FMB: a Functional Manipulation Benchmark for Generalizable Robotic Learning

Anonymous Author(s)
Affiliation
Address
`email`

Abstract: In this paper, we propose a benchmark for studying robotic learning for functional manipulation. We identify handling complex contact dynamics and generalization as two central challenges in practical robotic manipulation. While many prior works have addressed one challenge or the other, few have studied both in combination. We hypothesize that making progress on the combination of these challenges requires a set of real-world benchmark tasks that balance complexity with accessibility, providing a set of tasks that are sufficiently narrowly scoped that models and datasets of reasonable scale can be used to make progress, but sufficiently varied that they present a meaningful generalization challenge not just in terms of basic and imprecise skills such as grasping, but also more complex and precise behaviors that require functional manipulation, such as repositioning and reorienting an object for a precise assembly task. Our functional manipulation benchmark consists of a variety of 3D printed objects that can be reproduced precisely by other researchers, each one requiring a sequence of grasping, reorientation, and assembly behaviors. Generalization can be evaluated on test objects and varied positions, as well as more complex multi-stage assembly tasks. We also provide an imitation learning system that provides a basic set of policies for each skill, allowing researchers to use our tasks as a toolkit for studying any portion of the pipeline – for example by proposing a better design for a grasping controller and evaluating it in combination with our baseline reorientation and assembly controllers. Our dataset, object CAD files and evaluation videos can be found on our project website: https://sites.google.com/view/manipulationbenchmark

Keywords: manipulation, imitation learning, benchmarking

1 Introduction

Manipulation is one of the foundational problems in robotics research, but enabling robots to perform dexterous manipulation skills that reflect the capabilities of humans is still out of reach. In fact, even matching the performance of human *teleoperation* remains a major challenge, particularly in environments that require generalization and are not constrained to a specific fixed set of well-characterized objects. As Cui and Trinkle [1] point out, two primary sources of difficulty in robotic manipulation lie in handling complex contact mechanics and intelligently handling the variability in the environment and objects. While robotic learning techniques hold potential to address these challenges, effective progress will require tasks that are accessible enough for current methods while still exposing the key challenges of complex contact mechanics and object generalization.

While significant recent research in robotic learning has made progress on various aspects of the manipulation problem [2, 3], much of the emphasis on recent works has either been on broad generalization with relatively simple tasks, which often do not capture the many physical challenges of manipulation (e.g., focusing on picking or imprecise pick-and-place tasks) [4], or else training policies for narrow tasks that are physically more complex but not do demand extensive generalization [5].

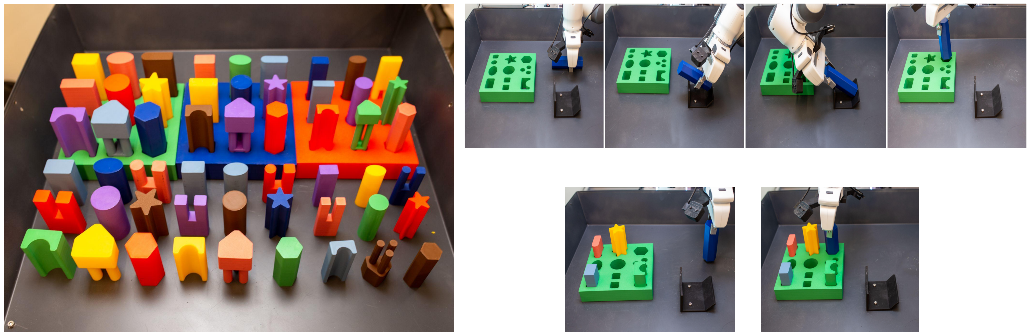

Figure 1: **Left:** The 3D-printed parts for the simpler insertion tasks. **Right:** An illustration of the steps for inserting a single part, which requires grasping the part, reorienting it (potentially using an environment fixture), and then inserting it into the appropriate slot. Note that the full task requires grasping, reorientation, and insertion to be performed in concert.

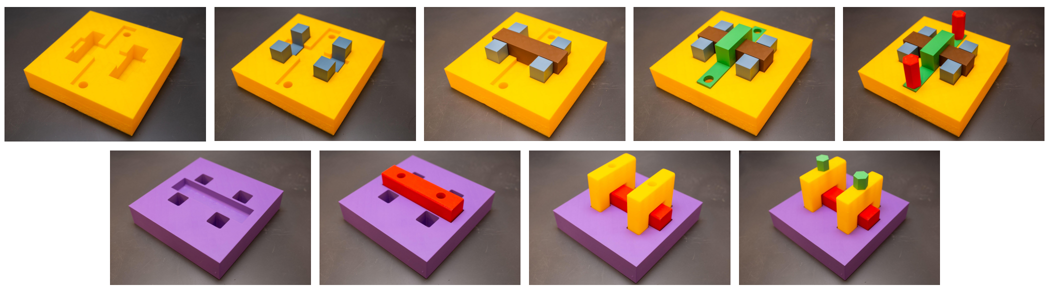

Figure 2: Two instantiations of the complex assembly task. These tasks require similar functional manipulation behaviors as the simpler set of tasks, but with multiple interlocking objects and a more complex higher-level structure that requires assembling the parts in the right order.

This is not unreasonable: it is very difficult to simultaneously make progress on broad generalization (which often requires huge datasets), and tackle the full physical complexity of dexterous manipulation. So how can we take a step toward facilitating robotic learning research that emphasizes both generalization and physically intricate skills, while still keeping the problem constrained enough so as to enable meaningful progress?

In this paper, we propose a family of benchmarks that aims to cover the important dimensions of physical complexity and object generalization, while still providing a degree of accessibility by carefully restricting the scope to a domain where we can make progress with reasonable sized datasets and models. We approach the design of this benchmark by defining functional object manipulation as the problem of picking up an object in a functionality relevant way, positioning it in an appropriate pose, and then using it for a physical interaction. While this definition is more restrictive, we believe it captures a broad range of practical manipulation tasks, and includes both the challenges of complex contact dynamics and object generalization.

The specific tasks we instantiate to capture functional manipulation are themed around assembly problems, including simpler pick-and-place tasks and more complex multi-part assemblies. These tasks, illustrated in Fig. 1, require picking up the individual pieces, reorienting them (potentially using environment affordances and regrasping), and then slotting them into their required location. Each phase requires addressing both the challenge of complex contacts and the challenge of generalization. The objects may vary between training and test-time, and their locations are randomized. The grasping phase requires selecting a grasp that is suitable for reorienting the object, the reorientation phase requires positioning the object so that contact with the environment changes its pose in the desired way, and the assembly phase requires compliant insertion and proper accounting for the contact forces on the object. Each phase requires handling different objects (including new test-time objects) and different poses. The robotic assembly task has been long seen as a representative manipulation benchmark [6, 7, 8, 9]; however, the generalization effect across such tasks has been less studied comprehensively. Thus, performing such tasks among the pool of diverse shapes would be an ideal candidate for benchmarking generalizable dexterity.

To ensure reproducibility and portability of our benchmark task, we use 54 3D-printed objects with diverse shapes and sizes that can be reproduced by other researchers, and a widely used Franka robotic arm. We collected a dataset of 9000 human demonstrations of grasping, repositioning, and inserting these objects, and trained a baseline imitation learning system to perform each stage of the task. Our dataset also contains a variety sensory modalities as presented in Fig. 3: we record RGB and depth images from eye-in-hand and eye-to-hand views. These make our environment modular so that other researchers can repurpose it for a variety of methods that they may wish to develop, and can focus on any stage or aspect of the task. For example, some researchers might choose to focus on better functional grasping methods, while the other stages are handled by our baseline system, while others might focus on compliant insertion, utilizing our baseline system for the grasping stage. Our tasks are also designed to accommodate pretraining with finetuning to other downstream behaviors. To this end, we also provide a set of more complex assembly objects, as shown in Fig. 2, which can be handled by policies adapted from pretraining on the main dataset. We describe our benchmark tasks, and conduct a comprehensive evaluation studying the performance of imitation learning methods trained on our data, evaluating both training object and test set performance. Our hope is that our functional manipulation benchmark (FMB) will provide a toolkit for robotic learning researchers to study manipulation both in terms of complex contact dynamics and generalization.

2 Related Work

Considerable recent progress on robotic manipulation has studied generalization, though often in the context of simpler tasks such as grasping [10, 2], pushing [10], and imprecise repositioning [10]. A number of other works have studied tasks that are dynamic [11], precise (e.g., insertion) [12], or otherwise physically challenging [5]. However, few works have studied these factors in combination. We believe many of the central challenges in robotic manipulation lie at the confluence of these two challenges: tasks that require handling complex contact dynamics, not by memorizing the particular pattern needed for a single narrow task, but by learning general behaviors for handling object interaction that can generalize to new objects. Our aim is to propose a benchmark that can study this combination of challenges, while keeping the scope narrow enough that it remains accessible to many researchers.

Our tasks combine aspects of grasping, repositioning, and peg insertion or assembly. A number of works have studied these individual stages [2, 13]. Our goal is not to attain the best possible performance in narrow settings for any of these stages (e.g., ultra-high-precision industrial insertion), but to use these tasks as a lens through which to gauge general manipulation capabilities learned via general-purpose robotic learning methods.

A number of prior works have proposed datasets for robotic learning, including datasets consisting of demonstrations [4] and autonomously collected data [2, 14], as well as annotated datasets of grasp points [15], object geometries [16, 17], and simulated environments [18]. However, there has been comparatively little work on standard and accessible object sets that are combined with multi-stage tasks for studying generalization. The YCB object set [19] comes with a number of evaluation protocols [19], but these protocols generally focus on object repositioning tasks that do not evaluate the complex contacts challenges that we discuss in the previous section. A number of existing demonstration datasets cover many different behaviors [4, 20], but also focus on behaviors that emphasize basic pick-and-place skills rather than precise or contact-rich manipulation. Some works have focused on insertion skills in particular (e.g., connector insertion) [21]. While our benchmark is related, we aim specifically to cover a range of skills, including grasping and repositioning, that we believe cover a basis of basic manipulation capabilities. We also emphasize generalization as a primary challenge for our benchmark.

We use 3D printed objects to facilitate reproducibility. Other prior works have also proposed standard meshes and 3D printed parts for benchmarking and reproducibility [19], typically focusing on object

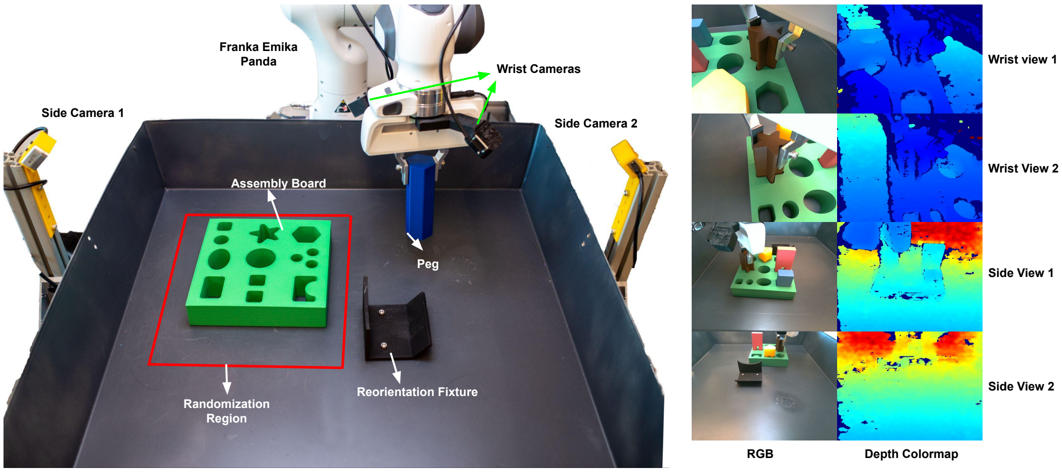

Figure 3: Illustration of the robot setup, with a standard Franka arm equipped with four cameras (two on the wrist and two attached to the environment), each with RGB and depth, positioned in front of a workspace containing an object, reorientation fixture, and assembly board. The board is placed into a random pose within the randomization region, and the object is located in a randomized pose on the table, from where it must be picked up, reoriented, and inserted.

grasping. These efforts are related, but our aim is to provide parts that are specifically well suited for evaluating all of the stages: grasping, reorientation, and assembly, rather than only grasping.

3 Functional Manipulation Benchmark

In this section, we introduce the basic principles behind FMB and the protocols to evaluate different methods on this benchmark. We are mostly concerned with studying the generalization of each individual functional manipulation task as well as the combinatorial ways of composing them to achieve novel behaviors. Therefore we collect a diverse dataset of robotic behaviors with different objects, viewpoints, and robot initial poses. We also additionally provide novel objects for the purpose of benchmarking the generalization capability of individual skills, as well as the ability for a method to compose these skills to solve unseen long-horizon tasks.

3.1 Object Set

We designed 54 3D-printed objects of different sizes, shapes, and colors, with examples shown in Figure 1.

In total, we have 9 different basic shapes, and for each shape there are 6 different sizes. The parts are assigned 8 different random colors. There are three boards with matching holes for the objects. We additionally designed two more complex boards to facilitate multi-stage assembly tasks, shown in Figure 2, where multiple parts must be fitted together. The tolerance for mating these objects is consistently 1mm to 1.5mm. All of our CAD files including those for environment fixtures and camera mount are publicly available on our project website.

3.2 Functional Manipulation Tasks

In this section, we describe the individual tasks that we propose to evaluate with our benchmark. For each type of tasks, we provide demonstration trajectories collected with a Franka robot (see Figure 3), and an evaluation protocol. The modular design of our benchmark facilitates extension to add new tasks with the provided objects, but the tasks we describe here are suitable both for evaluating generalization and for testing a range of manipulation capabilities.

Grasping. The grasping task in our benchmark is a *functional* grasping task, in the sense that the robot must grasp the object in a way that facilitates downstream reorientation, rather than simply picking the object in any pose. We illustrate this task in Fig. 4. A top-down grasp is reasonable if the object is placed in a vertical pose, as shown on the right side of Fig. 4. However, a horizontal grasp is

Figure 4: Objects may need to be grasped from a variety of poses, particularly when using the reorientation fixture, where they might lie at an angle.

much more desirable if the object is positioned as on the left side of Fig. 4. In case such a grasp is infeasible due to the robot's kinematic constraints, the robot needs to perform additional repositioning steps to adjust the feasible grasp pose. The robot must learn grasping skills that deploy the appropriate grasp for the object's current configuration, and also generalize across different object shapes, colors, and sizes. Our demonstration dataset for the grasping task consists of 50 trajectories per object, with varying object rest poses in the bin, for a total of 2700 trajectories performing functional grasping over the 54-object set.

Repositioning. A repositioning step is sometimes necessary to adjust the grasping pose so that the object is held in a way that is suitable for downstream assembly. Manipulating and reorienting objects by leveraging environment affordances (e.g., tilting the object in the gripper by levering it against a table or wall) may often be necessary for fluent and complex manipulation, and this reorientation task exercises this capability. We provide a simple fixture that can serve as an environment affordance to rest the object at angle, as shown in Fig. **??**. To reorient the objects into the right pose, the robot may need to use this fixture, resting the object on it and then regrasping it in a more appropriate pose for reorientation. We collected 3000 demonstrations for placing and regrasping, which can be used to learn strategies for using environmental affordances for regrasping and reorientation. Since objects land in the fixture in a relatively deterministic fashion, we partially script our demonstration collection process while maintaining a certain degree of randomness for the purpose of data diversity.

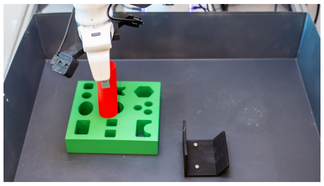

Figure 5: An illustration of the assembly task.

Assembly. Our assembly tasks consider assembling objects of diverse shapes into their matching slots, which requires performing fine-grained precise manipulation. An illustrative example is shown in Fig. 5. Here, having completed the preceding two steps, the robot is holding an object, and needs to insert it into the matching slot in the blue board. For each object, we collect 50 human demonstrations that include various robot initial poses and board positions, for a total of 3000 demonstrations performing the assembly task from various initial conditions. Note that the board is located in different places on the table for different episodes, requiring a reactive strategy that localizes the board and the appropriate opening, and guides the object into the correct location.

Long-horizon manipulation. Aside from performing individual steps, such as grasping, reorientation, and assembly, our benchmark and demonstrations can be used to learn the entire long-horizon sequence, performing the steps in turn to insert one or multiple objects into the board. The difficulty of this task mainly comes from the compounding errors accumulated over each individual step which gets even more magnified when switching between tasks.

Multi-step interlocking assembly. We also present two sets of novel objects for benchmarking much broader generalization capability. The pieces in Fig. 2 are largely different from our original set of objects, and would require adaptation to perform grasping or insertion which can be achieved by pretraining on the collected dataset and finetuning on a few new demonstrations. The major challenge with these tasks is that these objects need to be put together in a specific order, such as in an interlocking fashion. While it may not be too hard to perform individual steps alone, the difficulty increases rapidly when a policy needs to simultaneously reason the manipulation sequence as well as accounting for compounding manipulation errors introduced by individual steps.

3.3 Robotic system and data collection

We now describe the robotic system and the process we used to collect the training data.

Robotic system overview Our system can be seen in Fig. 3. We use a Franka Panda robot to collect our dataset, since it is widely adopted for research and offers a torque control interface which is very desirable in contact-rich manipulation tasks. To record demonstrations, we use a SpaceMouse to control the robot at 10 HZ. In total, we have four Intel RealSense D405 cameras, two of which are mounted on the robot end-effector, and the rest are placed on each side of the bin to provide a complementary view of objects in the bin. We concurrently capture RGB and depth images from these cameras.

Data collection protocol. Our dataset consists of 2700 demonstrations for the full long-horizon task of grasping, reorientation, and assembly for these 54 objects. For each object, we collect around 50 demonstrations per task. Each such demonstration trajectory is around 20 to 30 seconds long, and thus it's more practical to break them into individual "primitives" of shorter horizons. In fact, we automatically add indicators at the end of a manipulation skill such as grasping so that we can segment these long-horizon trajectories. In our dataset, these primitives include `grasping`, `reorientation`, `move`, `insertion`; so in that sense, we have 8100 demonstrations of each primitive with horizons around 5 seconds. For the grasping task, the object of interest is randomized around a 20cm x 30cm rectangular area in the bin; whereas for the insertion task, the board is randomized around a 40cm by 60cm area. We also include distractors (i.e. objects not needed for a task) when performing the insertion task, half of the insertion demonstrations were carried out when there are distractors present to gain robustness.

4 Using the FMB in imitation learning

To illustrate the utility of our benchmark in imitation learning. We describe a few example usages of our dataset and the corresponding evaluation protocol. The detailed evaluation protocol and metric can be found on our website `https://sites.google.com/view/manipulationbenchmark`. Although in principle our data can also be easily altered to study other approaches such as offline reinforcement learning.

4.1 Training and Evaluation of Individual Skills

Generally speaking, we expect to see the emergence of generalization by training on a large, diverse dataset. To verify this hypothesis, we refer to two ways of testing generalization. For grasping and insertion, we can hold out a specific object in the training set, train a policy without seeing any data associated with that object, and then test on the held-out object. Alternatively, we also provide five novel objects that are not contained in the dataset for which we can directly evaluate trained policies.

4.2 Pre-training and Finetuning

Pretraining and finetuning visuomotor skills is an open and important research question. By having a large-scale diverse dataset of robot manipulation behaviors, it's possible for us to study this problem. We can pretrain on a set of robot behaviors associated with some objects, and then finetune on data from objects that are not present in the pretraining dataset. If that object is entirely novel, such as the more complex assembly objects in Figure 2, we can collect some additional demonstrations using our setup for finetuning.

4.3 Composing Skills to Solve Long-Horizon tasks

FMB also supports studying long-horizon tasks in various ways: one can train "flat" style imitation learning methods on all the data or hierarchical style methods that trigger individual primitives in some intelligent ways. In additional to the original "grasp-reorient-assembly" task, it's also possible to study more complex novel tasks such as the one shown in Fig. 2, by finetuning on the new objects.

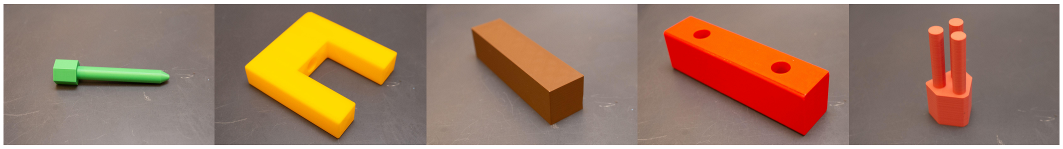

Figure 6: Unseen test objects used for evaluating generalization in our protocol. The robot must generalize to new combinations of shapes, colors, and sizes using the diverse training set.

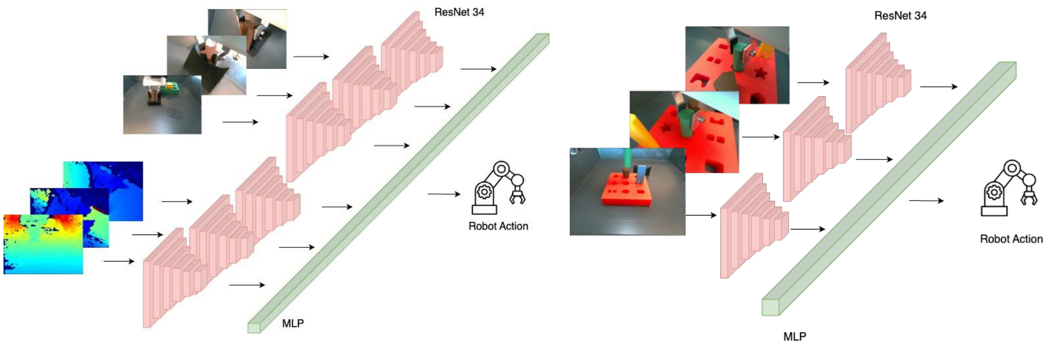

Figure 7: Architecture diagrams for the grasping and reorientation tasks (left), and the assembly task (right). The models encode each observation with a ResNet34 encoder, and then fuse the modalities with fully connected layers.

5 Experiments

In this section, we conduct experiments first to verify our proposed tasks are actually feasible. We then seek to answer the following questions: (1) For each individual task, does training on a diverse manipulation dataset generalize across object properties? (2) When do multi-modal inputs help for which manipulation skill? (3) What are the necessary ingredients for solving long-horizon complex manipulation tasks?

5.1 Grasping Task

To show that it is feasible to perform the grasping tasks using our dataset, we first train a grasping BC policy specifically for the oval object. We obtain 12 successful grasps out of 30 trials, which amounts to 40% success rate. During the evaluation, we test on all six oval objects, performing five trials per each object so that generalization can be fairly tested.

Then we train two grasping BC policies on all the data and test the trained policy on both in-distribution objects as well as novel objects. We present results in Fig. 8. One grasping policy is trained with RGB images, the other one we provide additional depth information; their neural network architecture can be seen in Fig. 7. We find that depth information is crucial in helping achieve better grasping performance.

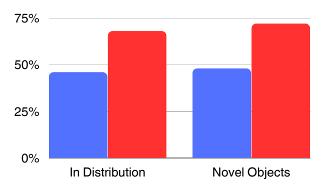

Figure 8: Comparison of grasping success rates on in distribution and novel objects when trained with (Red) and without (Blue) depth information.

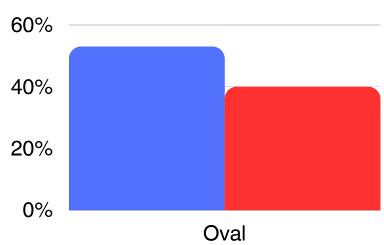

Figure 9: Success rates of grasping oval when training without oval data (Blue) and only oval data (Red).

5.2 Repositioning Task

For the repositioning task, we train BC policies to first place the object on the fixture and then try to re-grasp the object from the other end. The policy's success rate is 0% if trained to solve place and reorient at once with all the data. If we train only on placing data, the policy can achieve 33.3% success rate out of 30 trials; however, the re-grasping policy is 5% success rate trained on corresponding re-grasping data. The failure mode includes missing the object, flipping over the object, and the arm wasn't able to turn over. This is reasonable since we only train policies on visual data without robot state, and the re-grasping part is particularly challenging due to its multi-modal nature.

5.3 Assembly Task

For the assembly task, we first train three separate insertion policies on the shapes of round, hexagon, and two-square. We train these policies on a small portion of our dataset that only contains their corresponding shapes, e.g., the policy to insert round objects is trained on only round data. We also vary each policy's input modality differentiating by depth information. The results can be seen in Fig. 10. We can see the success rate does decrease as the shape becomes more complex, this implies the chosen assembly task is indeed a challenging robotic manipulation task thus worthwhile benchmarking. We also find that naïvely adding depth information doesn't help for the insertion tasks.

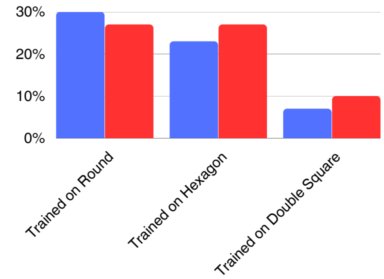

Figure 10: Comparing insertion success rates when training only on one peg data with (Red) and without (Blue) depth then evaluating only on trained peg.

To carry out an initial study on the generalization of training on diverse shapes, we train an insertion BC policy with all the data and test it on the round and star shapes; which we get 4/30 and 0/30 success rates respectively. However, when we train individual policies just with data from that particular shape, we are able to get 9/30 and 0/30. This is reasonable because this task is very precise, naively mixing the data will cause the uni-modal BC model confused so that performance gets hurt. We present this result in Fig. 11.

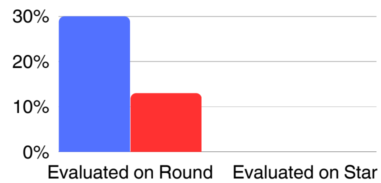

Figure 11: Comparing insertion on specific pegs when training on round with RGB only(Blue) vs training on all data with RGB only (Red).

5.4 Long-Horizon Task

In addition to training BC policies on only primitives, we train an end-to-end long-horizon Behavioral Cloning (BC) policy with all the long horizon demos and transitions. We provide all the RGB camera views for this policy. The goal of this policy is to successfully grasp, reorient, and perform assembly. We evaluate the end-to-end Behavioral Cloning policy on 5 different pegs for a total of 10 trials; this policy achieves a success rate of 0/10. This is reasonable due to the accumulation of errors in long-horizon end-to-end behavioral cloning.

We explore alternatives to naive Behavioral Cloning and modify our approach to include our previously trained grasping and insertion policies. By manually triggering the grasp and insertion policies, as well as using a scripted reorientation motion, we achieve a success rate of 2/10. We present this result in Fig. 12.

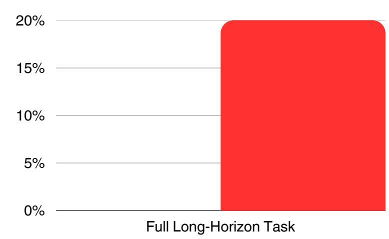

Figure 12: Comparison between end-to-end behavioral cloning (Blue) for grasping, reorientation, and assembly and manually triggered learned skills (Red) and scripted skills

6 Discussion and Limitation

In this paper, we present a benchmark for functional manipulation. We opensource all the data as well as the object CAD files to facilitate reproducibility. We evaluate imitation learning methods with different input modalities and their abilities to generalize across objects. We hope our benchmark FMB would encourage and contribute to in robotic manipulation research.

Limitations and future work. Although our dataset has a variety of diverse objects, we still only have one scene; it would be helpful if we can include more background scenes. Additionally, 3D-printed objects are easy to reproduce, however, it would be more useful if we include real standardized objects in the future so we can study much broader generalization.

References

[1] J. Cui and J. Trinkle. Toward next-generation learned robot manipulation. *Science Robotics*, 6(54), 2021. URL https://www.science.org/doi/abs/10.1126/scirobotics.abd9461.

[2] S. Levine, P. Pastor, A. Krizhevsky, and D. Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection, 2016.

[3] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, and S. Levine. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation, 2018.

[4] F. Ebert, Y. Yang, K. Schmeckpeper, B. Bucher, G. Georgakis, K. Daniilidis, C. Finn, and S. Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021.

[5] OpenAI, I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, J. Schneider, N. Tezak, J. Tworek, P. Welinder, L. Weng, Q. Yuan, W. Zaremba, and L. Zhang. Solving rubik's cube with a robot hand, 2019.

[6] I.-W. Kim, D.-J. Lim, and K.-I. Kim. Active peg-in-hole of chamferless parts using force/moment sensor. *Proceedings 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human and Environment Friendly Robots with High Intelligence and Emotional Quotients (Cat. No.99CH36289)*. doi:10.1109/iros.1999.812802.

[7] W. Newman, Y. Zhao, and Y.-H. Pao. Interpretation of force and moment signals for compliant peg-in-hole assembly. *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No.01CH37164)*. doi:10.1109/robot.2001.932611.

[8] V. Gullapalli, R. Grupen, and A. Barto. Learning reactive admittance control. *Proceedings 1992 IEEE International Conference on Robotics and Automation*. doi:10.1109/robot.1992.220143.

[9] M. Majors and R. Richards. A neural-network-based flexible assembly controller. In *1995 Fourth International Conference on Artificial Neural Networks*, pages 268–273, 1995. doi:10.1049/cp:19950566.

[10] S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, and C. Finn. Robonet: Large-scale multi-robot learning, 2020.

[11] D. Seita, N. Jamali, M. Laskey, A. K. Tanwani, R. Berenstein, P. Baskaran, S. Iba, J. Canny, and K. Goldberg. Deep transfer learning of pick points on fabric for robot bed-making, 2019.

[12] K. Zakka, A. Zeng, J. Lee, and S. Song. Form2fit: Learning shape priors for generalizable assembly from disassembly, 2020.

[13] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics, 2017.

[14] L. Pinto and A. Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours, 2015.

[15] H.-S. Fang, C. Wang, M. Gou, and C. Lu. Graspnet: A large-scale clustered and densely annotated dataset for object grasping, 2020.

[16] S. Tyree, J. Tremblay, T. To, J. Cheng, T. Mosier, J. Smith, and S. Birchfield. 6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark, 2022.

[17] J. J. P, Y.-W. Chao, and Y. Xiang. Fewsol: A dataset for few-shot object learning in robotic environments, 2023.

[18] S. James, P. Wohlhart, M. Kalakrishnan, D. Kalashnikov, A. Irpan, J. Ibarz, S. Levine, R. Hadsell, and K. Bousmalis. Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks, 2019.

[19] 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 &amp Automation Magazine*, 22(3):36–52, sep 2015. doi:10.1109/mra.2015.2448951. URL `https://doi.org/10.1109%2Fmra.2015.2448951`.

[20] A. Mandlekar, J. Booher, M. Spero, A. Tung, A. Gupta, Y. Zhu, A. Garg, S. Savarese, and L. Fei-Fei. Scaling robot supervision to hundreds of hours with roboturk: Robotic manipulation dataset through human reasoning and dexterity, 2019.

[21] G. D. Magistris, A. Munawar, T.-H. Pham, T. Inoue, P. Vinayavekhin, and R. Tachibana. Experimental force-torque dataset for robot learning of multi-shape insertion, 2018.