RoboVerse: A Unified Benchmark for Scalable and Generalizable Vision-Language Robotic Manipulation

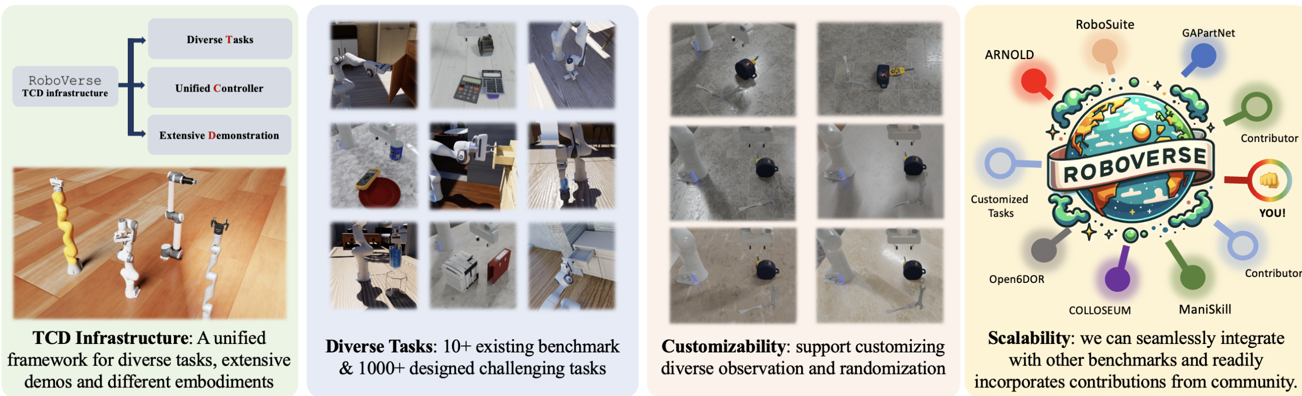

Figure 1. **Overview of `RoboVerse` benchmark.** We provide a unified infrastructure for robotic manipulation in simulation environments. This design unifies diverse tasks, extensive demonstrations, and different robot embodiments in existing robotic manipulation benchmarks. We further enrich the available robotic manipulation demonstration by scaling existing tasks with domain randomization and incorporating newly designed tasks. Our benchmark and dataset exhibit remarkable flexibility for each task, allowing for utilization across different observation modalities, diverse randomization strategies, and scalability with the joint efforts of the robotics community via an easy-to-use coding pipeline.

Abstract

The importance of diverse, high-quality datasets is underscored by their role in training foundational models, especially in fields like natural language processing and computer vision. However, scaling up data and models for robotics presents unique challenges due to the confinement of prior models to specific datasets and domains and the limitations inherent in collecting diverse real-world demonstrations. To overcome these limitations, we propose leveraging simulators as an alternative. Simulators can generate vast, diverse datasets and allow for flexible manipulation of various elements, such as observation representations and action formats, thereby offering a scalable and adaptable approach for training robotic models. To this end, we propose `RoboVerse` benchmark, in which we provide a unified infrastructure for diverse tasks, extensive demonstrations, and different robot embodiments. We also collect a large-scale dataset merging both existing benchmarks and newly designed tasks. Furthermore, our framework exhibits remarkable flexibility, allowing for utilization across different observation modalities, diverse randomization strategies, and scalable data augmentation.

1. Introduction

Recent advancements in foundational models highlight the growing importance of comprehensive and high-quality datasets in improving model performance and generalization capabilities. However, directly adapting such data scaling effects to robotics research faces several significant challenges in data collection. Predominantly, with data from different sources utilizing different input modalities (*e.g.*, RGB images, point clouds, *etc.*) and robot embodiments (*e.g.*, Franka Emika, UR10e, *etc.*), setting a universal standard for both data representation and task is difficult. Consequently, linking research findings across different experimental settings for a cohesive conclusion is challenging.

To address this limitation, prior works have attempted to collect large-scale manipulation demonstrations in both real-world and simulation environments. In real-world settings, the RT series [1–4, 14] have RGB recordings of robot manipulation at the cost of extensive data collection efforts. Simulation-based benchmarks such as ManiSkill [10, 12] and RoboSuite [16] collect demonstrations of specific tasks in different simulation environments, making it challenging to transfer policies learned in different benchmarks.

Recognizing the substantial effort needed to collect real-world data and the lack of unification in robot simulation benchmarks, our `RoboVerse` benchmark encompasses the following appealing features:

- **Unified Structure**: We devise a unified task structure, dataset format, and evaluation system to support a wide range of tasks, complemented by a suite of tools to facilitate task design and demonstration generation.
- **Flexibility and Diversity**: The flexibility of our benchmark allows for effortless customization of new tasks, observation representations, and action representations to suit specific needs. We unify existing demonstrations in simulation environments and also collect a large amount of tasks and demonstrations with rich annotations for downstream policy learning.
- **Comprehensive Modalities**: Utilizing simulation environments, we provide both multi-view RGB recordings, and point clouds as well as detailed language annotations catering to a variety of tasks.
- **Scalability**: We provide extensive APIs to make our benchmark scalable for robotics manipulation research, enabling the seamless addition of new tasks, demonstrations, and the training of new models.

Utilizing `RoboVerse`, we can thoroughly evaluate the performance and generalization capabilities of existing methods across input modalities, tasks, and robot embodiments through both interpolation and extrapolation. Additionally, with our integrated vision-language-action demonstrations, we can craft a versatile robotic manipulation policy for diverse tasks and complex scenarios.

2. `RoboVerse`

We propose `RoboVerse`, a comprehensive and multi-task benchmark, developed using the IsaacSim simulator [13].

2.1. Benchmark

Unified Infrastructure. We introduce our Task-Controller-Demonstration (TCD) Infrastructure for the `RoboVerse` benchmark. We have developed a base class for tasks, which includes all necessary functions and variables specific to each task. Upon defining a task, we inherit from this base class, customizing configurations and environment setup functions accordingly. Additionally, we have designed and implemented the controller infrastructure to manage embodiment, serving as an intermediary between the demonstration and the environment. It's worth noting that our controller is adaptable, accommodating various embodiments and standardizing their control mechanisms. Furthermore, we have collected extensive demonstrations for each task within our infrastructure, providing ample support for their utilization, re-rendering, replay, and evaluation. We have also established a standardized format for storing demonstrations across all tasks.

Task. Our `RoboVerse` benchmark integrates three key components: (1) pick and place, (2) articulated object manipulation, and (3) complex manipulation tasks. Specifically, we incorporate tasks, benchmarks, and demonstrations from various sources into our benchmark, including: (1) ManiSkill [10, 12], (2) SceneDiffuser [11], (3) GAPartNet [8], (4) PartManip [7], (5) ARNOLD [9], (6) Open6DOR [5], (7) COLOSSEUM [15], (8) SAGE[6]. Furthermore, we customize certain specific tasks to address shortcomings identified in previous benchmarks. For example, we combine object pick and place with articulated object manipulation to create tasks such as "open the top drawer of the cabinet on the left of the table, and place the apple into it." Additionally, we employ heuristics and reinforcement learning algorithms to execute these tasks and gather demonstrations.

Multiple Embodiment Support. We support multiple embodiments in our `RoboVerse` benchmark, including Franka Emika FR3, Kinova Gen3, KUKA IIWA, Kinova Jaco, UR10e, and so on.

Language Description. We provide precise and detailed language descriptions for each task.

2.2. Demonstration

We gather large-scale demonstrations for the benchmark, each comprising task configurations, trajectories, language annotations, and other useful information. We introduce a unified representation for these demonstrations, focusing on the trajectory of the end-effector pose and gripper state. This standardized format enables the seamless reuse of demonstrations across various embodiments and scenes. Additionally, we offer comprehensive APIs for domain randomization, facilitating the creation of a more diverse and realistic dataset. Specifically, we currently support randomization for (1) object colors, (2) ground plane, (3) lighting, (4) scene, (5) camera pose, (6) physical parameters, and (7) cross-embodiment.

For existing benchmarks with demonstrations, we directly adopt theirs and transfer them to our simulator environment, adjusting them to fit into the shared format. For other benchmarks or our newly designed tasks, we either utilize their existing policies, amalgamate several existing methods, or design heuristics to collect demonstrations.

3. Future Work

We are continuously expanding our benchmark and gathering more demonstrations to enhance our dataset. Leveraging these extensive demonstrations, we aim to quantitatively evaluate several key aspects crucial to the research community. Our goals include identifying optimal visual representations that ensure high performance and generalization, addressing data balancing challenges across various tasks, and assessing the model's generalization capabilities within and beyond its training distribution. Additionally, we plan to explore the contribution of simulation data to real-world applications, focusing on strategies for data balancing and identifying effective training paradigms.

References

[1] Michael Ahn, Debidatta Dwibedi, Chelsea Finn, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Karol Hausman, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, et al. Autort: Embodied foundation models for large scale orchestration of robotic agents. *arXiv preprint arXiv:2401.12963*, 2024. 1

[2] Suneel Belkhale, Tianli Ding, Ted Xiao, Pierre Sermanet, Quon Vuong, Jonathan Tompson, Yevgen Chebotar, Debidatta Dwibedi, and Dorsa Sadigh. Rt-h: Action hierarchies using language. *arXiv preprint arXiv:2403.01823*, 2024.

[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] 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. 1

[5] Yufei Ding, Haoran Geng, Chaoyi Xu, Xiaomeng Fang, Jiazhao Zhang, Songlin Wei, Qiyu Dai, Zhizheng Zhang, and He Wang. Open6dor: Benchmarking open-instruction 6-dof object rearrangement and a vlm-based approach. In *First Vision and Language for Autonomous Driving and Robotics Workshop*. 2

[6] Haoran Geng, Songlin Wei, Congyue Deng, Bokui Shen, He Wang, and Leonidas Guibas. Sage: Bridging semantic and actionable parts for generalizable manipulation of articulated objects. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*. 2

[7] Haoran Geng, Ziming Li, Yiran Geng, Jiayi Chen, Hao Dong, and He Wang. Partmanip: Learning cross-category generalizable part manipulation policy from point cloud observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2978–2988, 2023. 2

[8] Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7081–7091, 2023. 2

[9] Ran Gong, Jiangyong Huang, Yizhou Zhao, Haoran Geng, Xiaofeng Gao, Qingyang Wu, Wensi Ai, Ziheng Zhou, Demetri Terzopoulos, Song-Chun Zhu, et al. Arnold: A benchmark for language-grounded task learning with continuous states in realistic 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[10] 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. *arXiv preprint arXiv:2302.04659*, 2023. 1, 2

[11] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16750–16761, 2023. 2

[12] Tongzhou Mu, Zhan Ling, Fanbo Xiang, Derek Yang, Xuanlin Li, Stone Tao, Zhiao Huang, Zhiwei Jia, and Hao Su. Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations. *arXiv preprint arXiv:2107.14483*, 2021. 1, 2

[13] NVIDIA. Isaacsim simulator. 2

[14] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023. 1

[15] Wilbert Pumacay, Ishika Singh, Jiafei Duan, Ranjay Krishna, Jesse Thomason, and Dieter Fox. The colosseum: A benchmark for evaluating generalization for robotic manipulation. *arXiv preprint arXiv:2402.08191*, 2024. 2

[16] 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. In *arXiv preprint arXiv:2009.12293*, 2020. 1