Data-Efficient and Contact-Rich Manipulation Through Diffusion Augmentation and Vision-Language Models

Daniel Seita[1]

[1]Department of Computer Science, University of Southern California
1031 Downey Way, Los Angeles, CA 90089, USA
Email: seita@usc.edu; Website: https://danielseita.github.io/; Lab: https://slurm-lab-usc.github.io/

Abstract

Recent progress in robot learning has produced impressive results, yet many systems still require learning from large datasets of demonstrations and are less effective in clutter or with highly deformable objects. This talk presents work on data-efficient manipulation using (i) diffusion-based augmentation that synthesizes geometrically consistent images and action labels to reduce demonstration requirements and (ii) Vision-Language Models (VLMs) that inject high-level semantics for contact-rich motion planning in clutter. We will also introduce ManipBench, which evaluates VLMs' abilities for low-level manipulation. Together, we show how to move the community towards achieving robot manipulators that can learn and operate with reduced demonstration requirements across cluttered and real-world environments.

Talk Overview

Part I: Diffusion for Bimanual Data Augmentation. Learning bimanual manipulation via visual imitation learning remains difficult due to the need for large datasets and the complexity of coordinating two arms. To address this, we developed D-CODA (**D**iffusion for **CO**ordinated **D**ual-Arm Data **A**ugmentation) (Liu et al. 2025), which trains a diffusion model to synthesize *viewpoint-consistent, wrist-camera images* for both arms along with new valid action labels. By distinguishing between contactless and contact-rich states, and enforcing constraints during augmentation, D-CODA generates diverse yet feasible data for policy training. Across simulated and real-world tasks, the augmented demonstrations result in improved performance. We have also extended this idea for bimanual data augmentation from *third-person* viewpoints (Chen et al. 2026).

Part II: VLMs for Contact-Rich Manipulation. Traditional motion planners treat all collisions as failures, which is overly restrictive in clutter where incidental contact may be necessary. IMPACT (**I**ntelligent **M**otion **P**lanning with **A**cceptable **C**ontact **T**rajectories) leverages VLMs to infer object semantics and assign contact-tolerance costs *without collecting additional training data* (Ling et al. 2026). These costs are encoded into 3D maps that integrate with motion planners, enabling trajectories that push aside robust objects while avoiding fragile ones. In simulation and real-world evaluations, IMPACT results in higher success rates and trajectories that human evaluators prefer over baselines.

Part III: Benchmarking VLMs for Manipulation. Finally, while VLMs are increasingly used for high-level planning, their ability to support low-level manipulation reasoning is less understood. To fill this gap, we introduced ManipBench (Zhao et al. 2025), a benchmark with 12,617 multiple-choice questions spanning pick-and-place, articulated object manipulation, deformable objects, and dynamic manipulation. We evaluated **33** models across **10** families, revealing wide variance in low-level reasoning ability and a clear correlation between benchmark scores and real-world task performance. These results, along with those from our prior benchmark on physical reasoning (Chow et al. 2025), provides the community with resources to evaluate and guide progress in using VLMs as "agents" for manipulation.

Conclusion. Together, these efforts demonstrate how diffusion models, VLM-based semantics, and rigorous benchmarks can complement each other: augmenting scarce data, enabling contact-aware planning, and measuring low-level reasoning. By unifying these directions, my research aims to make robotic manipulation more robust, generalizable, and useful for complex and cluttered real-world settings.

References

Chen, J.; Liu, I.-C. A.; Sukhatme, G.; and Seita, D. 2026. ROPA: Synthetic Robot Pose Generation for RGB-D Data Augmentation. *Under Review*.

Chow, W.; Mao, J.; Li, B.; Seita, D.; Guizilini, V.; and Wang, Y. 2025. PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding. *International Conference on Learning Representations (ICLR)*.

Ling, Y.; Owalekar, K.; Adesanya, O.; Bıyık, E.; and Seita, D. 2026. IMPACT: Intelligent Motion Planning with Acceptable Contact Trajectories via Vision-Language Models. *Under Review*.

Liu, I.-C. A.; Chen, J.; Sukhatme, G.; and Seita, D. 2025. D-CODA: Diffusion for Coordinated Dual-Arm Data Augmentation. *Conference on Robot Learning (CoRL)*.

Zhao, E.; Raval, V.; Zhang, H.; Mao, J.; Shangguan, Z.; Nikolaidis, S.; Wang, Y.; and Seita, D. 2025. ManipBench: Benchmarking Vision-Language Models for Low-Level Robot Manipulation. *Conference on Robot Learning (CoRL)*.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.