Sensorized Manipulation Challenge Kit for Benchmarking Robotic Manipulation

Ben Abbatematteo[1], Callum Robbins*[2], Keith Sherry*[2], Jitpuwapat Mokkamakkul[1],
Eric Rosen[1], Skye Thompson[1], Matthew Stein[2], George Konidaris[1]

***Abstract*— We introduce the Sensorized Manipulation Challenge Kit (*SMaCK*) for benchmarking robotic manipulation capabilities. Existing sets of physical objects for benchmarking robotic manipulation focus on physical dexterity (e.g. grasping, in-hand manipulation) but do not assess planning ability (e.g. unlocking a box before the handle can be lifted), or goal-directed exploration, critical features of embodied intelligence. *SMaCK* is a collection of puzzle boxes that systematically test manipulation *and* reasoning capabilities and which are cheap and easy to fabricate. The boxes include sensors for measuring object state information (e.g: inertial data, pose, articulated joint state) that can be recorded and used to either evaluate or train a manipulation policy. We also include simulated MuJoCo domains using the CAD descriptions of the boxes, and evaluate a deep reinforcement learning agent on the simplest benchmark. Instructions for building the *SMaCK* and our codebase can be found at `https://github.com/babbatem/smack/`.**

I. INTRODUCTION

As robots become increasingly physically capable, benchmarks will serve an essential role in quantitatively comparing the performance of different approaches to robotic manipulation. Benchmarking is notoriously challenging in manipulation, in particular, as individual research groups typically choose objects and tasks to suit each experiment. This prohibits the analysis of related approaches on a common benchmark, obfuscating the relationship between the approaches.

Manipulation is unique in that it requires the coordination of long sequences of fine motor behavior in order to reach a goal. Solving these problems in unstructured environments requires agents to perceive objects in a scene, reason about their properties, and plan for the future, all while controlling complicated physical interactions. As such, effective manipulation benchmarks must test not only physical dexterity but also an agent's capability for exploration, abstract reasoning, and planning. While many benchmarks have been proposed in the literature [1]–[7], existing object sets fail to satisfy these criteria. Benchmarks must also provide instrumented physical objects in addition to CAD descriptions so that real systems can be tested and meaningful human baselines can be established.

We therefore introduce *SMaCK*, the Sensorized Manipulation Challenge Kit, a set of puzzle boxes that consist of articulated mechanisms that test an agent's ability to plan and execute multi-step manipulation sequences to achieve a goal. The objects sense, record, and report their full state in real-time, allowing quantitative comparisons between algorithms and to human baselines without the limitations of vision-based tracking systems. By adding articulated parts and locking dependencies, the problems systematically scale in *breadth* and *depth*, enabling evaluation of manipulation, exploration, and planning capabilities along two axes of difficulty. The designs scale in size, enabling the comparison between agents of different sizes. We provide physical instantiations of these objects, instructions for their fabrication, CAD files, and simulated learning environments. We also evaluate a standard reinforcement learning agent's ability to interact with the most basic box.

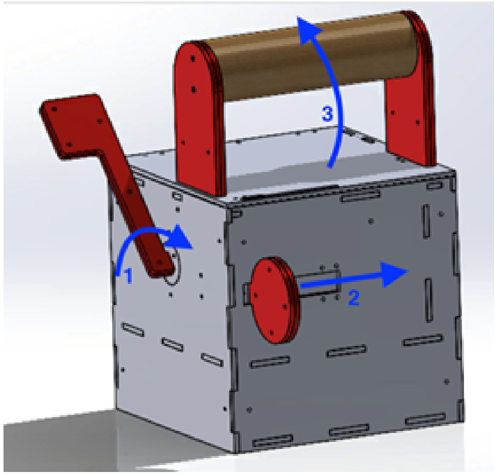

Fig. 1. The proposed puzzle boxes test physical dexterity and planning ability, and sense their full state. The box shown requires a three-step plan to open: rotating a lever, then sliding a locking mechanism, and finally opening the lid.

II. RELATED WORK

There are relatively few existing real-world manipulation benchmarks. While simulation benchmarks offer convenience and a multitude of tasks [8]–[10], they fail to serve as a benchmark for real-world manipulation capabilities, and to facilitate comparisons to humans.

Physical object sets such as the YCB object set [1] and the objects from the Amazon Picking Challenge [2], [3] have been used to evaluate object recognition and grasping performance. Datasets consisting of CAD models and 3D scans have been similarly used to train and evaluate grasping systems, e.g. BigBird [5], KIT [4], the Columbia Grasp Database [7], but typically lack a readily available

* denotes equal contribution.

[1] Department of Computer Science, Brown University, Providence RI

[2] School of Engineering, Computing, and Construction, Roger Williams University, Bristol, RI

set of physical objects to test on. The NIST Manufacturing Objects and Assemblies Dataset (NIST-MOAD) [6] dataset is designed to benchmark assembly tasks typical in manufacturing. The Sensorized Objects benchmark [11] provides rigid objects to benchmark in-hand manipulation performance. Our initial design concepts were based on this work, including the use of the Arduino Nano IOT for its built-in IMU combined and BLE communication features. Our development expanded on these efforts towards the goal of producing stand-alone sensorized objects with clearly defined manipulation objectives. Our work is inspired by the puzzle boxes from the work of Baum et al [12] in which the authors describe a stationary wooden puzzle to be solved by a cockatoo and subsequently a robot by sliding and pulling various levers and handles.

III. SENSORIZED OBJECT DESIGN AND IMPLEMENTATION

Effective manipulation benchmarks must test not only physical dexterity but also cognitive abilities like exploration and planning. They must include physical objects so that robotic systems can be tested and compared to humans; they must also be instrumented such that the state of the objects can be tracked without complicated computer vision or motion capture systems. *SMaCK* consists of several puzzle boxes designed to meet these criteria.

A. Benchmark Task Design

The set of puzzle boxes is designed to test both physical dexterity and abstract reasoning abilities like goal-directed exploration and planning. The goal of each puzzle is to open the lid of box; complexity is added by introducing dependencies between additional articulated parts. This allows experimenters to compare algorithms and agents across a set of tasks that systematically scale in difficulty.

As shown in Figure 1, the puzzle boxes are rigid cubes with optional levels of complexity. A base set of boxes includes models at three levels of complexity: a) box with hinged lid that may be swung open by lifting handle (step 3 in Figure 1), denoted “depth 1 box” b) box with slider that must be slid to the right, unlocking the lid (step 2 then 3), denoted “depth 2 box” c) box with lever that must be rotated clockwise to permit movement of the slider (steps 1 then 2 then 3), denoted “depth 3 box”. In this way, the *depth* of the planning problem increases as more sequential steps are required to achieve the goal. The three basic benchmark tasks are then:

- depth 1 box (basic)
- depth 2 box (slider) (Fig. 2 right).
- depth 3 box (slider + lever) (Fig. 2 middle)

We can also increase the *breadth* of each planning problem by introducing “distractor” parts - for example, adding an additional lever opposite the original one with no function, or adding a “distractor” slide to the front of the object that does not serve to lock the box. This increases the effective action space of the agent and serves as an additional axis along which we may evaluate performance. One such box is pictured in Figure 2 (left)—one lever locks the sliding mechanism (which in turn locks the lid) and one lever does nothing. The agent must determine which of the parts are actually functional and the dependencies that arise in order to succeed in opening the lid. By adding articulated parts without locking mechanisms, we can add breadth to the problems, introducing several additional benchmark tasks:

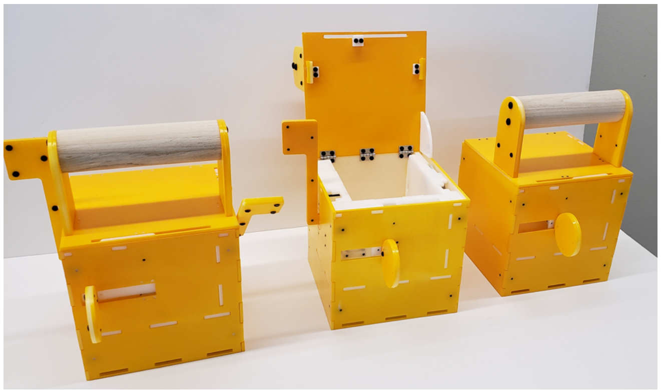

Fig. 2. Completed *SMaCK* boxes (right to left) slider, slider + lever, slider + lever + dummy lever. Boxes are painted yellow for improved camera visibility.

- depth 1 box (basic) + distractor slide (breadth 2)
- depth 1 box (basic) + distractor lever (breadth 2)
- depth 1 box (basic) + distractor slide + distractor lever (breadth 3)
- depth 1 box (basic) + 2 distractor levers (breadth 3)
- depth 1 box (basic) + distractor slide + 2 distractor levers (breadth 4)
- depth 2 box (slider) + distractor lever (breadth 3)
- depth 2 box (slider) + 2 distractor levers (breadth 4)
- depth 3 box (slider + lever) + distractor lever (breadth 4) (Fig. 2 left)

This allows us to create 8 additional tasks from the original three by adding distractor parts and disabling some locking mechanisms.

B. Design and Fabrication

A research objective is to compare quantitative performance measurement of goal-directed behavior by humanoid robots to that of other intelligent agents. As one population for comparison is humans, the manipulation objects must be scalable in proportion to varying hand size. A child’s hand is about one-half the size of an adult human hand. In contrast, the Brown University humanoid manipulates objects using Schunk Grippers, roughly twice the size of human hands. The sensorized objects must be scalable both in size and durability to permit exploratory manipulation by a variety of hands.

We designed stand-alone puzzle boxes to meet the various requirements of durability, scalability and clear manipulation objective. Boxes are self-contained objects with the clear manipulation objective of opening their lid. Boxes may be picked up, shaken, and manipulated, all the while sensing their orientation and the positions of their moving parts.

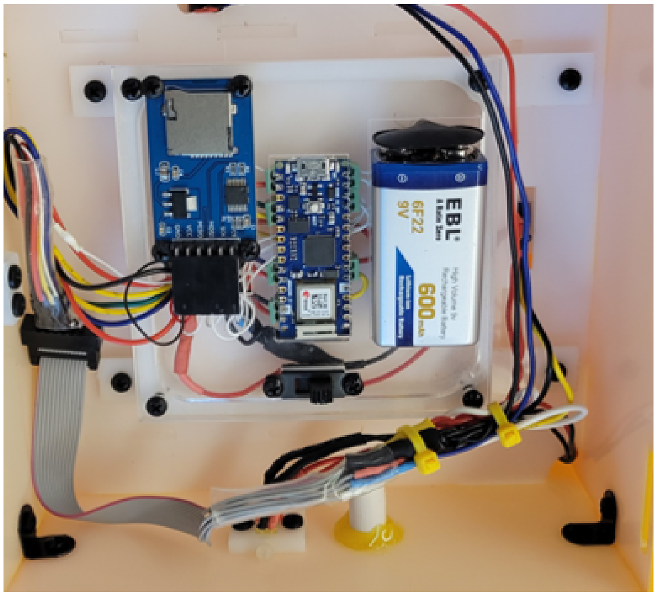

Fig. 3. Electronics insert located in false bottom of all boxes. Central component is the Arduino Nano IoT.

To date we have scaled the designs for human hands, the humanoid grippers at Brown, and the Boston Dynamics Spot manipulator.

Boxes are constructed of one-eighth inch Delrin™[1] sheets laser-cut and hand assembled using 4-40 threaded nylon fasteners and a few metal components such as hinges and support braces. The handle is the only wood component as it light, inexpensive and easily gripped. Delrin is self-lubricating, tough and easily formed into working mechanical components. Boxes are relatively tough, withstanding a drop from shoulder height to a hard floor for example. For stronger than human manipulators, all structural components are specifically designed to be produced from one-eighth inch aluminum sheets on a standard 3-axis mill.

Box lid angle is sensed by a cam depressing a linear potentiometer as shown in the left image of Figure 4. Motion of the slider unlocks the lid and also turns a rotary potentiometer through a rack and pinion mechanism shown in the center image of Figure 2. Rotating the handle unlocks the slider and also rotates a potentiometer through a meshing gear pair as shown in the right image of Figure 2. Total cost for mechanical components is approximately $110 requiring two 1/8x12x24in Delrin sheets at $40 each, fasteners and assorted hardware.

As shown in Figure 2, boxes are constructed with interior walls to protect sensing electronics. Each box has a false bottom containing an electronics insert, as shown in Figure 3. The core of the insert is an Arduino Nano IoT equipped with 8, 10-bit A/D channels, built-in LSM6DS3 IMU and Nina W102 uBlox Bluetooth communication module. Potentiometers measuring lid, slider and lever positions are powered by the Nano and read using the analog input channels. The LSM6DS3 IMU provides box roll, pitch and yaw relative to gravity orientation. BLE communication permits real-time transmission of one-byte IMU and sensor readings at 8HZ to moderately close receivers, for example in the same room, with proportionate degradation in data rate with distance. The unit is powered by a single 9V battery that, in our testing, lasts for several weeks. An SD Card permits data recording at 100Hz for later transfer. Experimentation is controlled through a magnetic sensor and LED displays, where an experimenter can initiate a test by placing a magnetic pendant on the sensor. On initiating a test, an LED will flash a countdown sequence to permit later synchronization between sensor data and potentially a video of the manipulation. Total cost for electrical components is approximately $95 including Arduino Nano, instrumentation potentiometers and SD card reader.

Publicly available documents at https://github.com/babbatem/smack/ provide detailed fabrication instructions. The document includes templates both in SolidWorks Drawing and PDF format for immediate use by a laser cutter. Cutting and etching the parts from Delrin sheets requires about three hours on our 50W Epilog Fusion Laser Cutter. A texture is etched on the Delrin to improve the adherence of spray paint. Completed parts can be assembled in about five hours using hand tools. A parts list and circuit diagram for the electronic insert is also included in the documentation. Hand assembly requires about eight hours using soldering and wire management tools.

IV. Simulated Learning Environments

The CAD files were converted to URDF format and imported into the MuJoCo simulator [13]. Learning environments were implemented through the ROBOSUITE benchmark suite [9] allowing for different robots, observation modes, and controllers. Locking physics are implemented by modulating joint stiffness as the CAD files contain penetrating geometries. Thus far, the three basic benchmark tasks are implemented.

A reinforcement learning agent was trained to complete the simplest benchmark task, opening the depth 1 basic box's lid. The learning algorithm was TD3 [14]. The observation space consisted of the agent's proprioceptive state (joint positions, velocities, tactile data, force/torque data from the wrist) as well as low-dimensional object state data (the angle of the lid). The action space employed was operational space control with variable impedance [15], allowing the agent to learn to control the relative displacement and orientation of its end-effector and regulate its stiffness. The reward function incentivized opening the lid and reaching for the handle. The success rate, aggregated over 10 seeds, is plotted in Figure 5. After 5000 episodes, the agent is only 30% successful. Preliminary experiments with the more complicated boxes suggest that these problems are prohibitively challenging for naive reinforcement learning approaches. Humans, in contrast, are typically able to solve the puzzles in under a minute. Future work will investigate hierarchical reinforcement learning approaches and model-based methods for

[1] Delrin® is the brand name for polyoxymethylene (POM)

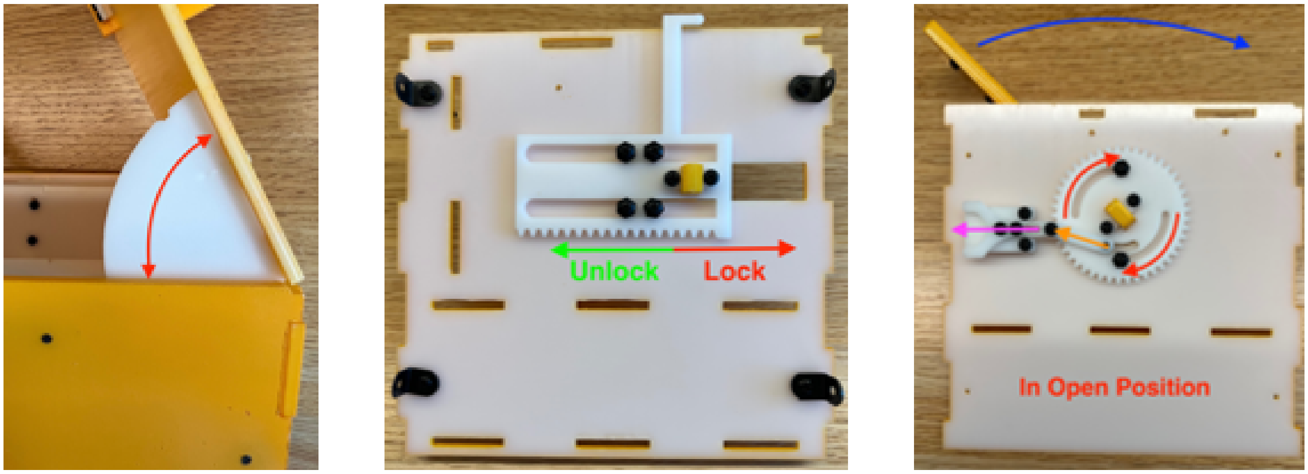

Fig. 4. Sensor mechanisms of boxes, (left to right) lid cam, slider rack/pinion, level gear mesh.

perceiving and manipulating the objects in order to establish competent baselines.

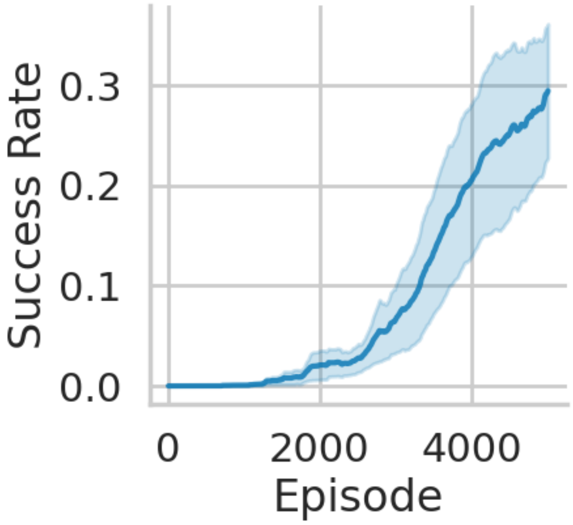

Fig. 5. Success rate for the basic box opening task, smoothed and aggregated over 10 seeds.

V. SUMMARY

SMaCK, the Sensorized Manipulation Challenge Kit, is a robot manipulation benchmark that poses constrained manipulation problems that require goal-directed exploration and planning, which systematically scale in difficulty. Future work will quantify the performance of humans and robots on these tasks to establish a baseline.

In our testing, the one-byte potentiometer readings of lid, slider and lever positions reflect the time history with anticipated accuracy. IMU readings appear reliable for modest handling of the box but can be confused with aggressive handling, (i.e. shaking or flipping) and are subject to some drifting over time. For example, when tilting and then returning the box to its original location the IMU will show a net change in orientation. We have constructed more than ten operational boxes out of Delrin, but to date have not attempted aluminum construction.

ACKNOWLEDGMENTS

This research was supported in part by ONR grants N00014-21-1-2584 and N00014-22-1-2592, NSF CAREER award #1844960, and an Amazon Research Award and to Konidaris. We would like to thank Eliza Nelson for her helpful insights.

REFERENCES

[1] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, “Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set,” *IEEE Robotics & Automation Magazine*, vol. 22, no. 3, pp. 36–52, 2015.

[2] N. Correll, K. E. Bekris, D. Berenson, O. Brock, A. Causo, K. Hauser, K. Okada, A. Rodriguez, J. M. Romano, and P. R. Wurman, “Analysis and observations from the first amazon picking challenge,” *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 1, pp. 172–188, 2016.

[3] C. Eppner, S. Höfer, R. Jonschkowski, R. Martín-Martín, A. Sieverling, V. Wall, and O. Brock, “Lessons from the amazon picking challenge: Four aspects of building robotic systems.” in *Robotics: science and systems*, 2016, pp. 4831–4835.

[4] A. Kasper, Z. Xue, and R. Dillmann, “The kit object models database: An object model database for object recognition, localization and manipulation in service robotics,” *The International Journal of Robotics Research*, vol. 31, no. 8, pp. 927–934, 2012.

[5] A. Singh, J. Sha, K. S. Narayan, T. Achim, and P. Abbeel, “Bigbird: A large-scale 3d database of object instances,” in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014, pp. 509–516.

[6] K. Kimble, J. Albrecht, M. Zimmerman, and J. Falco, “Performance measures to benchmark the grasping, manipulation, and assembly of deformable objects typical to manufacturing applications,” *Frontiers in Robotics and AI*, vol. 9, 2022.

[7] C. Goldfeder, M. Ciocarlie, H. Dang, and P. K. Allen, “The columbia grasp database,” in *2009 IEEE international conference on robotics and automation*. IEEE, 2009, pp. 1710–1716.

[8] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, “Rlbench: The robot learning benchmark & learning environment,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3019–3026, 2020.

[9] Y. Zhu, J. Wong, A. Mandlekar, R. Martín-Martín, A. Joshi, S. Nasiriany, and Y. Zhu, “robosuite: A modular simulation framework and benchmark for robot learning,” *arXiv preprint arXiv:2009.12293*, 2020.

[10] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, “Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning,” in *Conference on robot learning*. PMLR, 2020, pp. 1094–1100.

[11] G. Gao, G. Gorjup, R. Yu, P. Jarvis, and M. Liarokapis, “Modular, accessible, sensorized objects for evaluating the grasping and manipulation capabilities of grippers and hands,” *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6105–6112, 2020.

[12] M. Baum, M. Bernstein, R. Martin-Martin, S. Höfer, J. Kulick, M. Toussaint, A. Kacelnik, and O. Brock, “Opening a lockbox through physical exploration,” in *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*. IEEE, 2017, pp. 461–467.

[13] E. Todorov, T. Erez, and Y. Tassa, “Mujoco: A physics engine for model-based control,” in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 5026–5033.

[14] S. Fujimoto, H. Hoof, and D. Meger, “Addressing function approximation error in actor-critic methods,” in *International conference on machine learning*. PMLR, 2018, pp. 1587–1596.

[15] R. Martín-Martín, M. A. Lee, R. Gardner, S. Savarese, J. Bohg, and A. Garg, “Variable impedance control in end-effector space: An action space for reinforcement learning in contact-rich tasks,” in *2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2019, pp. 1010–1017.