Data-Efficient Robot Learning for Manipulation in Complex Environments

Daniel Seita
November 16, 2025
http://danielseita.github.io seita@usc.edu

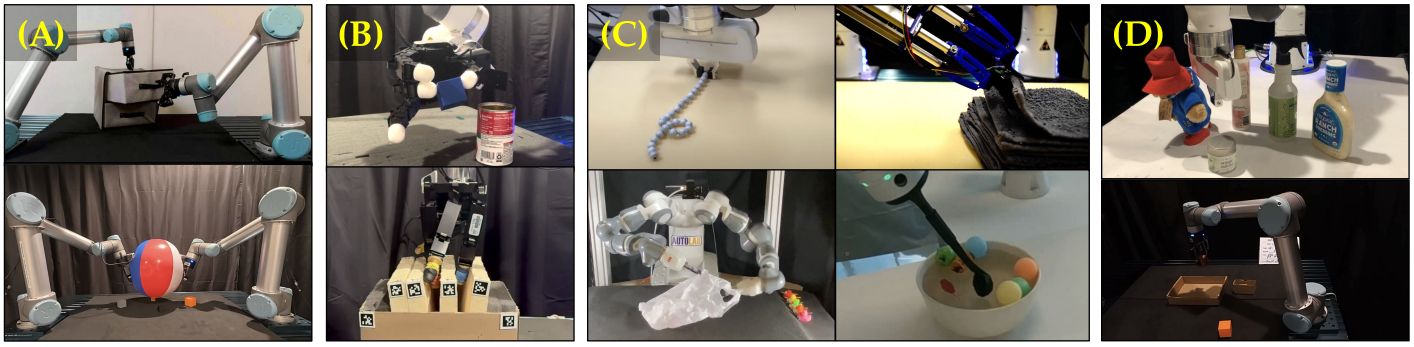

Figure 1: Overview of my prior work on robot learning for robot manipulation. (A, Section 2.1): we have designed methods for data-efficient bimanual manipulation via 3D representations [20] and data augmentation [19]. (B, Section 2.2): we have studied manipulation using dexterous multi-fingered hands [9, 15]. (C, Section 2.3): we have conducted deformable object manipulation tasks involving ropes [23], fabrics [28], bags [2], and scooping [30]. (D, Section 2.4): we have applied [18] and benchmarked [33] VLMs for robotics.

1 Introduction

Autonomous robots have the potential to transform industries ranging from healthcare to manufacturing, but a fundamental challenge remains: learning robust manipulation policies requires vast amounts of data. Unlike computer vision or natural language processing, where large-scale datasets enable foundation models to achieve remarkable generalization, robotics faces unique constraints [7]. Physical data collection is expensive, time-consuming, and often dangerous, and even the largest community-scale robotics datasets [6] focus primarily on simple pick-and-place tasks of rigid objects in free space. This data scarcity bottleneck limits the deployment of learning-based manipulation systems in real-world settings. **To achieve greater progress in manipulation tasks involving deformable materials, dense clutter, and other contact-rich settings, my thesis is that we will always need some form of faster and automated data collection beyond standard demonstrations.**

My research addresses this challenge by developing *data-efficient robot learning methods for robot manipulation* that reduce the need for extensive physical demonstrations (see Figure 1 for representative examples of prior work). I pursue this goal through complementary approaches that reduce reliance on additional physical demonstrations. These include: (i) data augmentation techniques that synthetically expand training datasets while preserving action label correctness, (ii) leveraging pre-trained Vision-Language Models (VLMs) for zero-shot semantic reasoning, and (iii) automated data collection systems through simulation, self-supervised learning, and real-world infrastructure that generate large-scale, high-quality training data. A common theme my group focuses on is applying these techniques on two particularly challenging application domains: *deformable object manipulation* and *manipulation in dense clutter*. Both domains require robots to reason about complex physical interactions, make contact-rich decisions, and handle high-dimensional state spaces, making them ideal testbeds for data-efficient learning approaches. My work demonstrates that by combining these principled approaches, we can substantially reduce the data requirements for robot learning while improving generalization capabilities.

2 Prior Research

2.1 Data-Efficient Bimanual Manipulation

Bimanual manipulation represents a critical capability for general-purpose robots, enabling coordinated lifting, assembly, and deformable object manipulation tasks essential for warehouse automation and fulfillment operations. However, learning bimanual policies is challenging due to higher-dimensional action spaces and the need for coordinated control between two arms. Imitation learning approaches [4, 32] require substantial amounts of demonstrations, which is time-consuming to collect in the real world. To address these challenges, my research has pursued *data efficient methods to reduce demonstration requirements*, particularly in contact-rich scenarios. First, we developed VoxAct-B [20], which improves sample efficiency directly from real demonstrations by using 3D voxel-based representations and spatial equivariance. This enables "acting" and "stabilizing" behaviors with fewer examples, where equivariance allows networks to better generalize to different locations in the scene. I earlier explored a similar concept for data-efficient learning in 2D tabletop manipulation [23]. Second, we have proposed novel *data augmentation* methods for bimanual manipulation across different camera viewpoints. D-CODA [19] introduces *contact-aware wrist-camera augmentation* that uses diffusion models to synthesize new egocentric views while enforcing constraints so that robot-object contacts remain physically consistent. ROPA [1] generalizes this idea to *third-person RGB-D augmentation* by generating new bimanual poses and viewpoints through diffusion-based synthesis and constrained kinematics, again ensuring that gripper-object contact geometry is preserved.

2.2 Dexterous Multi-Fingered Manipulation

Dexterous multi-fingered hands with high degrees-of-freedom (DOF) offer robots greater flexibility in manipulation by enabling diverse contact strategies beyond parallel-jaw grippers. However, learning to control high-DOF systems remains challenging, particularly when robots must decide which parts of the hand to use and how to coordinate multiple contact points. My work develops methods that equip dexterous hands with policies for contact-rich manipulation, often trained in simulation and deployed in the real world. For example, we have used differentiable force-closure reasoning, physics-based validation, and diffusion-based generation to allow a four-fingered robot hand to sequentially grasp objects [9]. We have also studied different modes of dexterous manipulation, such as one where the robot must *isolate, grasp, and retrieve objects* in tightly packed clutter [15]. To do this, we introduced a displacement-based state representation and a multi-phase reinforcement learning procedure that transfers directly from simulation to real hardware. Other modes of dexterous interaction that we have explored include nonprehensile pushing and pulling, which expand the range of tasks that a single hand can perform [16]. Finally, while these preceding works use existing hands (e.g., the Allegro Hand), we have also designed our own dexterous robotic hand to provide a multisensory platform that integrates vision, thermal, force, and tactile information, enabling richer contact understanding and supporting future research on multimodal dexterous manipulation [34]. Together, our efforts demonstrate how dexterous hands can achieve flexible and contact-aware manipulation behaviors that extend the capabilities of parallel-jaw grippers.

2.3 Manipulation of Deformable Objects using Machine Learning

Deformable object manipulation presents unique data efficiency challenges due to high-dimensional state spaces and complex dynamics [21]. My research has addressed these challenges through complementary approaches that reduce reliance on physical demonstrations.

Learning from Simulation with Domain Randomization. To minimize physical data collection, we developed methods that leverage simulation for deformable manipulation. For fabric smoothing [24], which is common in applications such as handling laundry and making beds [25], we used domain randomization [29] to transfer policies trained entirely in simulation to physical robots. We

extended this with VisuoSpatial Foresight (VSF) [10, 11], a model-based method that learns dynamics models from simulated interactions for multi-step fabric manipulation. By training in simulation, VSF achieved precise fabric smoothing and folding directly on physical robots.

Generalizable Representations for Multi-Task Learning. Rather than collecting task-specific data, we have developed representations that enable generalization. Our dense descriptor approach [8] learns pixel-wise fabric correspondences that transfer across different fabrics and tasks, allowing a single representation to support diverse pick-and-place operations without retraining. We complemented visual representations with tactile feedback [28], using sensors to enable precise multi-layer fabric grasping, a capability that visual data alone could not provide due to occlusions.

Self-Supervised Learning with Automated Labeling. To scale beyond manual demonstrations, we developed self-supervised approaches for *dynamic manipulation*. For cable [17, 31] and fabric flinging [3], we designed parametric trajectory controllers that robots could execute autonomously, with automatic labeling after each motion by detecting positions or measuring coverage. This automated procedure enabled large-scale data collection without human supervision.

Efficient Learning for Granular Manipulation. We have extended our data-efficient learning approaches to granular media manipulation with legged quadruped robots [12, 13]. We designed a custom gantry system that safely automated data collection by mimicking the robot's leg actions, enabling scalable data gathering without risking damage to the physical robot. From this real data, we then trained two models: one for how leg actions reshape the granular surface and move nearby obstacles, and one for how those same actions affect the robot's own state as it excavates [13]. By combining these models, the robot can plan sequences of leg actions that account for environment changes, robot motions, and the complexity of granular flow. This shows how to make learning feasible even in high-contact and deformable environments where simulation is unreliable.

2.4 Benchmarking and Deploying Vision-Language Models (VLMs) for Robotics

Foundation models such as VLMs have substantially advanced perception and reasoning in AI. They offer a pathway to data efficiency by providing semantic priors and zero-shot reasoning capabilities that reduce the need for task-specific training data [7]. To understand their physical and manipulative capabilities in robotics, we developed several visual question-answering benchmarks. These include PhysBench [5] to benchmark VLMs' intuitive physical world understanding and ManipBench [33] to evaluate low-level affordance reasoning such as contact prediction and object-object interactions. We also studied VLMs in human-robot settings through HRIBench [26], which analyzes accuracy-latency trade-offs for real-time human perception. Finally, we showed that VLMs can provide useful semantic and task-related information in robotics settings, first through fabric manipulation [22] and then about *object properties and acceptable contacts*, enabling contact-rich planning in clutter [18]. Collectively, these projects establish the current capabilities but also reveal important limitations of VLMs for robust physical prediction in robotics. This provides a foundation for my future work on adapting and improving these models for high-contact manipulation in dense clutter.

3 Future Research

My prior work has demonstrated that data augmentation, semantic reasoning with foundation models, and learning from simulation can substantially reduce the need for real-world robotics data. However, physical data will always remain limited relative to the scale required for foundation-model-level generalization. Therefore, *I propose to develop methods that produce more useful data, reuse existing data across embodiments, and exploit contact and multimodal structure in the physical world*. Below, I discuss some major research directions that I will pursue **in the next 1-5 years**:

Multimodal Data Augmentation and Cross-Embodiment Transfer. My prior work has studied *visual* data augmentation [1, 19], but manipulation inherently requires reasoning across multiple

modalities including vision, touch, force, and language. To this end, I plan to develop augmentation methods that utilize generative models to synthesize *tactile sensor signals* that are consistent with visual data and that preserve contact geometry. For *language* augmentation, our recent work shows that Vision-Language-Action (VLA) models are highly sensitive to small variations in phrasing, with minor prompt changes causing large performance drops [27]. I will develop techniques to automatically generate diverse and appropriate task descriptions from existing demonstrations, and then fine-tune VLAs on this data. *Cross-embodiment learning* offers another path to data efficiency, and I will develop methods to transfer bimanual demonstrations across robot morphologies through embodiment-invariant manipulation primitives combined with visual augmentation techniques.

Robot Foundation Models for High-Contact Manipulation. I plan to make fundamental progress on tasks that involve frequent and complex contacts between a robot's tool (or end-effector) and objects, or with object-object contact, which frequently occurs in dense clutter. My research will adapt VLAs to better perform manipulation in these situations, where robots must reason about safety, physical affordances, and acceptable contact. Building on our work that uses VLMs for contact-rich manipulation [18] and our recent benchmarks [5, 33], I plan to develop *object-centric representations that allow VLAs to predict contact outcomes and plan interactions in tight spaces*. This includes manipulation tasks that require reasoning about deformable layers, occluded objects, and complex object-object contact, as well as manipulation applications in surgical [14] and scooping [30] contexts. The goal is to make VLAs more reliable in settings where traditional motion planning struggles due to contact complexity, and we will propose new benchmarks for evaluating VLAs in clutter.

Whole-Body and Multifunctional Manipulation. Inspired by how humans leverage their entire body for manipulation (e.g., using arms to stabilize, and elbows or backs of hands to push objects), I aim to develop robots that can *flexibly deploy multiple parts of their embodiment* for diverse manipulation strategies. My recent work on sequential multi-object grasping [9], nonprehensile manipulation with dexterous hands [16], manipulation with legs [12, 13], and building a multisensory robotic hand [34] provides a foundation for this direction. Building on these capabilities, I will develop planning and learning algorithms that enable robots to: (i) use different hand surfaces (fingertips, palm, back of hand) for simultaneous grasping and pushing, (ii) leverage arm links for bracing and obstacle manipulation in clutter, and (iii) determine when to switch between prehensile (grasping) and nonprehensile (pushing/pulling) strategies. By combining semantic reasoning from VLMs with whole-arm feedback from multimodal sensors, I hope to allow robots to make informed decisions about which body parts to use and which objects are safe to contact.

3.1 Long-Term Vision: Integrated Data-Efficient Learning for Flexible Robot Manipulation

My **5-10 year vision** is to build robots that learn general-purpose manipulation skills through *data-efficient pipelines* that combine augmentation, cross-embodiment transfer, semantic reasoning, and multimodal sensing. I envision robots that can *autonomously compose these capabilities*, determining when to grasp, when to push, when to use arm or hand surfaces, and when to rely on tactile or visual feedback based on task requirements and contact affordances. This framework would enable robots in home assistance to handle diverse objects including deformables with minimal supervision, and in manufacturing to rapidly adapt to new product variants. By making robot learning more data-efficient while increasing robot flexibility through whole-arm and multifunctional manipulation, my research aims to accelerate deployment of autonomous systems that operate reliably in unstructured contact-rich environments. We are at an exciting juncture in robotics research, where advances in foundation models, data augmentation, and multimodal sensing are converging to enable more capable and adaptable robots. I am eager to continue pushing the boundaries of data-efficient robot learning and to mentor the next generation of researchers in this rapidly evolving and growing field.

References

[1] Jason Chen, I-Chun Arthur Liu, Gaurav Sukhatme, and **Daniel Seita**. "ROPA: Synthetic Robot Pose Generation for RGB-D Data Augmentation". In: *arXiv preprint arXiv:2509.19454* (2025).

[2] Lawrence Yunliang Chen, Baiyu Shi, **Daniel Seita**, Richard Cheng, Thomas Kollar, David Held, and Ken Goldberg. "AutoBag: Learning to Open Plastic Bags and Insert Objects". In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2023.

[3] Lawrence Yunliang Chen*, Huang Huang*, Ellen Novoseller, **Daniel Seita**, Jeffrey Ichnowski, Michael Laskey, Richard Cheng, Thomas Kollar, and Ken Goldberg. "Efficiently Learning Single-Arm Fling Motions to Smooth Garments". In: *International Symposium on Robotics Research (ISRR)*. 2022.

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

[5] Wei Chow*, Jiageng Mao*, Boyi Li, **Daniel Seita**, Vitor Guizilini, and Yue Wang. "PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding". In: *International Conference on Learning Representations (ICLR)* (2025).

[6] Open X-Embodiment Collaboration. "Open X-Embodiment: Robotic Learning Datasets and RT-X Models". In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2024.

[7] Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, Brian Ichter, Danny Driess, Jiajun Wu, Cewu Lu, and Mac Schwager. "Foundation Models in Robotics: Applications, Challenges, and the Future". In: *arXiv preprint arXiv:2312.07843* (2023).

[8] Aditya Ganapathi, Priya Sundaresan, Brijen Thananjeyan, Ashwin Balakrishna, **Daniel Seita**, Jennifer Grannen, Minho Hwang, Ryan Hoque, Joseph Gonzalez, Nawid Jamali, Katsu Yamane, Soshi Iba, and Ken Goldberg. "Learning Dense Visual Correspondences in Simulation to Smooth and Fold Real Fabrics". In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2021.

[9] Sicheng He, Zeyu Shangguan, Kuanning Wang, Yongchong Gu, Yuqian Fu, Yanwei Fu, and **Daniel Seita**. "Sequential Multi-Object Grasping with One Dexterous Hand". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2025).

[10] Ryan Hoque*, **Daniel Seita***, Ashwin Balakrishna, Aditya Ganapathi, Ajay Tanwani, Nawid Jamali, Katsu Yamane, Soshi Iba, and Ken Goldberg. "VisuoSpatial Foresight for Multi-Step, Multi-Task Fabric Manipulation". In: *Robotics: Science and Systems (RSS)*. 2020.

[11] Ryan Hoque*, **Daniel Seita***, Ashwin Balakrishna, Aditya Ganapathi, Ajay Tanwani, Nawid Jamali, Katsu Yamane, Soshi Iba, and Ken Goldberg. "VisuoSpatial Foresight for Physical Sequential Fabric Manipulation". In: *Autonomous Robots*. 2021.

[12] Haodi Hu, Feifei Qian†, and **Daniel Seita**†. "Learning Granular Media Avalanche Behavior for Indirectly Manipulating Obstacles on a Granular Slope". In: *Conference on Robot Learning (CoRL)* (2024).

[13] Haodi Hu, Yue Wu, **Daniel Seita**†, and Feifei Qian†. "Granular Loco-manipulation: Repositioning Rocks and Boulders Through Strategic Sand Avalanche using a Locomoting, Multi-legged Robot". In: *Conference on Robot Learning (CoRL)* (2025).

[14] Minho Hwang, Jeffrey Ichnowski, Brijen Thananjeyan, **Daniel Seita**, Samuel Paradis, Danyal Fer, Thomas Low, and Ken Goldberg. "Automating Surgical Peg Transfer: Calibration with Deep Learning Can Exceed Speed, Accuracy, and Consistency of Humans". In: *IEEE Transactions on Automation Science and Engineering (TASE)*. 2022.

[15] Hao Jiang, Yuhai Wang†, Hanyang Zhou†, and **Daniel Seita**. "Learning to Singulate Objects in Packed Environments using a Dexterous Hand". In: *International Symposium on Robotics Research (ISRR)* (2024).

[16] Yunshuang Li, Yiyang Ling, Gaurav Sukhatme, and **Daniel Seita**. "Learning Geometry-Aware Nonprehensile Pushing and Pulling with Dexterous Hands". In: *arXiv preprint arXiv:2509.18455* (2025).

[17] Vincent Lim*, Huang Huang*, Lawrence Yunliang Chen, Jonathan Wang, Jeffrey Ichnowski, **Daniel Seita**, Michael Laskey, and Ken Goldberg. "Planar Robot Casting with Real2Sim2Real Self-Supervised Learning". In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2022.

[18] Yiyang Ling*, Karan Owalekar*, Oluwatobiloba Adesanya, Erdem Bıyık, and **Daniel Seita**. "IMPACT: Intelligent Motion Planning with Acceptable Contact Trajectories via Vision-Language Models". In: *arXiv preprint arXiv:2503.10110* (2025).

[19] I-Chun Arthur Liu, Jason Chen, Gaurav Sukhatme, and **Daniel Seita**. "D-CODA: Diffusion for Coordinated Dual-Arm Data Augmentation". In: *Conference on Robot Learning (CoRL)* (2025).

[20] I-Chun Arthur Liu, Sicheng He, **Daniel Seita**†, and Gaurav Sukhatme†. "VoxAct-B: Voxel-Based Acting and Stabilizing Policy for Bimanual Manipulation". In: *Conference on Robot Learning (CoRL)* (2024).

[21] Alberta Longhini, Yufei Wang, Irene Garcia-Camacho, David Blanco-Mulero, Marco Moletta, Michael Welle, Guillem Alenyà, Hang Yin, Zackory Erickson, David Held, Júlia Borràs, and Danica Kragic. "Unfolding the Literature: A Review of Robotic Cloth Manipulation". In: *arXiv preprint arXiv:2407.01361* (2024).

[22] Vedant Raval*, Enyu Zhao*, Hejia Zhang, Stefanos Nikolaidis, and **Daniel Seita**. "GPT-Fabric: Smoothing and Folding Fabric by Leveraging Pre-Trained Foundation Models". In: *International Symposium on Robotics Research (ISRR)* (2024).

[23] **Daniel Seita**, Pete Florence, Jonathan Tompson, Erwin Coumans, Vikas Sindhwani, Ken Goldberg, and Andy Zeng. "Learning to Rearrange Deformable Cables, Fabrics, and Bags with Goal-Conditioned Transporter Networks". In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2021.

[24] **Daniel Seita**, Aditya Ganapathi, Ryan Hoque, Minho Hwang, Edward Cen, Ajay Kumar Tanwani, Ashwin Balakrishna, Brijen Thananjeyan, Jeffrey Ichnowski, Nawid Jamali, Katsu Yamane, Soshi Iba, John Canny, and Ken Goldberg. "Deep Imitation Learning of Sequential Fabric Smoothing From an Algorithmic Supervisor". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2020.

[25] **Daniel Seita***, Nawid Jamali*, Michael Laskey*, Ron Berenstein, Ajay Kumar Tanwani, Prakash Baskaran, Soshi Iba, John Canny, and Ken Goldberg. "Deep Transfer Learning of Pick Points on Fabric for Robot Bed-Making". In: *International Symposium on Robotics Research (ISRR)*. 2019.

[26] Zhonghao Shi, Enyu Zhao, Nathaniel Dennler, Jingzhen Wang, Xinyang Xu, Kaleen Shrestha, Mengxue Fu, **Daniel Seita**, and Maja Matarić. "HRIBench: Benchmarking Vision-Language Models for Real-Time Human Perception in Human-Robot Interaction". In: *International Symposium on Experimental Robotics (ISER)* (2025).

[27] Siddharth Srikanth, Freddie Liang, Ya-Chuan Hsu, Varun Bhatt, Shihan Zhao, Henry Chen, Bryon Tjanaka, Minjune Hwang, Akanksha Saran, **Daniel Seita**†, Aaquib Tabrez†, and Stefanos Nikolaidis†. "Red-Teaming Vision-Language-Action Models via Quality Diversity Prompt Generation for Robust Robot Policies". In: *Under Review* (2026).

[28] Sashank Tirumala*, Thomas Weng*, **Daniel Seita***, Oliver Kroemer, Zeynep Temel, and David Held. "Learning to Singulate Layers of Cloth Using Tactile Feedback". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2022.

[29] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. "Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2017.

[30] Kuanning Wang, Yongchong Gu, Yuqian Fu, Zeyu Shangguan, Sicheng He, Xiangyang Xue, Yanwei Fu, and **Daniel Seita**. "SCOOP'D: State-based Sim2Real Generative Policy for Generalizable Mixed-Liquid-Solid Scooping". In: *arXiv preprint arXiv:2510.11566* (2025).

[31] Harry Zhang, Jeff Ichnowski, **Daniel Seita**, Jonathan Wang, Huang Huang, and Ken Goldberg. "Robots of the Lost Arc: Self-Supervised Learning to Dynamically Manipulate Fixed-Endpoint Cables". In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2021.

[32] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware". In: *Robotics: Science and Systems (RSS)*. 2023.

[33] 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". In: *Conference on Robot Learning (CoRL)* (2025).

[34] Hanyang Zhou*, Haozhe Lou*, Wenhao Liu*, Enyu Zhao, Yue Wang, and **Daniel Seita**. "The MOTIF Hand: A Robotic Hand for Multimodal Observations with Thermal, Inertial, and Force Sensors". In: *International Symposium on Experimental Robotics (ISER)* (2025).