Gaussian Splatting to create a high-fidelity digital twin of the robot and the environment. This enables policies trained purely in simulation to be deployed directly to the real world.

ting is also employed by Splatsim [34] to render real-world environments. This approach is a significant step towards bridging the Sim2Real gap. The availability of an open and user-friendly teleoperation platform for data collection is also of paramount importance. AgentWorld [14] implements human teleoperation based on Isaac Sim [38]. Although these existing works have designed and realized high-fidelity, interactive simulation platforms, RealMirror distinguishes itself by demonstrating zero-shot generalization of dexterous hand skills to real-world scenarios, thereby validating the efficacy of data collection to real-world deployment.

## III. RealMirror

RealMirror is a comprehensive, open-source embodied AI platform designed for humanoid VLA research. It integrates an efficient, low-cost system for data collection, model training, and inference, enabling end-to-end VLA development without requiring a real robot. To facilitate model evolution, RealMirror provides a dedicated benchmark consisting of over 1,000 high-quality simulation trajectories across multiple tasks and humanoid robot platforms. Finally, the platform incorporates a Sim2Real transfer framework, leveraging generative models [16] and 3DGS [44], [45] to reconstruct realistic environments and robot models, thereby enabling zero-shot transfer from simulation to real-world execution.

Table I summarizes the key features of RealMirror compared with existing robotic simulation platforms, and Fig. 2 provides an overview of the platform architecture.

### A. Scene Construction for Physical Simulation

To support diverse embodied AI tasks, we construct simulation environments based on the NVIDIA Isaac Sim platform [38]. Specifically, we design a wide range of indoor scenes that incorporate complex layouts, multiple manipulable objects, and realistic physical interactions. By integrating CAD models and assets from various asset libraries [38], [46], [47], RealMirror enables the creation of customizable environments with varying difficulty levels and task requirements. In addition, we assign appropriate physical properties (e.g., mass, friction, and collision parameters) to the assets, ensuring their plausibility in simulation and compatibility with humanoid robot embodiments. These simulation-ready environments serve as the foundation for large-scale data collection, training, and evaluation in RealMirror.

### B. Efficient System for Data Collection, Training, and Inference

We have developed an efficient, end-to-end system for VLA data collection, training, and inference. By leveraging a teleoperation system, the platform enables high-quality data acquisition, while the unified training and inference framework supports multiple state-of-the-art VLA models. Deep integration with Isaac Sim allows for closed-loop evaluation, ensuring both efficiency and reproducibility. Overall, this system significantly reduces the cost and complexity of data collection and provides a reliable foundation for subsequent VLA research and model benchmarking.

*1) Teleoperation-Based Data Collection:* We develop a teleoperation-based data collection system to efficiently gather high-quality training data for humanoid VLA tasks. The system consists of two main components. First, we implement a motion control pipeline in the simulation environment with multi-level filtering mechanisms, including: IK joint control jump filtering, end-effector pose communication and drift compensation, IK solver threshold filtering, and cross-frame end-effector pose threshold filtering. These layers of filtering ensure smooth and physically plausible motion during teleoperation. Second, we build a lightweight

| Name | Data Collection | | | Platform Capabilities | |
|---|---|---|---|---|---|
| | Tele-operation | Dexterous-hand | Num of Trajectories | End-to-end Framework | Zero-shot Sim2Real |
| Maniskill2[12] | ✗ | ✗ | 30k | ✗ | ✗ |
| RLBench[39] | ✗ | ✗ | – | ✗ | ✗ |
| BiGym[40] | ✓ | ✗ | >2000 | ✗ | ✗ |
| Behavior-1K[41] | ✗ | ✗ | – | ✗ | ✗ |
| MimicGen[42] | ✗ | ✗ | 50k | ✗ | ✗ |
| GRUtopia[43] | ✓ | ✗ | – | ✗ | ✗ |
| AgentWorld[14] | ✓ | ✓ | >1000 | ✗ | ✗ |
| RealMirror (Ours) | ✓ | ✓ | >1000 | ✓ | ✓ |

TABLE I: Comparison of robotic simulation platforms in terms of data collection methods and platform-level capabilities. RealMirror distinguishes itself by providing a complete, end-to-end framework that supports the full VLA research lifecycle, from data collection to direct real-world deployment.

WebXR-based communication system for real-time control. By streamlining the data transmission protocol, our system operates at a 90 Hz transmission frequency and achieves a 114ms reduction in end-to-end latency from teleoperation command to robot execution, compared to the general communication framework [15].

Together, these components enable efficient data acquisition. For example, a single-arm pick-and-place task averages 7.83 seconds per trajectory, encompassing the full end-to-end workflow: environment initialization, object manipulation, and data packaging for subsequent training.

*2) Unified Training and Inference Framework:* Our framework provides a unified training and inference system for VLA research. On the training side, we support multiple representative VLA models, including ACT [4], Diffusion Policy [5], and SmolVLA [6]. In addition, all our algorithms are adapted to incorporate a temporal ensembling mechanism to enhance action prediction robustness and reduce compounding errors. To ensure a fair and efficient comparison, we build upon the LeRobot library [48] and extend its functionalities to support humanoid robot embodiments. On the inference side, the trained models are integrated with Isaac Sim [38] for interactive evaluation. During inference, the system continuously receives multi-modal inputs, including BGR observations, robot proprioceptive states, and natural language instructions. The VLA model then predicts the corresponding action sequences, which are published back to the simulator for real-time execution. This closed-loop integration enables seamless evaluation of different models under consistent conditions, thereby facilitating fair benchmarking and systematic analysis.

### C. Benchmark for Humanoid VLA

We introduce a systematic benchmark for humanoid robot VLA research. The benchmark provides a standardized platform for training, evaluating, and comparing VLA algorithms across a wide range of tasks and scenarios. It comprises over 1,000 high-quality simulated trajectories, capturing diverse interactions, manipulation challenges, and dynamic environments. By combining task diversity, multi-modal inputs, and a unified evaluation pipeline, this benchmark enables systematic, reproducible, and rigorous research on humanoid VLA.

*1) Task Scenarios and Dataset:* We construct a high-quality humanoid VLA dataset comprising five task scenarios, each with 240 trajectories, totaling over 1,000 simulated trajectories. The task scenarios are designed to cover a broad range of skills, including *Pick and Place*, *Dual-arm collaboration*, *Push and Pull*, *Dynamic grasping*, and *Precision control*. Table II presents the distribution of demonstration trajectories in the benchmark dataset for the humanoid robot. Below, we describe each task scenario:

- ***Kitchen Cleanup***: Use the left hand to pick up chips, green tea, or lemon tea from the table, then transfer the item to the right hand and place it into the basket.
- ***Air Fryer Manipulation***: Use the left hand to lift chicken rolls, chicken legs, or carrots from a plate, then use the right hand to open the air fryer and place the food inside.
- ***Assembly Line Sorting***: Sort three types of items (oil, cola, and Sprite) on the conveyor belt, ensuring that each sorted item lands correctly in its designated box.
- ***Cup-to-Cup Transfer***: Pour berries from the cup on the right into the cup on the left.
- ***Can Stacking***: Stack cans from both sides into the center and ensure they are placed stably.

*2) Supported VLA Models:* We train several representative VLA algorithms on our benchmark, covering different design philosophies: **ACT** [4] is based on a VAE architecture for VLA modeling, emphasizing fine-grained bimanual manipulation. **Diffusion Policy** [5] predicts actions via a generative diffusion process, offering robust visuomotor control. **SmolVLA** [6] employs a dual-system architecture, where S2 VLM extracts visual and textual information and S1 Action Expert predicts corresponding actions, enabling broad generalization.

*3) Evaluation Metrics:* Evaluation metrics focus on task success rate, which provides a clear and consistent measure of VLA performance across different tasks and scenarios. We provide automated evaluation tools to ensure objective, reproducible, and fair assessment of all models, enabling

TABLE II: Mapping of benchmark tasks to the core skills they assess, and the corresponding number of demonstration trajectories.

| Task | Assessed Skills | Num of Traj. |
|---|---|---|
| Kitchen Cleanup | - Pick and Place<br>- Dual-arm collaboration | 240 |
| Air Fryer Manipulation | - Pick and Place<br>- Push and Pull<br>- Dual-arm collaboration | 240 |
| Assembly Line Sorting | - Pick and Place<br>- Dual-arm collaboration<br>- Dynamic grasping | 240 |
| Cup-to-Cup Transfer | - Dual-arm collaboration<br>- Precision control | 240 |
| Can Stacking | - Pick and Place<br>- Precision control | 240 |
| **Total** | | **1200** |

standardized benchmarking for future research. Simultaneously, to intuitively analyze how the capabilities of different models are distributed across the same scenario, we developed a heatmap analysis tool, as depicted in Fig. 1.

### D. Sim2Real Transfer Framework

While Isaac Sim provides a physically robust foundation, bridging the visual gap between simulation and reality is crucial for effective policy transfer. We employ a multi-pronged strategy that ensures the VLA model observes a simulation environment virtually indistinguishable from the real world. By integrating high-fidelity background reconstruction, photorealistic robot modeling, and differentiated interactive object generation, our pipeline enables zero-shot transfer of trained VLA policies to real-world deployment.

*1) Visual Augmentation for Realism:* We adopt a hybrid generative approach to enhance visual realism. Different scene components, including the static background, the robot, and interactive objects, are treated with specialized techniques to ensure high-fidelity perception.

**Static Environment Rendering with 3DGS:** We capture the target real-world workspace from multiple viewpoints and reconstruct the entire static scene using 3DGS. This high-fidelity background, including lab benches, walls, and distant clutter, is integrated into the simulator as a non-interactive canvas, preserving subtle lighting and material effects that are difficult to reproduce with standard rendering.

**High-Fidelity Articulated Robot Model:** We reconstruct the physical humanoid robot with 3DGS and segment it into individual links. Each link is rigidly aligned to its corresponding USD model in Isaac Sim using a scale $S$, rotation $R$, and translation $T$ transformation:

$$P_{\text{USD}} = S \cdot R \cdot P_{\text{3DGS}} + T \tag{1}$$

This process overlays a photorealistic "skin" onto the kinematically and physically accurate robot skeleton, ensuring visual fidelity in simulation.

*2) Differentiated Strategy for Interactive Objects:* Interactive objects are treated according to their physical and visual requirements:

**High-Precision Objects:** Objects requiring precise contact physics, such as dexterous end-effectors or flat tabletops, use a digital twin approach. A clean CAD model governs all dynamics and collisions, while a 3DGS reconstruction provides visual realism aligned to the CAD model.

**Low-Precision Objects:** Objects with lower physics requirements prioritize visual diversity. Using few-shot 3D generative models, we produce textured 3D meshes from a few images, which serve as both visual and collision representations, enabling rapid scene population with diverse assets.

*3) Coordinate Alignment and Camera Calibration:* We align the coordinate systems of Isaac Sim, the 3DGS reconstructed scene, and the robot's real-world pose to ensure consistency between the physical and simulated environments. First, we employ the Iterative Closest Point (ICP) algorithm to align the CAD assets in Isaac Sim with the 3DGS reconstructed environment. For camera calibration, we record the robot's observed images $I_{\text{robot}}$ during task execution and solve for the camera pose using Structure-from-Motion (SfM) together with a set of real-world images. Because the Isaac Sim coordinates are already aligned with the 3DGS scene, placing a virtual camera at the solved $I_{\text{robot}}$ pose within Isaac Sim achieves accurate calibration between the real and simulated cameras.

In summary, our Sim2Real transfer framework combines high-fidelity visual augmentation, precise robot modeling, and differentiated interactive object treatment to bridge the gap between simulation and reality. With coordinate alignment and camera calibration, trained VLA policies can be directly deployed on real robots without additional fine-tuning. This approach provides a robust foundation for efficient and scalable humanoid VLA research, ensuring that learned policies are both transferable and reliable across diverse scenarios.

## IV. EXPERIMENTS

The experiment consists of two main phases. First, we established a VLA benchmark by training and automatically evaluating various VLA models on our platform to compare their performance. Second, our Sim2Real experiments assess the effectiveness of models trained on simulation data when deployed on a real robot.

### A. VLA Benchmark

*1) Experimental Setup:* Our datasets were collected using a PICO Neo3 Pro headset and a workstation with Ada5880. To comprehensively evaluate our benchmark, we selected three representative VLA models for the experiments: ACT [4], Diffusion Policy [5], and SmolVLA [6]. All models process synchronized multi-view BGR images, while SmolVLA additionally utilizes natural language task descriptions. The unified action space is 26-dimensional, with 13 dimensions for each of the two robotic arms (7 for the arm and 6 for the hand). For training, each model was trained for 100,000 steps with a batch size of 16. Additionally, a Temporal Ensembler Mechanism was adopted to enhance the smoothness of actions.

TABLE III: Comparison of task success rates (%) for each model. The best performance in each task is highlighted in **bold**.

| Task | ACT [4] | Diffusion Policy [5] | SmolVLA [6] |
|---|---|---|---|
| Kitchen Cleanup | **100.00** | 99.00 | 99.75 |
| Air Fryer Manipulation | 77.75 | **85.50** | 83.00 |
| Assembly Line Sorting | **95.00** | 88.00 | 86.00 |
| Cup-to-Cup Transfer | 55.50 | 63.50 | **68.00** |
| Can Stacking | 39.50 | 39.75 | **62.00** |
| **Avg** | 73.55 | 75.15 | **79.75** |

*2) Automatic Evaluation Protocol:* The performance of models was quantitatively evaluated by their success rate across five distinct manipulation scenarios. The criteria for each scenario were defined as follows.

- ***Kitchen Cleanup*** (400 trials): The task is successful when a specified item is picked up and placed into the designated basket by coordinating both of its arms.
- ***Air Fryer Manipulation*** (400 trials): A successful trial involves the robot correctly opening the air fryer drawer, placing a food item inside, and then closing the drawer.
- ***Can Stacking*** (400 trials): This scenario requires the robot to grasp two cans on the desk and stably stack them employing a bimanual manipulation approach.
- ***Cup-to-Cup Transfer*** (200 trials): A trial is deemed successful upon the transfer of a berry from the right cup to the left cup, performed while the cups are lifted in the air.
- ***Assembly Line Sorting*** (100 trials): The criterion for successful completion is the correct sorting of three consecutive items from a conveyor belt in a single trial.

*3) Experimental Results:* The evaluation results for the three models are summarized in Table III, showing their success rates across the five scenarios.

From a task-centric perspective, the results show that while SmolVLA achieved the highest average success rate (79.75%), its primary advantage was in high-precision scenarios like *Cup-to-Cup Transfer* (68.00%) and *Can Stacking* (62.00%). In contrast, ACT demonstrated near-perfect performance in *Kitchen Cleanup* (100.00%) and the dynamic *Assembly Line Sorting* (95.00%), while Diffusion Policy's strength was most apparent in *Air Fryer Manipulation* (85.50%). This comparative analysis reveals that while all three models possess domain-specific strengths, SmolVLA achieved the most robust performance across the evaluated tasks.

In addition to evaluating the models based on task completion success rates, our benchmark also analyzed their performance from a skill-based perspective. For this, we abstracted five core robotic skills: Pick and Place, Dual-arm collaboration, Push and Pull, Dynamic grasping, and Precision control. This approach provides a finer-grained understanding of each model, as shown in Fig. 3. The analysis reveals that SmolVLA exhibited the most balanced performance, achieving the highest average success rates in both Pick and Place and Precision Control. The ACT model showed a significant advantage in Dynamic Grasping, with a 95% success rate that was notably higher than the other two models. Meanwhile, Diffusion Policy performed best in the Push and Pull skill. Several qualitative results for the different VLA algorithms are shown in Fig. 4.

This two-tiered evaluation framework, assessing performance at both the task and skill levels, is the cornerstone of the RealMirror benchmark's utility. It allows us to directly correlate a model's success in a complex task with its proficiency in underlying robotic skills. By providing this deeper, diagnostic insight, our benchmark empowers the community to move beyond simple leaderboards. It offers a systematic foundation for comparing VLA models, identifying their specific architectural trade-offs, and ultimately guiding future research toward targeted algorithmic improvements.

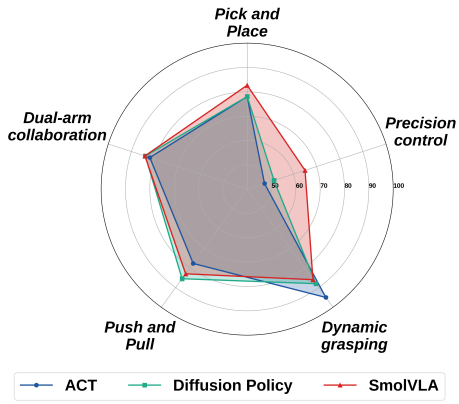


Fig. 3: Model performance comparison across different robotic skills.

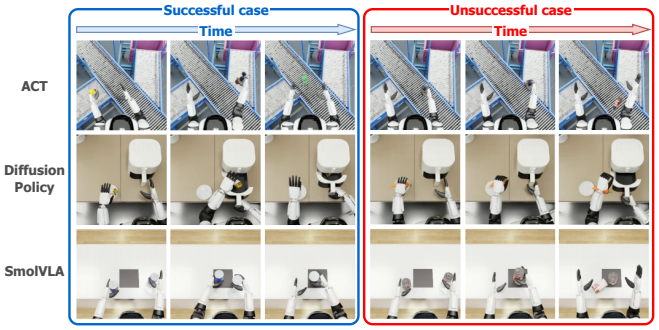


Fig. 4: Qualitative results for different VLA algorithms in our benchmark.

## B. Sim2Real Experiments

Traditional robotic simulations suffer from a substantial Sim2Real gap, primarily in visual fidelity. This gap stems from discrepancies in lighting, material properties, textures, and geometry between the simulated and real worlds, which severely restricts the direct transferability of models trained on simulation data to real robots. To address this challenge, we integrated generative models and 3DGS to reconstruct highly realistic environments and robot models. The real-world experiments were implemented with a ZHIYUAN A2 robot. As illustrated in Fig. 5, our reconstructed scenes are visually more realistic than those generated by traditional simulations and closely approximate real-world visual fidelity.

This visual realism provides a robust foundation for zero-shot Sim2Real transfer. We chose the ACT algorithm as our representative model because of its demonstrated efficiency and strong real-time inference performance. Meanwhile, we selected two distinct tasks: 1) picking a chip from the edge of a table and placing it in the center. 2) Pouring a ball from a cup held in the right hand into a cup held in the left. This is a simplified version of the Cup-to-Cup Transfer task, where we replaced the berry with a ball to reduce the randomness of the movement of the object. For each task, a dataset was collected and utilized to train the models within the simulation environment. These trained models were subsequently evaluated on both the simulation platform and a real robot.

The results are shown in Fig. 6. The models successfully executed both tasks with smooth and stable movements, from basic object picking and placing to more complex operations like dual-arm collaborative ball transfer. In addition to qualitative results, we also conducted quantitative experiments. The model achieved a 92.86% accuracy on the basic object picking and placing task, while it reached 71.43% Sim2Real accuracy for the complex ball transfer task without fine-tuning.

In summary, models trained exclusively on our simulation data can perform tasks seamlessly on a real robot without any fine-tuning, demonstrating the feasibility of zero-shot Sim2Real transfer.


Fig. 5: Comparison of traditional simulated, real, and our simulated scene.

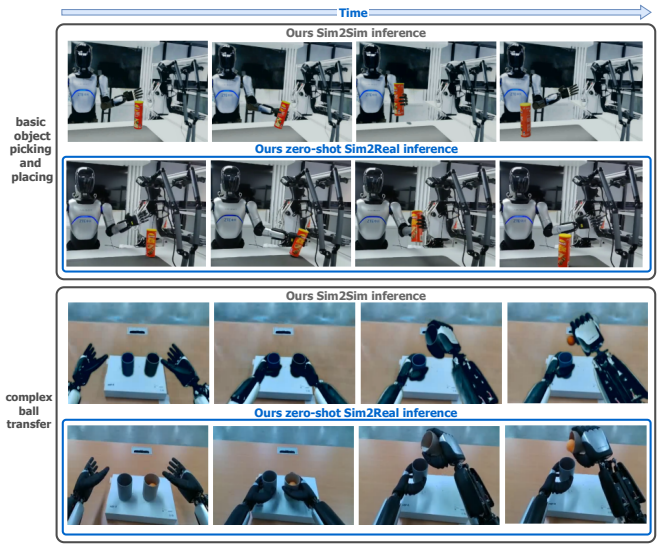


Fig. 6: Comparison of Sim2Sim and Sim2Real inference results.

## V. CONCLUSIONS

In response to the urgent necessity for a professional R&D platform in humanoid robotics and dexterous manipulation, we introduce RealMirror, a comprehensive, open-source platform. This platform provides a seamless, end-to-end solution for data collection, model training, model inference, and performance evaluation. Meanwhile, we demonstrate that models trained in our high-fidelity simulation environment can achieve zero-shot Sim2Real transfer to the real world. We are currently advancing the platform by developing a novel automatic data collection method and expanding its simulation scenarios, which will enable the generation of large-scale, high-quality simulation data. Additionally, further zero-shot Sim2Real transfer experiments are being conducted to validate the universality and robustness of our platform across a wider range of scenarios. We believe these systematic advancements will significantly accelerate the development of embodied AI.

## REFERENCES


[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[2] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[3] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[4] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[5] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.

[6] M. Shukor, D. Aubakirova, F. Capuano, P. Kooijmans, S. Palma, A. Zouitine, M. Aractingi, C. Pascal, M. Russi, A. Marafioti, *et al.*, "Smolvla: A vision-language-action model for affordable and efficient robotics," *arXiv preprint arXiv:2506.01844*, 2025.

[7] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. 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)*. IEEE, 2024, pp. 6892–6903.

[8] Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, S. Gao, X. He, X. Hu, X. Huang, *et al.*, "Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems," *arXiv preprint arXiv:2503.06669*, 2025.

[9] K. Wu, C. Hou, J. Liu, Z. Che, X. Ju, Z. Yang, M. Li, Y. Zhao, Z. Xu, G. Yang, *et al.*, "Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation," *arXiv preprint arXiv:2412.13877*, 2024.

[10] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, "Libero: Benchmarking knowledge transfer for lifelong robot learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 44 776–44 791, 2023.

[11] T. Mu, Z. Ling, F. Xiang, D. Yang, X. Li, S. Tao, Z. Huang, Z. Jia, and H. Su, "Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations," in *35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[12] J. Gu, F. Xiang, X. Li, Z. Ling, X. Liu, T. Mu, Y. Tang, S. Tao, X. Wei, Y. Yao, *et al.*, "Maniskill2: A unified benchmark for generalizable manipulation skills," *arXiv preprint arXiv:2302.04659*, 2023.

[13] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3019–3026, 2020.

[14] Y. Zhang, Z. Yu, J. Lai, C. Lu, and L. Han, "Agentworld: An interactive simulation platform for scene construction and mobile robotic manipulation," *arXiv preprint arXiv:2508.07770*, 2025.

[15] U. Robotics, "Xr teleoperation." https://github.com/unitreerobotics/xr_teleoperate, 2024.

[16] T. Hunyuan3D, S. Yang, M. Yang, Y. Feng, X. Huang, S. Zhang, Z. He, D. Luo, H. Liu, Y. Zhao, *et al.*, "Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material," *arXiv preprint arXiv:2506.15442*, 2025.

[17] N. Moenne-Loccoz, A. Mirzaei, O. Perel, R. de Lutio, J. Martinez Esturo, G. State, S. Fidler, N. Sharp, and Z. Gojcic, "3d gaussian ray tracing: Fast tracing of particle scenes," *ACM Transactions on Graphics (TOG)*, vol. 43, no. 6, pp. 1–19, 2024.

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

[19] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*. PMLR, 2023, pp. 2165–2183.

[20] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong, *et al.*, "Openvla: An open-source vision-language-action model," in *Conference on Robot Learning*. PMLR, 2025, pp. 2679–2713.

[21] L. X. Shi, B. Ichter, M. Equi, L. Ke, K. Pertsch, Q. Vuong, J. Tanner, A. Walling, H. Wang, N. Fusai, *et al.*, "Hi robot: Open-ended instruction following with hierarchical vision-language-action models," *arXiv preprint arXiv:2502.19417*, 2025.

[22] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, *et al.*, "$\pi$0. 5: a vision-language-action model with open-world generalization, 2025," *URL https://arxiv. org/abs/2504.16054*, 2025.

[23] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, *et al.*, "*pi*_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[24] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, *et al.*, "Gr00t n1: An open foundation model for generalist humanoid robots," *arXiv preprint arXiv:2503.14734*, 2025.

[25] W. Zhao, J. P. Queralta, and T. Westerlund, "Sim-to-real transfer in deep reinforcement learning for robotics: a survey," in *2020 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 2020, pp. 737–744.

[26] S. Höfer, K. Bekris, A. Handa, J. C. Gamboa, M. Mozifian, F. Golemo, C. Atkeson, D. Fox, K. Goldberg, J. Leonard, *et al.*, "Sim2real in robotics and automation: Applications and challenges," *IEEE transactions on automation science and engineering*, vol. 18, no. 2, pp. 398–400, 2021.

[27] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 5026–5033.

[28] S. W. Abeyruwan, L. Graesser, D. B. D'Ambrosio, A. Singh, A. Shankar, A. Bewley, D. Jain, K. M. Choromanski, and P. R. Sanketi, "i-sim2real: Reinforcement learning of robotic policies in tight human-robot interaction loops," in *Conference on Robot Learning*. PMLR, 2023, pp. 212–224.

[29] I. Radosavovic, T. Xiao, B. Zhang, T. Darrell, J. Malik, and K. Sreenath, "Learning humanoid locomotion with transformers," *CoRR*, 2023.

[30] A. Agarwal, A. Kumar, J. Malik, and D. Pathak, "Legged locomotion in challenging terrains using egocentric vision," in *Conference on robot learning*. PMLR, 2023, pp. 403–415.

[31] X. Hu, S. Li, T. Huang, B. Tang, R. Huai, and L. Chen, "How simulation helps autonomous driving: A survey of sim2real, digital twins, and parallel intelligence," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 593–612, 2023.

[32] Y. Yuan, H. Che, Y. Qin, B. Huang, Z.-H. Yin, K.-W. Lee, Y. Wu, S.-C. Lim, and X. Wang, "Robot synesthesia: In-hand manipulation with visuotactile sensing," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 6558–6565.

[33] T. G. W. Lum, M. Matak, V. Makoviychuk, A. Handa, A. Allshire, T. Hermans, N. D. Ratliff, and K. Van Wyk, "Dextrah-g: Pixels-to-action dexterous arm-hand grasping with geometric fabrics," *arXiv preprint arXiv:2407.02274*, 2024.

[34] M. N. Qureshi, S. Garg, F. Yandun, D. Held, G. Kantor, and A. Silwal, "Splatsim: Zero-shot sim2real transfer of rgb manipulation policies using gaussian splatting," *arXiv preprint arXiv:2409.10161*, 2024.

[35] W. Liu, Y. Wan, J. Wang, Y. Kuang, X. Shi, H. Li, D. Zhao, Z. Zhang, and H. Wang, "Fetchbot: Object fetching in cluttered shelves via zero-shot sim2real," *arXiv preprint arXiv:2502.17894*, 2025.

[36] R. Ye, W. Xu, H. Fu, R. K. Jenamani, V. Nguyen, C. Lu, K. Dimitropoulou, and T. Bhattacharjee, "Rcare world: A human-centric simulation world for caregiving robots," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 33–40.

[37] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, M. Lingelbach, J. Sun, *et al.*, "Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 80–93.

[38] NVIDIA, "Isaac sim 4.5 - robotics simulation and synthetic data generation," *https: //developer.nvidia.com/isaacsim*, 2025.

[39] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3019–3026, 2020.

[40] N. Chernyadev, N. Backshall, X. Ma, Y. Lu, Y. Seo, and S. James, "Bigym: A demo-driven mobile bi-manual manipulation benchmark," in *8th Annual Conference on Robot Learning*.

[41] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, M. Lingelbach, J. Sun, *et al.*, "Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 80–93.

[42] A. Mandlekar, S. Nasiriany, B. Wen, I. Akinola, Y. Narang, L. Fan, Y. Zhu, and D. Fox, "Mimicgen: A data generation system for scalable robot learning using human demonstrations," *arXiv preprint arXiv:2310.17596*, 2023.

[43] H. Wang, J. Chen, W. Huang, Q. Ben, T. Wang, B. Mi, T. Huang, S. Zhao, Y. Chen, S. Yang, *et al.*, "Grutopia: Dream general robots in a city at scale," *arXiv preprint arXiv:2407.10943*, 2024.

[44] Q. Wu, J. Martinez Esturo, A. Mirzaei, N. Moenne-Loccoz, and Z. Gojcic, "3dgut: Enabling distorted cameras and secondary rays in gaussian splatting," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[45] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[46] Z. Jin, Z. Che, Z. Zhao, K. Wu, Y. Zhang, Y. Zhao, Z. Liu, Q. Zhang, X. Ju, J. Tian, Y. Xue, and J. Tang, "Artvip: Articulated digital assets of visual realism, modular interaction, and physical fidelity for robot learning," 2025. [Online]. Available: https://arxiv.org/abs/2506.04941

[47] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks for manipulation research," in *2015 International Conference on Advanced Robotics (ICAR)*, 2015, pp. 510–517.

[48] R. Cadene, S. Alibert, A. Soare, Q. Gallouedec, A. Zouitine, S. Palma, P. Kooijmans, M. Aractingi, M. Shukor, D. Aubakirova, M. Russi, F. Capuano, C. Pascal, J. Choghari, J. Moss, and T. Wolf, "Lerobot: State-of-the-art machine learning for real-world robotics in pytorch," https://github.com/huggingface/lerobot, 2024.