THE COLOSSEUM: A Benchmark for Evaluating Generalization for Robotic Manipulation

Wilbert Pumacay*[1], Ishika Singh*[2], Jiafei Duan*[3], Ranjay Krishna[3,4], Jesse Thomason[2], Dieter Fox[3,5]
[1]Universidad Católica San Pablo, [2]University of Southern California,
[3]University of Washington, [4]Allen Institute for Artifical Intelligence, [5]NVIDIA
*Equal contribution
robot-colosseum.github.io

Abstract

To realize effective large-scale, real-world robotic applications, we must evaluate how well our robot policies adapt to changes in environmental conditions. Unfortunately, a majority of studies evaluate robot performance in environments closely resembling or even identical to the training setup. We present THE COLOSSEUM*, a novel simulation benchmark, with 20 diverse manipulation tasks, that enables systematical evaluation of models across 14axes of environmental perturbations. These perturbations include changes in color, texture, and size of objects, tabletops, and backgrounds; we also vary lighting, distractors, physical properties perturbations and camera pose. Using* THE COLOSSEUM*, we compare 5 state-of-the-art manipulation models to reveal that their success rate degrades between 30-50% across these perturbation factors. When multiple perturbations are applied in unison, the success rate degrades $\geq$75%. We identify that changing the number of distractor objects, target object color, or lighting conditions are the perturbations that reduce model performance the most. To verify the ecological validity of our results, we show that our results in simulation are correlated ($\bar{R}^2 = 0.614$) to similar perturbations in real-world experiments. We open source code for others to use* THE COLOSSEUM*, and also release code to 3D print the objects used to replicate the real-world perturbations. Ultimately, we hope that* THE COLOSSEUM *will serve as a benchmark to identify modeling decisions that systematically improve generalization for manipulation.*

1. Introduction

The promise of robotics requires ubiquity. For effective real-world deployment, robots must operate in a variety of environments. When asked to turn on a stove, a robot should be able to turn the stove's knob, regardless of the size of the knob, irrespective of the kitchen's backdrop, invariant to the kitchen counter's texture, during the day, or even under a dim evening light. Unfortunately, a majority of studies evaluate robot performance in environments closely resembling or even identical to the training setup [1–3, 11].

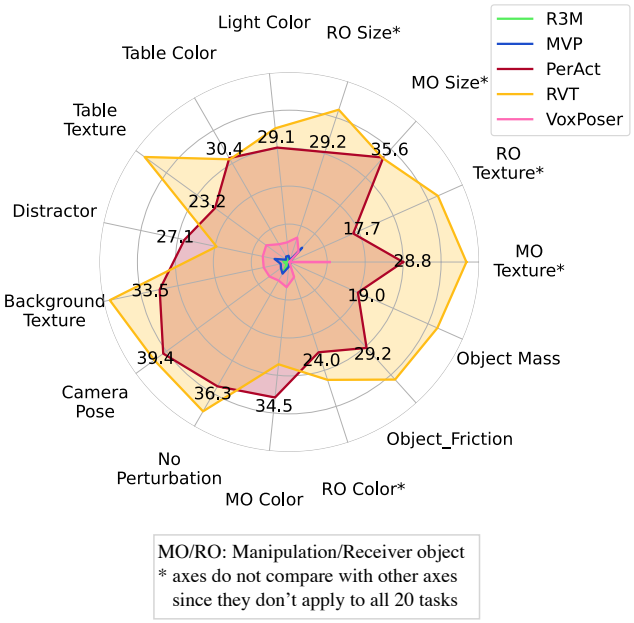

Figure 1. **Evaluating generalization with THE COLOSSEUM.** Task-averaged success rate for 5 SotA robotic manipulation policies over 14 perturbation factors and 20 robotic manipulation tasks. Changes in RGB input space affects all models due to end-to-end RGB-based training. Image-based models are also affected by camera pose change, while models without in-the-wild pretraining suffer in the presence of distractors.

We introduce THE COLOSSEUM, a comprehensive benchmark aimed at systematically evaluating the generalization of robot manipulation to environmental perturbations. THE COLOSSEUM introduces perturbations across 20 different tasks from the RLBench [6] framework, spanning 14 dimensions of perturbations. These perturbations

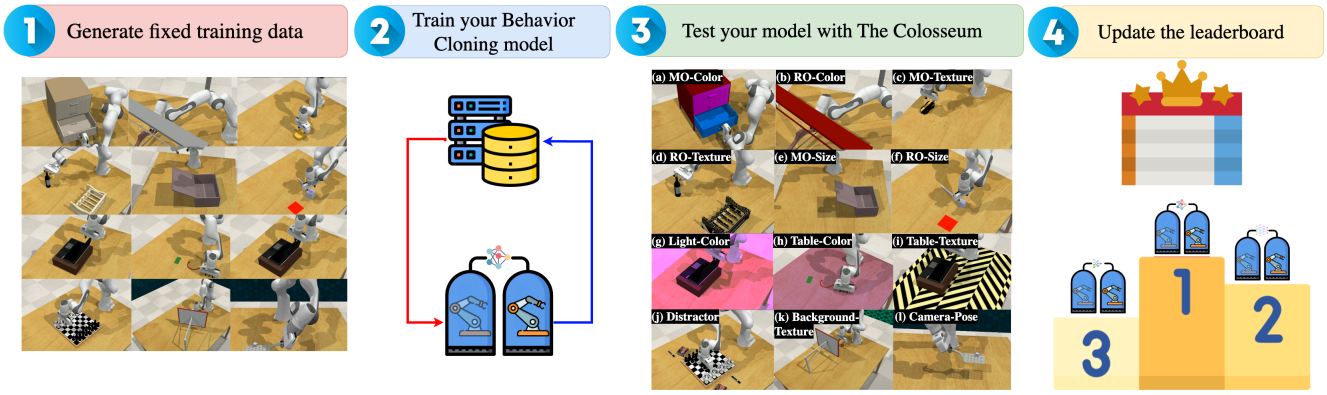

Figure 2. **The Colosseum Challenge.** This challenge is designed to enhance generalization of Behavior Cloning (BC) models in robotic manipulation tasks. It involves four key phases: 1) Participants generate a standard training dataset from 20 tasks with 100 demonstrations each, without `perturbation_factors`. 2) Participants train their BC models using this standardized dataset. 3) The models are restricted to evaluate over a fixed 25 episodes across 14 different `perturbation_factors`. 4) Models are ranked on a leaderboard based on the percentage change in their performance across these factors. We've shown that simulation aligns with real-world evaluation, so participants can expect similar generalization when participating in the simulation benchmark.

include object color, object texture, object size, table color, table texture, the presence of distractor objects, changes to the camera pose, and changes to physical properties like friction and mass, inspired by those observed in real-world robotics datasets [4, 7, 9]. The Colosseum also includes a parallel real world evaluation with task setups and objects reproducible via open-sourced 3D printing models.

We evaluate 5 state of the art robot manipulation models [3, 5, 8, 10, 11] using The Colosseum and draw insights into answers for critical research questions on generalization for BC policies. We establish a strong correlation between falling task success under perturbations in simulations and those observed in real-world scenarios for the same tasks, suggesting that The Colosseum evaluations in simulation give reliable insight into real world generalization at a fraction of the setup cost. The Colosseum challenge and leaderboard (Figure 2) will provide as a unified platform to develop, evaluate, and compare future robotic manipulation methods that stand the test of robustness and generalization.

2. Results

We summarize our key results: For 2D learning models (`R3M-MLP` and `MVP-MLP`), we observe that object and light color, texture, and camera pose are the most affecting factors. Since these models are trained end-to-end with RGB inputs, and the color or texture related perturbations shift the input space, thereby affecting the output space as well. Moreover, training with specific `Camera_Poses` when using RGB as input also affects the performance when camera poses are perturbed. For zero-shot manipulation models using Large Pretrained World Models, we observed that the system demonstrates robust generalization capabilities across various conditions, particularly excelling in tasks where it is predisposed to succeed. Specifically, for the two tasks in which `VoxPoser` excels, it maintains consistent performance across all variants. For 3D learning models (`RVT` and `PerAct`), we observe that the most affecting factors are color-related including object, table and light colors as well as presence of `Distractors`. Since `RVT` and `PerAct` are both trained end-to-end with RGB images or voxel grid with RGB channels, the color perturbations remain challenging for these models as well. These models lack any real-world pretraining, thus, the presence of `Distractors` puts the scene out of distribution, significantly affecting their performance. We observe that these model are robust to changes in `Camera_Pose`, because they do not directly learn on captured view. They instead preprocess the input RGBD views into a voxel grid or re-rendered novel views. On physical perturbations, `RVT` performs better than `PerAct`, perhaps because modelling in `RVT` is more robust for keypoint prediction than `PerAct` under these perturbations. Physical perturbation results for other models are inconclusive, as they cannot perform the tasks that support these perturbations. 3D baselines are better performing generally (Figure 1), and much more robust to environment perturbations as compared to 2D baselines. We also observe that `RVT`, trained only with RGB views, generally gets more affected with `perturbation_factors` as compared to `PerAct`, trained with complete 3D scene, notably in the case of `Distractors`. This result indicates value in learning with 3D scenes as input, for the resultant model is more robust to such environmental perturbations, as it might be learning 3D features of the objects instead of just their 2.5-dimensional projections. For more details and resources, please refer to the website linked above.

References

[1] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: Robotics Transformer for Real-World Control at Scale. In *arXiv preprint arXiv:2212.06817*, 2022. 1

[2] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[3] Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. Rvt: Robotic view transformer for 3d object manipulation. *arXiv:2306.14896*, 2023. 1, 2

[4] Kristen Grauman, Andrew Westbury, and et al. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[5] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023. 2

[6] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020. 1

[7] Alexander Khazatsky, Karl Pertsch, and et al. Droid: A large-scale in-the-wild robot manipulation dataset. 2024. 2

[8] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A Universal Visual Representation for Robot Manipulation. In *6th Annual Conference on Robot Learning*, 2022. 2

[9] 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. 2

[10] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-World Robot Learning with Masked Visual Pre-training. In *6th Annual Conference on Robot Learning*, 2022. 2

[11] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022. 1, 2