# Robot Policy Evaluation for Sim-to-Real Transfer: A Benchmarking Perspective

Xuning Yang, Clemens Eppner, Jonathan Tremblay, Dieter Fox, Stan Birchfield, Fabio Ramos
NVIDIA

***Abstract*—Current vision-based robotics simulation benchmarks have significantly advanced robotic manipulation research. However, robotics is fundamentally a real-world problem, and evaluation for real-world applications has lagged behind in evaluating generalist policies. In this paper, we discuss challenges and desiderata in designing benchmarks for generalist robotic manipulation policies for the goal of sim-to-real policy transfer. We propose 1) utilizing high visual-fidelity simulation for improved sim-to-real transfer, 2) evaluating policies by systematic increasing task complexity and scenario perturbation to assess robustness, and 3) quantifying performance alignment between real-world performance and its simulation counterparts.**

## I. Introduction

Standardized evaluation has been crucial in the advancements of Large Language Models (LLMs) and Visual Language Models (VLMs). Strategic benchmarks such as Massive Multitask Language Understanding (MMLU) [6] and Holistic Evaluation of Language Models (HELM) [15] have presented a systematic way to represent language-based scenarios and evaluate trained policies against a set of diverse subjects. In addition, they include soft-metrics such as robustness, fairness, and bias to help understand the performance beyond just successful responses. These efforts lead to the development of the useful language AI applications that we use today.

In contrast, robotic evaluation for generalist manipulation policies has lagged behind, particularly for real-world applications. Current robotic benchmarks are characterized by specialized task suites with narrow focus, such as multi-task reinforcement learning [7, 24], VLM-based robotic reasoning [25], and limited testing tasks [16]. Moreover, most benchmarks lack considerations for robustness in deploying robot policies in the real world, which have been shown to significantly degrade policy performance [18].

The absence of a standardized, scalable robotic benchmark for sim-to-real transferability presents a critical bottleneck for visual policy for robotics. In this paper, we discuss the key desiderata for a robotic benchmark aimed at training generalist robot policies, ensuring that real-world challenges, such as robustness and task difficulty, are effectively represented. We describe several discrete and continuous metrics, as well as potential tools to be used for comparing simulation benchmarks to real robot performance. Lastly, we outline our approach for a scalable benchmark system using high-fidelity simulation for systematically evaluating robotic policies (Fig. 1). We hypothesize that systematic simulation has the potential to enable *scalable robotics benchmarking* as a viable proxy to extensive real-world experiments.

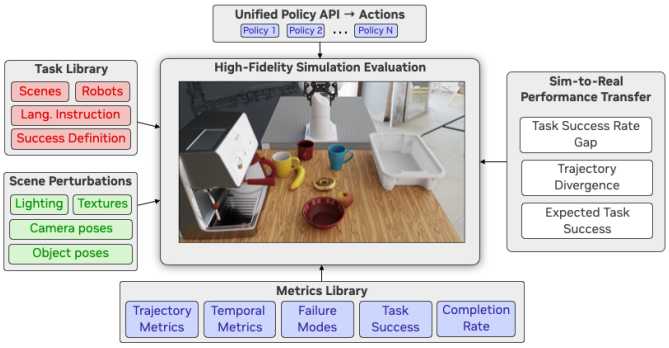


Fig. 1. Overview of the proposed evaluation benchmark.

## II. Challenges in Robotics Benchmarking

### A. Sim-to-Real

The sim-to-real gap remains a top challenge for vision-based policies. Transferring policies learned in simulation to real-world often fails due to various discrepancies in contact physics, visual appearance, and environmental dynamics with performance drop as high as 24–30% [13]. A common approach to both the visual and physical gap is to perform domain randomization [16, 26]. Another approach is to combine synthetic and real data, which requires fine-tuning on a set of environment-specific data in order to increase the performance of the model [8].

**Visual fidelity.** Visual fidelity plays a particularly critical role in this transfer. Traditional simulators often produce unrealistic visual observations that fail to capture the complexity of real-world lighting, textures, and environmental variations, as shown in Fig. 2. When trained with lower-quality simulation images, policies deployed in the real world face significant performance drops [9, 13]. However, with high-quality images, it is possible for the policy to transfer to the real world without any additional fine-tuning [9, 17]. This suggests that the level of photorealism in benchmarking environments is equally crucial for accurately evaluating sim-to-real policies.

**Scene variation.** Current datasets do not provide a systemic set of scene variations. However, recent works [18] have shown that changes in lighting and camera poses causes model success rate to degrade between 30–50%.

### B. Language Annotated Tasks

Robotics data has traditionally not focused on open-vocabulary instruction for tasks. However, recent robotic models have leveraged VLMs for generalizing to specific robotic

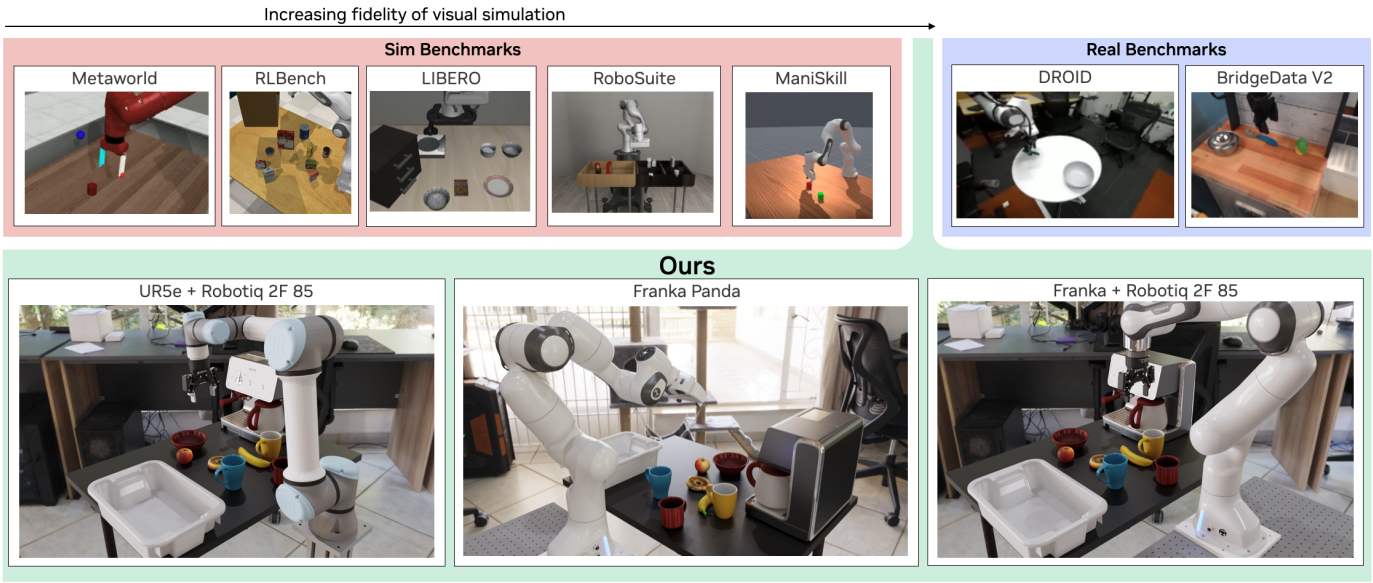


Fig. 2. Comparison of visual fidelity across various simulated benchmarks and real-world datasets.

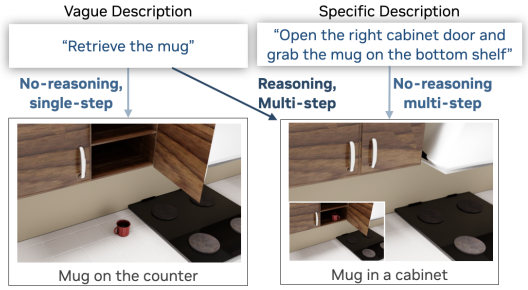


Fig. 3. A simple command such as *"retrieve the mug"* requires different levels of reasoning based on the scene. If the mug is in front of the robot, the task is fairly simple; however, if the mug is in a closed cabinet in a large kitchen, this would require open-world reasoning, including inferring the possible locations of the mug and recognizing and interacting with other objects in the scene such as cabinet doors in order to explore.

open-world task settings [22]. While some datasets incorporate natural language instructions [2, 7, 23], these datasets lack structured language annotations. As VLMs become prevalent in robotics, the language prompt itself may be a point of interest in future benchmarks, focusing on how well robots interpret and execute instructions that have varying levels of specificity [20]. Complexity of the task changes as a function of the scene and language instruction. Depending on the specificity of the instruction, the task may or may not require open-world reasoning, as illustrated in Fig. 3.

### C. Unified Platform

Existing datasets follow lax task definitions which are inconsistent across datasets, which vary widely and create difficulty in cross-platform and cross-embodiment comparisons. Additionally, robotic policies output various action spaces (*e.g.*, joint position/velocity, end effector pose [26], or action primitives [20, 14]), further complicating a standardized evaluation. To develop a unified benchmarking platform, it needs to be representative of the real-world conditions: the task structure needs to be systematic in categorizing tasks based on complexity and skill, and the policy interface needs to support multiple types of actions in order to enable a consistent comparison of policy architectures on identical tasks.

### D. Scale and Scope

Unlike datasets used to train LLM and VLMs, robotics data is difficult to obtain. Recent open-sourced initiatives such as RT-X [2] and DROID [10] have focused on using teleoperation to collect large-scale data in the real-world. For simulated data, aggregation frameworks such as RoboVerse [4] aim to unify benchmarks.

While large-scale datasets are useful in training, the suite of benchmarking tasks is not necessarily large in size but comprehensive in scope. Effective sim-to-real benchmarks need to have a curated set of tasks and environments that systematically cover a broad spectrum of skills, perceptual challenges, and task complexities. Khanna et al. [9] show that a small set of high-quality scenes ($\approx 200$) can outperform larger procedurally generated scenes ($\approx 10k$) in policy learning.

Thus, *scale* in benchmarking tasks needs to focus on diversity and real-world relevance. High-fidelity simulation enables systematic variations in representative tasks, which enables precise, repeatable benchmarking across a wide range of realistic scenarios. As real-robot data collection becomes less scalable as the field progresses, simulation offers a sustainable path forward.

## III. Desiderata for Robot Manipulation Benchmarking

Given the challenges above and the segregated landscape of existing frameworks, we present a recipe for a benchmarking framework for evaluating vision-based robotics policies. We employ the following definitions: A *task* $\mathcal{T}$ is a set of motions or *sub-tasks*, $\tau$, that completes a *language-based instruction*, $l$ [23]. We consider single-manipulator robot tasks, with the policy $\pi : \mathcal{O} \rightarrow \mathcal{A}$ where the action space $\mathcal{A}$ is policy dependent.

### A. Task Taxonomy

We introduce a novel task taxonomy that systematically categorizes tasks based on increasing complexity, required skills, and generalization. We categorize tasks into four difficulty levels:

- **T1 Single-motion tasks:** (e.g., pick, place, open, close) These typically involve a single, well-constrained action primitive involving a visually present object. These tasks test core visual reasoning and visuomotor capabilities. In particular, *pick/place* require the robot to reason about stable grasps; and *open/close* require reasoning about the joint constraints of the fixture (e.g., door hinges, sliders).
- **T2 Continuous-motion tasks:** (e.g., wiping, stirring, or pouring) These require smooth trajectories and precise control over a constrained space. These tasks require the robot to reason about tool-use and the space in which the continuous motion is constrained within.
- **T3 Multi-step tasks:** (e.g., put away, clean up) These combine multiple primitives into a temporally extended sequence of skills, which often require open-world reasoning of the scene and planning under partial observability and long-horizon dependency.
- **T4 Long-horizon tasks with memory:** Lastly, we consider cases where the robot needs to reason about its broader environment over its global memory[1]. These type of tasks require the robot to retain memory of objects' spatial relationships over time. These type of tasks are typical of mobile manipulators, where a task may involve retrieving objects from multiple locations.

### B. Robustness to Variations

To evaluate the robustness of a policy, it is important to apply a range of systematic perturbations to the environments. We introduce a suite of variations (Fig. 4) to simulate diverse deployment conditions, which may emerge during deployment in dynamically changing environments, following [13]:

- **V1 Object placement:** Object positions are perturbed within the workspace following a pre-specified distribution. These shifts assess the policy's ability to handle spatial displacements.
- **V2 Number of Objects:** This introduces distractor and occlusions that test the model's ability to distinguish relevant objects.
- **V3 Texture changes:** Surface textures of objects and background are randomized using a library of synthetic and real-world backdrops. These variations assess the policy's reliance on appearance-specific features and its ability to generalize across visually distinct but semantically identical environments.
- **V4 Lighting changes:** Altering conditions (e.g., ambient light intensity, directional light and shadow) challenges the visual encoder's robustness to changes in illumination.
- **V5 Camera pose variations:** Even small discrepancies in camera pose between training and deployment can lead to significant performance degradation, making robustness to pose variation critical for practical reliability. Therefore, perturbations in camera viewpoints tests the stability of the policy under pose deviations.

[1]This would be akin to Retrieval-Augmented Generation (RAG) [12] mechanisms employed in LLMs.

### C. Discrete and Continuous Metrics

Evaluations in robot policy learning have traditionally focused on task success rates, but this binary metric often fails to capture the full spectrum of policy performance [11]. To address this limitation, we define a more granular set of metrics including discrete and continuous metrics. These metrics allow us to understand policy behavior and limitations in end-to-end robot learning.

- **M1 Completion Rate $\mathcal{C}$:** The percentage of successful task completions in a total set of attempted tasks $\mathcal{C}(\pi)$. This quantifies the ability for the policy to complete the task from start to finish.
- **M2 Task Success $\mathcal{S}$:** We reframe this to describe the percentage of *sub-tasks* that have been completed: $\mathcal{S}(\mathcal{T}) = \frac{1}{\mathcal{T}} \sum_{\tau \in \mathcal{T}} \mathcal{S}(\tau)$ This approach provides a graded measure of success.
- **M3 Failure Modes:** These are systematically categorized to enable precise diagnosis of failure cases [1, 19]:
  a) *Grasp Failure:* The robot fails to establish initial contact, often due to inaccurate pose estimation, poor alignment, or insufficient gripper closure.
  b) *Grasp Stability Failure (Object Dropped):* The robot successfully grasps the object but subsequently loses it due to an unstable grasp.
  c) *Policy Generation Failure:* The policy outputs invalid and infeasible actions.
  d) *Reachability Failure:* The target action is unreachable due to the robot's kinematic constraints.
  e) *Reasoning Failure:* The robot exhibits incorrect high-level decision-making or planning, such as selecting inappropriate actions or misinterpreting task goals.
- **M4 Trajectory Metrics.** Trajectory metrics capture quality, efficiency, and desirability of robot motion.
  a) *Path Length:* The total distance traveled by the robot's end-effector during task execution.
  b) *Trajectory Smoothness:* Quantifies the consistency and fluidity of motion, measured by the higher derivatives of the trajectory.
  c) *Trajectory Optimality:* Quantifies whether the actions were time optimal; or if any corrective actions were taken.
- **M5 Temporal Metrics.** Temporal metrics capture time efficiency of task execution:
  a) *Total Time to Completion:* This measures system throughput and operational speed.
  b) *Average Policy Inference Time:* The measures the ability for the policy to be deployed online.
  c) *Episode Duration:* The total time span of an entire task attempt, including all actions and any recovery or correction phases.

### D. Sim-to-Real Transfer

We introduce quantitative metrics to evaluate the *transfer fidelity* of a policy's performance in sim vs. real.

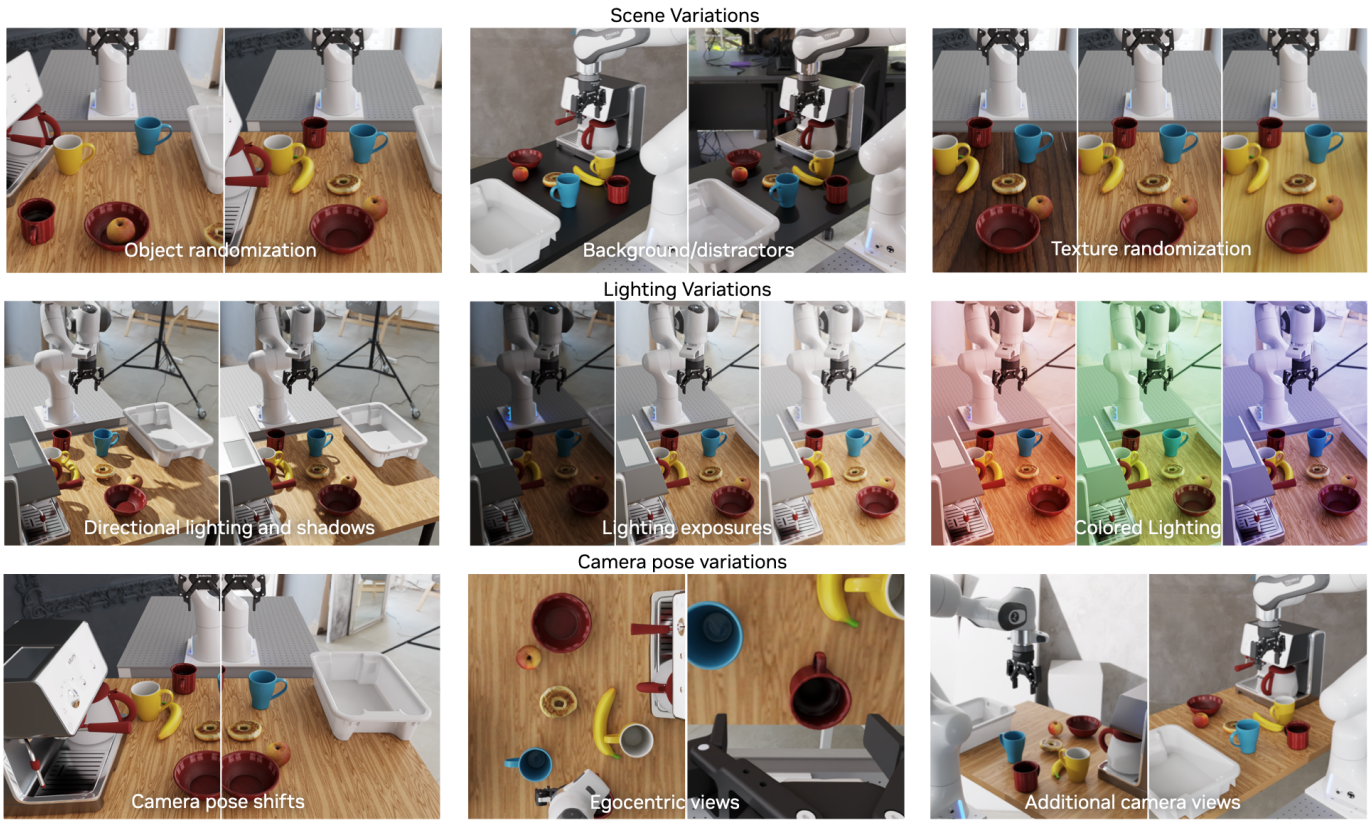


Fig. 4. Example scene variations, lighting variations, and camera pose variations. These are common perturbations present in real world settings.

**S1 Success Performance Matching in Fixed Baselines.** This refers to the difference in success rates between simulation and real-world deployment for a set of controlled scene equivalents. For a single task, this can be measured by $\|\mathcal{S}_{\text{sim}}(\pi) - \mathcal{S}_{\text{real}}(\pi)\|^2$, where $\|\cdot\|^2$ is the $L^2$ norm. However, for a range of tasks, Mean Maximum Rank Violation (MMRV) [13] has been used to describe performance shifts due to scene variations.

**S2 Trajectory Performance Matching.** We utilize *trajectory divergence*, a metric defined over the state evolution between trajectories executed in simulation and in real. For each motion $\tau = (s_0, s_1, \ldots, s_T)$, the divergence is given by $D(\{\tau^i_{\text{sim}}\}_{i=1}^N, \{\tau^j_{\text{real}}\}_{j=1}^M)$. Potential choices for $D(\cdot,\cdot)$ include the Maximum Mean Discrepancy (MMD) [5], energy statistics [21], and the classical Friedman-Rafsky test [3].

**S3 Expected Success Rate in Real.** Given the success rate obtained in simulation, it is possible to estimate the probability that the success rate for a policy $\pi$ in real is higher than a threshold $\theta$, as the posterior $p(\mathcal{S}_{\text{real}}(\pi) > \theta | \mathcal{S}_{\text{sim}}(\pi), \mathcal{T}, l) \propto p(\mathcal{S}_{\text{sim}}(\pi) | \mathcal{S}_{\text{real}}(\pi) > \theta, \mathcal{T}, l) p(\mathcal{S}_{\text{real}}(\pi))$, if the simulator provides ground-truth values. We aim to quantify this using our future experiments.

## IV. Proposed Benchmarking Framework

We discuss our initial efforts towards developing a vision-based robotic benchmarking framework, aimed at systematically evaluating robotic policies for improving sim-to-real transfer performance. The core objective of this framework is to 1) establish standardized protocols and metrics that evaluate vision-based policies in scalable high-fidelity simulation environments; and 2) quantify performance alignment between real-world experiments with its simulated equivalents, as described in Fig. 1.

We propose leveraging a high-fidelity visual simulator (IsaacSim) to bridge the visual perception gap between simulation and real-world. We procedurally generate tasks according to Sec. III-A, including scenes, language descriptions, task-success definitions. Additionally, the benchmark contains a suite of scene perturbations addendums, used to randomize the task library. and a suite of metrics as described in Sec. III-C.

Using our proposed benchmarking framework, we plan to perform a set of real-world experiments complementary to sim and a comprehensive analysis of the performance gap. By comparing sim-to-real performances using proposed metrics, our framework facilitates the identification of specific failure modes and evaluation domain gap. With this framework, we aim to increase systematic evaluation of robotic policies that scales as the field evolves. Ultimately, this framework is intended to serve as a reference pipeline for the broader community working on sim-to-real robot learning, including cross-comparison and reproducibility.

## References

[1] Christopher Agia, Rohan Sinha, Jingyun Yang, Zi ang Cao, Rika Antonova, Marco Pavone, and Jeannette Bohg. Unpacking failure modes of generative policies: Runtime

monitoring of consistency and progress, 2024. URL https://arxiv.org/abs/2410.04640.

[2] Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Muhammad Zubair Irshad, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Mart'in-Mart'in, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Vitor Guizilini, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open X-Embodiment: Robotic learning datasets and RT-X models. https://arxiv.org/abs/2310.08864, 2023.

[3] J. H. Friedman and L. C. Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *Annals of Statistics*, 7(4):697–717, 1979. doi: 10.1214/aos/1176344722.

[4] Haoran Geng, Feishi Wang, Songlin Wei, Yuyang Li, Bangjun Wang, Boshi An, Charlie Tianyue Cheng, Haozhe Lou, Peihao Li, Yen-Jen Wang, Yutong Liang, Dylan Goetting, Chaoyi Xu, Haozhe Chen, Yuxi Qian, Yiran Geng, Jiageng Mao, Weikang Wan, Mingtong Zhang, Jiangran Lyu, Siheng Zhao, Jiazhao Zhang, Jialiang Zhang, Chengyang Zhao, Haoran Lu, Yufei Ding, Ran Gong, Yuran Wang, Yuxuan Kuang, Ruihai Wu, Baoxiong Jia, Carlo Sferrazza, Hao Dong, Siyuan Huang, Yue Wang, Jitendra Malik, and Pieter Abbeel. Roboverse: Towards a unified platform, dataset and benchmark for scalable and generalizable robot learning, April 2025. URL https://github.com/RoboVerseOrg/RoboVerse.

[5] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander J Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[6] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In

*Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2009.03300.

[7] Stephen James, Zicong Ma, David R. Arrojo, and Andrew J. Davison. RLBench: The Robot Learning Benchmark & Learning Environment. *RAL*, 2020.

[8] Ryan Julian, Benjamin Swanson, Gaurav Sukhatme, Sergey Levine, Chelsea Finn, and Karol Hausman. Never stop learning: The effectiveness of fine-tuning in robotic reinforcement learning. In Jens Kober, Fabio Ramos, and Claire Tomlin, editors, *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pages 2120–2136. PMLR, 16–18 Nov 2021. URL https://proceedings.mlr.press/v155/julian21a.html.

[9] Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X. Chang, and Manolis Savva. Habitat Synthetic Scenes Dataset (HSSD-200): An Analysis of 3D Scene Scale and Realism Tradeoffs for ObjectGoal Navigation. *arXiv preprint*, 2023.

[10] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, Vitor Guizilini, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Muhammad Zubair Irshad, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. Droid: A large-scale in-the-wild robot manipulation dataset. 2024.

[11] Hadas Kress-Gazit, Kunimatsu Hashimoto, Naveen Kuppuswamy, Paarth Shah, Phoebe Horgan, Gordon Richardson, Siyuan Feng, and Benjamin Burchfiel. Robot learning as an empirical science: Best practices for policy evaluation, 2024. URL https://arxiv.org/abs/2409.09491.

[12] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL https://arxiv.org/abs/2005.11401.

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

[14] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *arXiv preprint arXiv:2209.07753*, 2022.

[15] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022. URL https://arxiv.org/abs/2211.09110.

[16] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning, 2023. URL https://arxiv.org/abs/2306.03310.

[17] Xiangyun Meng, Xuning Yang, Sanghun Jung, Fabio Ramos, Srid Sadhan Jujjavarapu, Sanjoy Paul, and Dieter Fox. Aim my robot: Precision local navigation to any object, 2024. URL https://arxiv.org/abs/2411.14770.

[18] Wilbert Pumacay, Ishika Singh, Jiafei Duan, Ranjay Krishna, Jesse Thomason, and Dieter Fox. The COLOSSEUM: A Benchmark for Evaluating Generalization for Robotic Manipulation. In *RSS*, 2024.

[19] Som Sagar, Jiafei Duan, Sreevishakh Vasudevan, Yifan Zhou, Heni Ben Amor, Dieter Fox, and Ransalu Senanayake. From mystery to mastery: Failure diagnosis for improving manipulation policies, 2025. URL https://arxiv.org/abs/2412.02818.

[20] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530, 2023. doi: 10.1109/ICRA48891.2023.10161317.

[21] Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

[22] Gemini Robotics Team. Gemini robotics: Bringing AI into the physical world. *arXiv:2503.20020*, 2025.

[23] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-

Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, 2023.

[24] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Avnish Narayan, Hayden Shively, Adithya Bellathur, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning, 2021. URL https://arxiv.org/abs/1910.10897.

[25] Enyu Zhao, Vedant Raval, Hejia Zhang, Jiageng Mao, Zeyu Shangguan, Stefanos Nikolaidis, Yue Wang, and Daniel Seita. ManipBench: Benchmarking vision-language models for low-level robot manipulation. *arXiv:2505.09698*, 2025.

[26] Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Kevin Lin, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. In *arXiv preprint arXiv:2009.12293*, 2020.