

# On Bringing Robots Home

Nur Muhammad (Mahi) Shafiullah\*† Anant Rai\* Haritheja Etukuru Yiqian Liu  
NYU NYU NYU NYU

Ishan Misra  
Meta

Soumith Chintala  
Meta

Lerrel Pinto  
NYU

<https://dobb-e.com>

## Abstract

Throughout history, we have successfully integrated various machines into our homes. Dishwashers, laundry machines, stand mixers, and robot vacuums are just a few recent examples. However, these machines excel at performing only a single task effectively. The concept of a “generalist machine” in homes – a domestic assistant that can adapt and learn from our needs, all while remaining cost-effective – has long been a goal in robotics that has been steadily pursued for decades. In this work, we initiate a large-scale effort towards this goal by introducing Dobb-E, an affordable yet versatile general-purpose system for learning robotic manipulation within household settings. Dobb-E can learn a new task with only five minutes of a user showing it how to do it, thanks to a demonstration collection tool (“The Stick”) we built out of cheap parts and iPhones. We use the Stick to collect 13 hours of data in 22 homes of New York City, and train Home Pretrained Representations (HPR). Then, in a novel home environment, with five minutes of demonstrations and fifteen minutes of adapting the HPR model, we show that Dobb-E can reliably solve the task on the Stretch, a mobile robot readily available on the market. Across roughly 30 days of experimentation in homes of New York City and surrounding areas, we test our system in 10 homes, with a total of 109 tasks in different environments, and finally achieve a success rate of 81%. Beyond success percentages, our experiments reveal a plethora of unique challenges absent or ignored in lab robotics. These range from effects of strong shadows to variable demonstration quality by non-expert users. With the hope of accelerating research on home robots, and eventually seeing robot butlers in every home, we open-source Dobb-E software stack and models, our data, and our hardware designs.

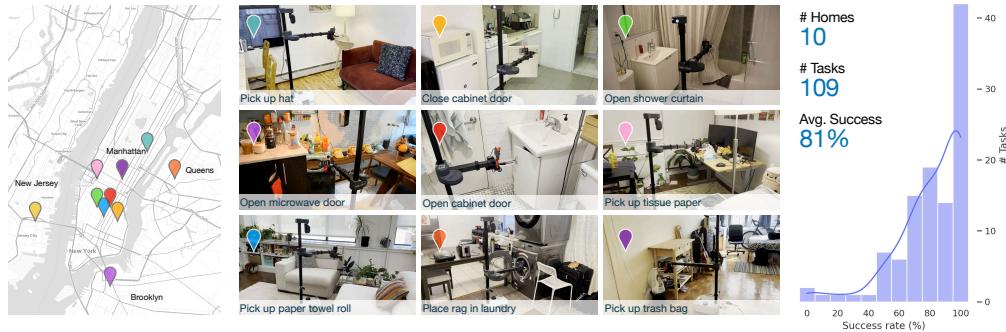


Figure 1: We present Dobb-E, a simple framework to train robots, which is then field tested in homes across New York City. In under 30 mins of training per task, Dobb-E achieves 81% success rates on simple household tasks.

\*Authors contributed equally.

†Corresponding author, email: mahi@cs.nyu.edu

## Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Technical Components and Method</b>	<b>6</b>
2.1	Hardware Design . . . . .	7
2.1.1	Collecting robot demonstrations . . . . .	7
2.1.2	Captured Data Modalities . . . . .	7
2.1.3	Robot Platform . . . . .	7
2.1.4	Camera Mounts . . . . .	8
2.1.5	Gripper Tips . . . . .	8
2.2	Pretraining Dataset – Homes of New York . . . . .	8
2.2.1	Gripper Data . . . . .	9
2.2.2	Dataset Format . . . . .	9
2.2.3	Dataset Quality Control . . . . .	10
2.2.4	Related Work . . . . .	10
2.3	Policy Learning with Home Pretrained Representations . . . . .	10
2.3.1	Visual Encoder Learning . . . . .	11
2.3.2	Downstream Policy Learning . . . . .	11
2.3.3	Related Work . . . . .	11
2.4	Deployment in Homes . . . . .	11
2.4.1	Protocol for Solving Home Tasks . . . . .	12
2.4.2	Policy Training Details . . . . .	12
2.4.3	Robot Execution Details . . . . .	12
2.4.4	Related Work . . . . .	12
<b>3</b>	<b>Experiments</b>	<b>12</b>
3.1	List of Tasks in Homes . . . . .	12
3.2	Understanding the Performance of Dobb-E . . . . .	17
3.2.1	Breakdown by Task Type . . . . .	17
3.2.2	Breakdown by Action Type . . . . .	17
3.2.3	Correlation between demo time and difficulty . . . . .	17
3.3	Failure Modes and Analysis . . . . .	17
3.3.1	Lighting and shadows . . . . .	17
3.3.2	Sensor limitations . . . . .	18
3.3.3	Robot hardware limitations . . . . .	19
3.3.4	Temporal dependencies . . . . .	20
3.4	Ablations . . . . .	21
3.4.1	Alternate visual representation models . . . . .	21
3.4.2	Number of demonstrations required for tasks . . . . .	21

3.4.3	Depth Perception . . . . .	21
3.4.4	Demonstrator Expertise . . . . .	22
3.4.5	Odometry . . . . .	22
<b>4</b>	<b>Open Problems and Request for Research</b>	<b>23</b>
4.1	Scaling to Long Horizon Tasks . . . . .	23
4.2	Incorporating Memory . . . . .	25
4.3	Improving Sensors and Sensory Representations . . . . .	25
4.4	Robustifying Robot Hardware . . . . .	25
<b>5</b>	<b>Reproducibility and Call for Collaboration</b>	<b>25</b>

## 1 Introduction

Since our transition away from a nomadic lifestyle, homes have been a cornerstone of human existence. Technological advancements have made domestic life more comfortable, through innovations ranging from simple utilities like water heaters to advanced smart-home systems. However, a holistic, automated home assistant remains elusive, even with significant representations in popular culture [1].

Our goal is to build robots that perform a wide-range of simple domestic tasks across diverse real-world households. Such an effort requires a shift from the prevailing paradigm – current research in robotics is predominantly either conducted in industrial environments or in academic labs, both containing curated objects, scenes, and even lighting conditions. In fact, even for the simple task of object picking [2] or point navigation [3] performance of robotic algorithms in homes is far below the performance of their lab counterparts. If we seek to build robotic systems that can solve harder, general-purpose tasks, we will need to reevaluate many of the foundational assumptions in lab robotics.

In this work we present Dobb-E, a framework for teaching robots in homes by embodying three core principles: efficiency, safety, and user comfort. For efficiency, we embrace large-scale data coupled with modern machine learning tools. For safety, when presented with a new task, instead of trial-and-error learning, our robot learns from a handful of human demonstrations. For user comfort, we have developed an ergonomic demonstration collection tool, enabling us to gather task-specific demonstrations in unfamiliar homes without direct robot operation.

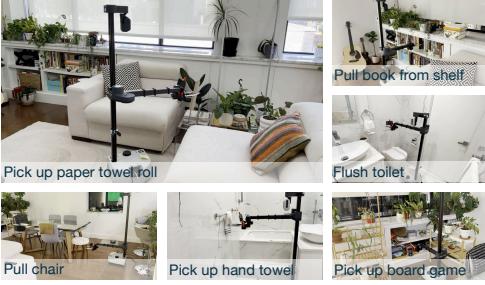
Concretely, the key components of Dobb-E include:

- **Hardware:** The primary interface is our demonstration collection tool, termed the “Stick.” It combines an affordable reacher-grabber with 3D printed components and an iPhone. Additionally, an iPhone mount on the robot facilitates direct data transfer from the Stick without needing domain adaptation.
- **Pretraining Dataset:** Leveraging the Stick, we amass a 13 hour dataset called Homes of New York (HoNY), comprising 5620 demonstrations from 216 environments in 22 New York homes, bolstering our system’s adaptability. This dataset serves to pretrain representation models for Dobb-E.
- **Models and algorithms:** Given the pretraining dataset we train a streamlined vision model, called Home Pretrained Representations (HPR), employing cutting-edge self-supervised learning (SSL) techniques. For novel tasks, a mere 24 demonstrations sufficed to finetune this vision model, incorporating both visual and depth information to account for 3D reasoning.
- **Integration:** Our holistic system, encapsulating hardware, models, and algorithms, is centered around a commercially available mobile robot: Hello Robot Stretch [4].

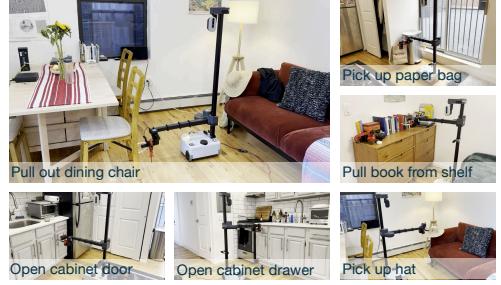


Figure 2: (A) We design a new imitation learning framework, starting with a data collection tool. (B) Using this data collection tool, users can easily collect demonstrations for household tasks. (C) Using a similar setup on a robot, (D) we can transfer those demos using behavior cloning techniques to real homes.

A. SoHo District, New York



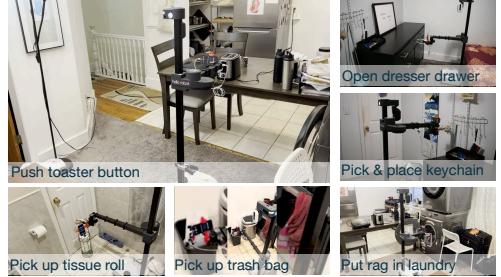
B. Upper East Side, New York



C. Midtown East, New York



D. Long Island City, Queens



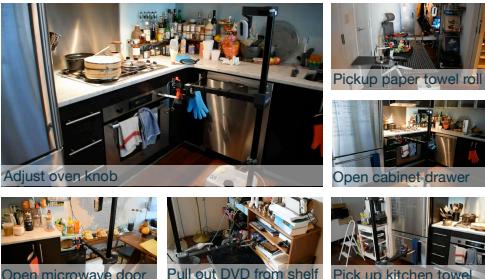
E. East Village, New York



F. Union Square, New York



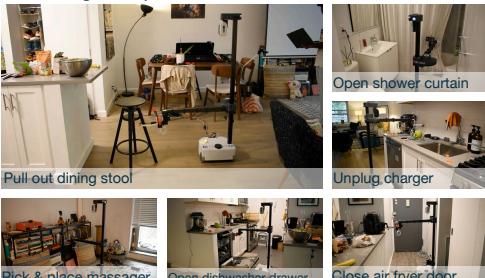
G. Dumbo, Brooklyn



H. Chelsea, New York



I. Washington Square Park, New York



J. Jersey City, New Jersey

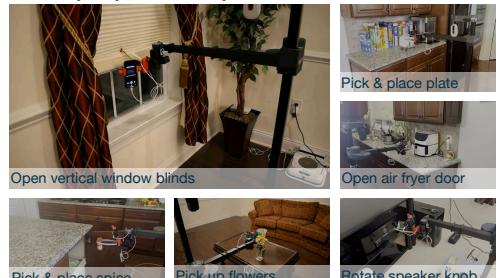


Figure 3: We ran experiments in a total of 10 homes near the New York City area, and successfully completed 102 out of 109 tasks that we tried. The figure shows a subset of 60 tasks, 6 tasks from 10 homes each, from our home robot experiments using Dobb-E.

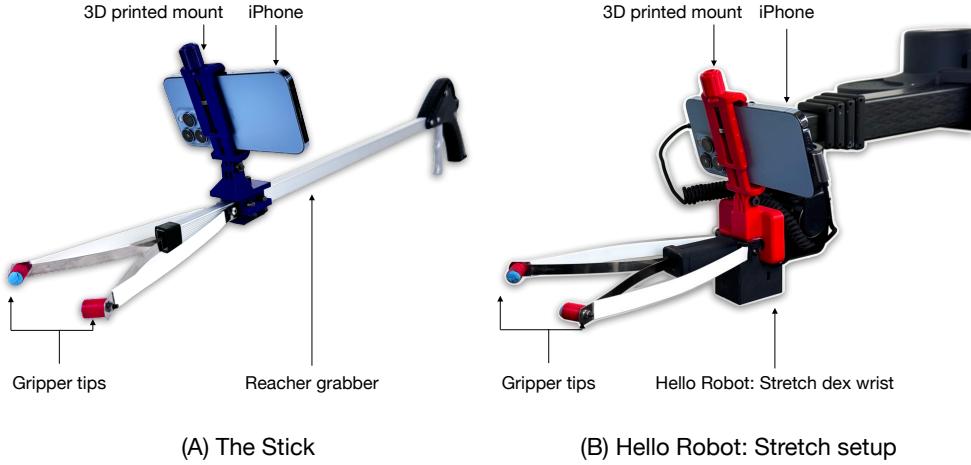


Figure 4: Photographs of our designed hardware, including the (A) Stick and the (B) identical iPhone mount for Hello Robot: Stretch wrist. From the iPhone’s point of view, the grippers look identical between the two setups.

We run Dobb-E across 10 homes spanning 30 days of experimentation, over which it tried 109 tasks and successfully learned 102 tasks with performance  $\geq 50\%$  and an overall success rate of 81%. Concurrently, extensive experiments run in our lab reveals the importance of many key design decisions. Our key experimental findings are:

- **Surprising effectiveness of simple methods:** Dobb-E follows a simple behavior cloning recipe for visual imitation learning using a ResNet model [5] for visual representation extraction and a two-layer neural network [6] for action prediction (see Section 2). On average, only using 91 seconds of data on each task collected over five minutes, Dobb-E can achieve a 81% success rate in homes (see Section 3).
- **Impact of effective SSL pretraining:** Our foundational vision model, HPR trained on home data improves tasks success rate by at least 23% compared to other foundational vision models [7–9], which were trained on much larger internet datasets (see Section 3.4.1).
- **Odometry, depth, and expertise:** The success of Dobb-E is heavily reliant on the Stick providing highly accurate odometry and actions from the iPhones’ pose and position sensing, and depth information from the iPhone’s Lidar. Ease of collecting demonstrations also makes iterating on research problems with the Stick much faster and easier (see Section 3.4).
- **Remaining challenges:** Hardware constraints such as the robot’s force, reach, and battery life, limit tasks our robot can physically solve (see Section 3.3.3), while our policy framework suffers with ambiguous sensing and more complex, temporally-extended tasks (see Sections 3.3.4, 4.1).

To encourage and support future work in home robotics, we have open-sourced our code, data, models, hardware designs, and are committed to supporting reproduction of our results. More information along with robot videos are available on our project website: <https://dobb-e.com>.

## 2 Technical Components and Method

To create Dobb-E we partly build new robotic systems from first principles and partly integrate state-of-the-art techniques. In this section we will describe the key technical components in Dobb-E. To aid in reproduction of Dobb-E, we have open sourced all of the necessary ingredients in our work; please see Section 5 for more detail.

At a high level, Dobb-E is an behavior cloning framework [10]. Behavior cloning is a subclass of imitation learning, which is a machine learning approach where a model learns to perform a task by observing and imitating the actions and behaviors of humans or other expert agents. Behavior cloning involves training a model to mimic a demonstrated behavior or action, often through the use of labeled training data mapping observations to desired actions. In our approach, we pretrain

a lightweight foundational vision model on a dataset of household demonstrations, and then in a new home, given a new task, we collect a handful of demonstrations and fine-tune our model to solve that task. However, there are many aspects of behavior cloning that we created from scratch or re-engineered from existing solutions to conform to our requirements of efficiency, safety, and user comfort.

Our method can be divided into four broad stages: (a) designing a hardware setup that helps us in the collection of demonstrations and their seamless transfer to the robot embodiment, (b) collecting data using our hardware setup in diverse households, (c) pretraining foundational models on this data, and (d) deploying our trained models into homes.

## 2.1 Hardware Design

The first step in scaling robotic imitation to arbitrary households requires us to take a closer look at the standard imitation learning process and its inefficiencies. Two of the primary inefficiencies in current real-world imitation learning lay in the process of collecting the robotic demonstrations and transferring them across environments.

**Collecting robot demonstrations** The standard approach to collect robot demonstrations in a robotic setup is to instrument the robot to pair it with some sort of remote controller device [11, 12], a full robotic exoskeleton [13–16], or simpler data collection tools [17–19]. Many recent works have used a video game controller or a phone [11], RGB-D cameras [20], or virtual reality device [12, 21, 22] to control the robot. Other works [23] have used two paired robots in a scene where one of the robots is physically moved by the demonstrator while the other robot is recorded by the cameras. However, such approaches are hard to scale up to households efficiently. Physically moving a robot is generally unwieldy, and for a home robotic task would require having multiple robots present at the site. Similarly, full exoskeleton based setups as shown in [13, 15, 16] are also unwieldy in a household setting. Generally, the hardware controller approach suffers from inefficiency because the human demonstrators have to map the controller input to the robot motion. Using phones or virtual reality devices are more efficient, since they can map the demonstrators’ movements directly to the robot. However, augmenting these controllers with force feedback is nearly impossible, often leading users to inadvertently apply extra force or torque on the robot. Such demonstrations frequently end up being unsafe, and the generally accepted solution to this problem is to limit the force and torque users can apply; however, this often causes the robot to diverge from the human behavior.

In this project, we take a different approach by trying to combine the versatility of mobile controllers with the intuitiveness of physically moving the robot. Instead of having the users move the entire robot, we created a facsimile of the Hello Robot Stretch end-effector using a cheap \$25 reacher-grabber stick that can be readily bought online, and augmented it ourselves with a 3D printed iPhone mount. We call this tool the “Stick,” which is a natural evolution of tools used in prior work [19, 24] (see Figure 4).

The Stick helps the user intuitively adapt to the limitations of the robot, for example by making it difficult to apply large amounts of force. Moreover, the iPhone Pro (version 12 or newer), with its camera setup and internal gyroscope, allows the Stick to collect RGB image and depth data at 30 frames per second, and its 6D position (translation and rotation). In the rest of the paper, for brevity, we will refer to the iPhone Pro (12 or later) simply as iPhone.

**Captured Data Modalities** Our Stick collects the demonstration data via the mounted iPhone using an off-the-shelf app called Record3D. The Record3D app is able to save the RGB data at  $1280 \times 720$  pixels recorded from the camera, the depth data at  $256 \times 192$  pixels from the lidar sensor, and the 6D relative translation and rotation data from the iPhone’s internal odometry and gyroscope. We record this data at 30 FPS onto the phone and later export and process it.

**Robot Platform** All of our systems are deployed on the Hello Robot Stretch, which is a single-arm mobile manipulator robot already available for purchase on the open market. We use the Stretch RE1 version in all of our experiments, with the dexterous wrist attachment that confers 6D movement abilities on the robot. We chose this robot because it is cheap, lightweight—weighing just 51 pounds (23 kilograms)—and can run on a battery for up to two hours. Additionally, Stretch RE1 has an Intel NUC computer on-board which can run a learned policy at 30 Hz.

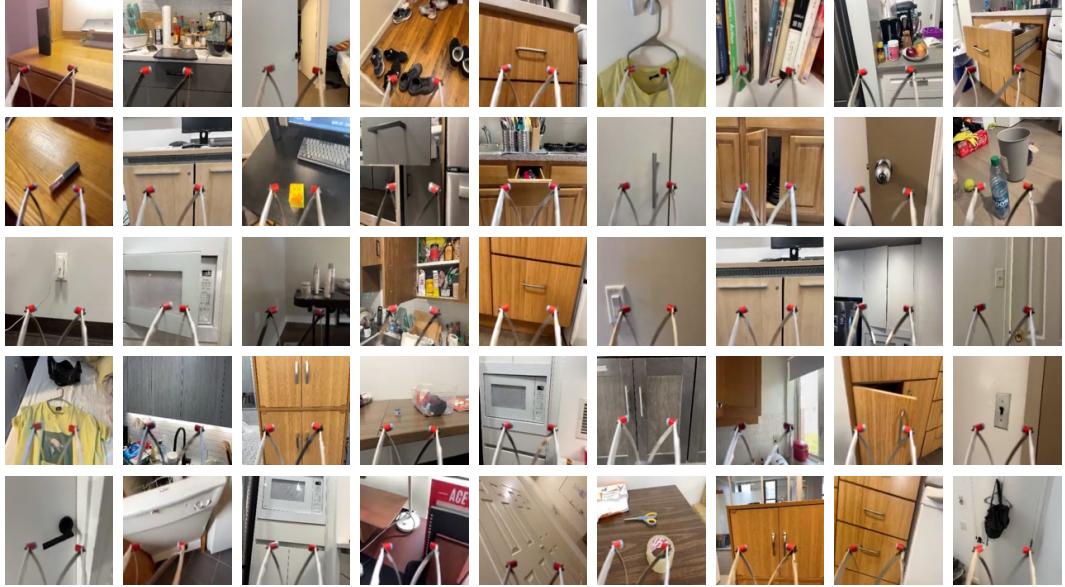


Figure 5: Subsample of 45 frames from Homes of New York dataset, collected using our Stick in 22 homes.

**Camera Mounts** We create and use matching mounts on the Stick and the Hello Robot arm to mount our iPhone, which serves as the camera and the sensor in both cases. One of the main advantages of collecting our data using this setup is that, from the camera’s point of view, the Stick gripper and the robot gripper looks identical, and thus the collected data and any trained representations and policies on such data can be directly transferred from the Stick to the robot. Moreover, since our setup operates with only one robot mounted camera, we don’t have to worry about having and calibrating a third-person, environment mounted camera, which makes our setup robust to general camera calibration issues and mounting-related environmental changes.

**Gripper Tips** As a minor modification to the standard reacher-grabber as well as the Hello Robot Stretch end-effector, we replace the padded, suction-cup style tips of the grippers with small, cylindrical tips. This replacement helps our system manipulate finer objects, such as door and drawer handles, without getting stuck or blocked. In some preliminary experiments, we find that our cylindrical tips are better at such manipulations, albeit making pick-and-place like tasks slightly harder.

## 2.2 Pretraining Dataset – Homes of New York

With our hardware setup, collecting demonstrations for various household tasks becomes as simple as bringing the Stick home, attaching an iPhone to it, and doing whatever the demonstrator wants to do while recording with the Record3D app. To understand the effectiveness of the Stick as a data collection tool and give us a launching pad for our large-scale learning approach, we, with the help of some volunteers, collected a household tasks dataset that we call Homes of New York (HoNY).

The HoNY dataset is collected with the help of volunteers across 22 different homes, and it contains 5620 demonstrations in 13 hours of total recording time and totalling almost 1.5 million frames. We asked the volunteers to focus on eight total defined broad classes of tasks: switching button, door opening, door closing, drawer opening, drawer closing, pick and place, handle grasping, and play data. For the play data, we asked the volunteers to collect data from doing anything arbitrary around their household that they would like to do using the stick. Such playful behavior has in the past proven promising for representation learning purposes [21, 24].

We instructed our volunteers to spend roughly 10 minutes to collect demonstrations in each “environment” or scene in their household. However, we did not impose any limits on how many different tasks they can collect in each home, nor how different each “environment” needs to be across tasks. Our initial demonstration tasks were chosen to be diverse and moderately challenging while still being possible for the robot.

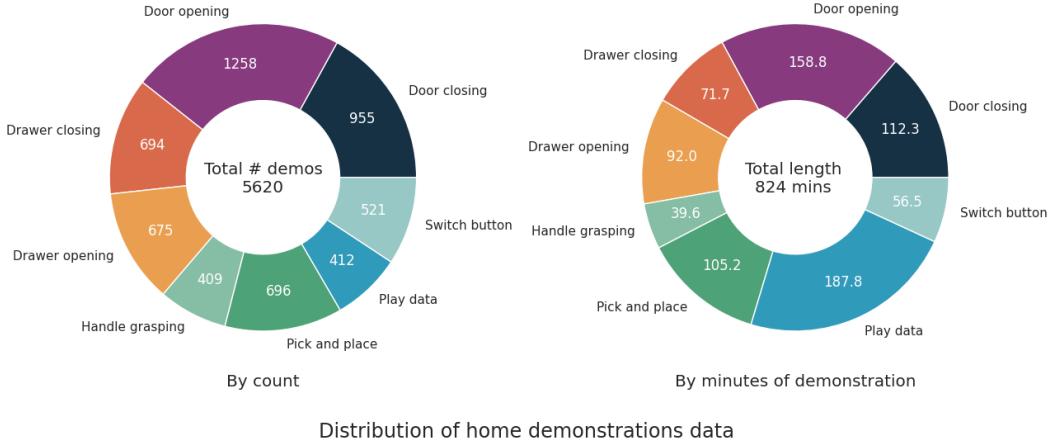


Figure 6: Breakdown of Homes of New York dataset by task: on the left, the statistics is shown by number of demonstrations, and on the right, the breakdown is shown by minutes of demonstration data collected.

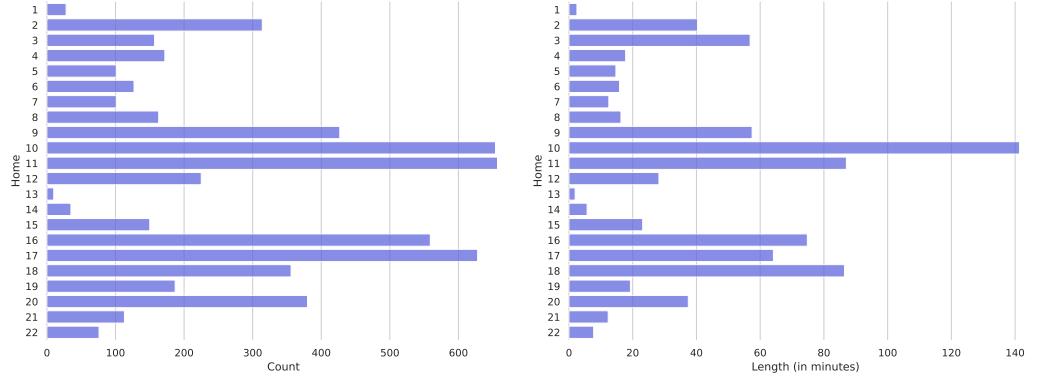


Figure 7: Breakdown of our collected dataset by homes. On the left, the statistics are shown by number of demonstrations, and on the right, the breakdown is shown by minutes of demonstration data collected. The Y-axis is marked with the home ID.

In Figure 6, we can see a breakdown of the dataset by the number of frames belonging to each broad class of tasks. As we can see, while there is some imbalance between the number of frames in each task, they are approximately balanced.

Moreover, our dataset contains a mixture of a diverse number of homes, as shown in Figure 7, with each home containing 67K frames and 255 trajectories on average.

**Gripper Data** While the iPhone can give us the pose of the end-effector, there is no way to trivially get the open or closed status of the gripper itself. To address this, we trained a model to track the gripper tips. We extracted 500 random frames from the dataset and marked the two gripper tip positions in pixel coordinates on those frames. We trained a gripper model on that dataset, which is a 3-layer ConvNet that tries to predict the distance between the gripper tips as a normalized number between 0 and 1. This model, which gets a 0.035 MSE validation error (on a scale from 0-1) on a heldout evaluation set, is then used to label the rest of the frames in the dataset with a gripper value between 0 and 1.

**Dataset Format** As mentioned in the previous section, we collect the RGB and depth data from the demonstration, as well as the 6D motion of the stick, at 30 Hz. For use in our models, we scale and reshape our images and depths into  $256 \times 256$  pixels. For the actions, we store the absolute 6D poses of the iPhone at 30 Hz. During model training or fine-tuning, we calculate the relative pose change as the action at the desired frequency during runtime.

Table 1: While previous datasets focused on the number of manipulation trajectories, we instead focus on diverse scenes and environments. As a result, we end up with a dataset that is much richer in interaction diversity.

Dataset	# Traj.	# Env.	# Homes	Public Data	Public Robot	Collection
MIME [53]	8.30k	1	-	✓	✓	human
RoboTurk [11]	2.10k	1	-	✓	✓	human
Learning in Homes [2]	28k	9	9	✓	✓	scripted
MT-Opt [47]	800k	1	-	✗	✓	scripted & learned
BC-Z [26]	26.0k	1	-	✓	✗	human
RT-1 [25]	130k	3	-	✓	✗	human
RH20T [27]	110k	50	10	✓	✓	human
RoboSet [55]	98.5k	11	-	✓	✓	scripted & human
BridgeData v2 [30]	60.1k	24	-	✓	✓	human & scripted
HoNY (Us)	5.6k	<b>216</b>	<b>22</b>	✓	✓	human

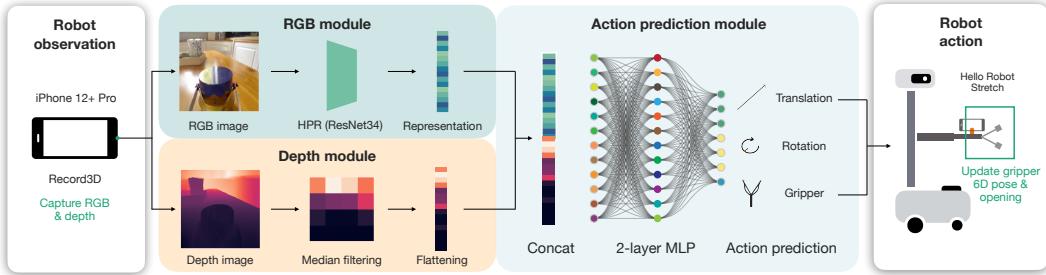


Figure 8: Fine-tuning the pretrained HPR model to learn a model that maps from the robot’s RGB and depth observations into robot actions: 6D relative pose and the gripper opening.

**Dataset Quality Control** We manually reviewed the videos in the dataset to validate them and filter them for any bad demonstrations, noisy actions, and any identifying or personal information. We filtered out any videos that were recorded in the wrong orientation, as well as any videos that had anyone’s face or fingers appearing in them.

**Related Work** Collecting large robotic manipulation datasets is new. Especially in recent years, there have been a few significant advances in collecting large datasets for robotics [2, 25–29, 29–52]. While our dataset is not as large as the largest of them, it is unique in a few different ways. Primarily, our dataset is focused on household interactions, containing 22 households, while most datasets previously were collected in laboratory settings. Secondly, we collect first-person robotic interactions, and are thus inherently more robust to camera calibration issues which affect previous datasets [11, 26, 47, 53–55]. Thirdly, using an iPhone gives us an advantage over previous work that used cheap handheld tools to collect data [17–19] since we can extract high quality action information quite effortlessly using the onboard gyroscope. Moreover, we collect and release high quality depth information from our iPhone, which is generally rare for standard robotic datasets. The primary reason behind collecting our own dataset instead of using any previous dataset is because we believe in-domain pretraining to be a key ingredient for generalizable representations, which we empirically verify in section 3.4.1 by comparing with previously released general-purpose robotic manipulation focused representation models. A line of work that may aid in future versions of this work are collections of first-person non-robot household videos, such as [56–58], where they can complement our dataset by augmenting it with off-domain information.

### 2.3 Policy Learning with Home Pretrained Representations

With the diverse home dataset, our next step in the process is to train a foundational visual imitation model that we can easily modify and deploy in homes. To keep our search space small, in this work we only consider simple visual imitation learning algorithms that only consider a single step at a time. While this inevitably limits the capabilities of our system, we leave temporally extended policies as a future direction we want to explore on home robots. Our policy is built of two simple components: a visual encoder and a policy head.

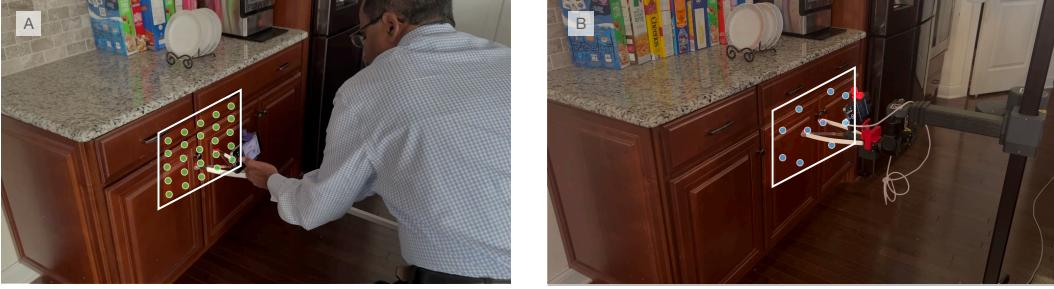


Figure 9: (a) The data collection grid: the demonstrator generally started data collection from a  $5 \times 5$  or  $4 \times 6$  grid of starting positions to ensure diversity of the collected demos. (b) To ensure our policies generalize to different starting positions, we start the robot policy roll-outs from 10 pre-scheduled starting positions.

**Visual Encoder Learning** We use a ResNet34 architecture as a base for our primary visual encoder. While there are other novel architectures that were developed since ResNet34, it satisfies our need for being performant while also being small enough to run on the robot’s onboard computer. We pretrain our visual encoder on our collected dataset with the MoCo-v3 self-supervised learning algorithm for 60 epochs. We call this model the Home Pretrained Representation (HPR) model, based on which all of our deployed policies are trained. We compare the effects of using our own visual encoder vs. a pretrained visual encoder trained on different datasets and algorithms, such as R3M [8], VC1 [9], and MVP [7], or even only pretraining on ImageNet-1K [59], in Section 3.4.1.

**Downstream Policy Learning** On every new task, we learn a simple manipulation policy based on our visual encoder and the captured depth values. For the policy, the input space is an RGB-D image (4 channels) with shape  $256 \times 256$  pixels, and the output space is a 7-dimensional vector, where the first 3 dimensions are relative translations, next 3 dimensions are relative rotations (in axis angle representation), and the final dimension is a gripper value between 0 and 1. Our policy is learned to predict an action at 3.75 Hz, since that is the frequency with which we subsample our trajectories.

The policy architecture simply consists of our visual representation model applied to the RGB channels in parallel to a median-pooling applied on the depth channel, followed by two fully connected layers that project the 512 dimensional image representation and 512 dimensional depth values down to 7 dimensional actions. During this supervised training period where the network learns to map from observation to actions, we do not freeze any of the parameters, and train them for 50 epochs with a learning rate of  $3 \times 10^{-5}$ . We train our network with a mean-squared error (MSE) loss, and normalize the actions per axis to have zero mean and unit standard deviation before calculating the loss.

Our pretrained visual encoders and code for training a new policy on your own data is available open-source with a permissive license. Please see Section 5 for more details.

**Related Work** While the pretraining-finetuning framework has been quite familiar in other areas of Machine Learning such as Natural Language [60, 61] and Computer Vision [5, 62], it has not caught on in robot learning as strongly. Generally, pretraining has taken the form of either learning a visual representation [7–9, 19, 24, 63–68] or learning a Q-function [69, 70] which is then used to figure out the best behavior policy. In this work, we follow the first approach, and pretrain a visual representation that we fine-tune during deployment. While there are recent large-scale robotic policy learning approaches [25, 49, 71], the evaluation setup for such policies generally have some overlap with the (pre-)training data. This work, in contrast, focuses on entirely new households which were never seen during pretraining.

## 2.4 Deployment in Homes

Once we have our Stick to collect data, the dataset preparation script, and the algorithm to fine-tune our pretrained model, the final step is to combine them and deploy them on a real robot in a home environment. In this work, we focus on solving tasks that mostly involve manipulating the environment, and thus we assume that the robot has already navigated to the task space and is starting while facing the task target (which for example could be an appliance to open or an object to manipulate).

**Protocol for Solving Home Tasks** In a novel home, to solve a novel task, we start by simply collecting a handful of demonstrations on the task. We generally collect 24 new demonstrations as a rule of thumb, which our experiments show is sufficient for simple, five second tasks. In practice, collecting these demos takes us about five minutes. However, some environments take longer to reset, in which case collecting demonstrations may also take longer. To confer some spatial generalization abilities to our robot policy, we generally collect the data starting from a variety of positions in front of the task setup, generally in a small  $4 \times 6$  or  $5 \times 5$  grid (Figure 9).

**Policy Training Details** Once the data is collected, it takes about 5 minutes to process the data from R3D files into our dataset format. From there, for 50 epochs of training it takes about 20 minutes on average on a modern GPU (RTX A4000). As a result, on average, within 30 minutes from the start of the data collection, we end up with a policy that we can deploy on the robot.

**Robot Execution Details** We deploy the policy on the robot by running it on the robot’s onboard Intel NUC computer. We use the iPhone mounted on the arm and the Record3D app to stream RGB-D images via USB to the robot computer. We run our policy on the input images and depth to get the predicted action. We use a PyKDL based inverse kinematics solver to execute the predicted relative action on the robot end-effector. Since the model predicts the motion in the camera frame, we added a joint in the robot’s URDF for the attached camera, and so we can directly execute the predicted action without exactly calculating the transform from the camera frame to the robot end-effector frame. For the gripper closing, we binarize the predicted gripper value by applying a threshold that can vary between tasks. We run the policy synchronously on the robot by taking in an observation, commanding the robot to execute the policy-predicted action, and waiting until robot completes the action to take in the next observation. For our evaluation experiments we generally use 10 initial starting positions for each robot task (Figure 9 (b)). These starting positions vary our robot gripper’s starting position in the vertical and horizontal directions. Between each of these 10 trials, we manually reset the robot and the environment.

**Related Work** While the primary focus of our work is deploying robots in homes, we are not the first one to do so. The most popular case would be commercial robots such as Roomba [72] from iRobot or Astro [73] from Amazon. While impressive as a commercial product, such closed-source robots are not conducive to scientific inquiry and are difficult to build upon as a community. Some application of robots in home includes early works such as [74] exploring applications of predefined behaviors in homes, [75, 76] exploring tactile perception in homes, or [2] exploring the divergence between home and lab data. More recently, ObjectNav, i.e. navigating to objects in the real world [3] has been studied by taking robots to six different houses. While [3] mostly experimented on short-term rental apartments and houses, we focused on homes that are currently lived in where cluttered scenes are much more common. There have been other works such as [77, 78] which focus on “in the wild” evaluation. However, evaluation-wise, such works have been limited to labs and educational institutions [77], or have focused on literal “in the wild” setups such as cross-country navigation [78].

### 3 Experiments

We experimentally validated our setup by evaluating it across 10 households in the New York and New Jersey area on a total of 109 tasks. On these 109 tasks, the robot gets an 81% success rate, and can complete 102 tasks with at least even odds. Alongside these household experiments, we also set up a “home” area in our lab, with a benchmark suite with 10 tasks that we use to run our baselines and ablations. Note that none of our experiments overlapped with the environments on which our HoNY dataset was collected to ensure that the experimental environments are novel.

#### 3.1 List of Tasks in Homes

In Table 2 we provide an overview of the 109 tasks that we attempted in the 10 homes, as well as the associated success rate on those tasks. Video of all 109 tasks can also be found on our website: <https://dobb-e.com/#videos>.

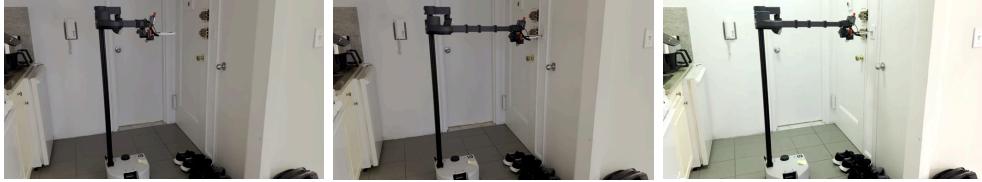
Table 2: A list of all tasks in the home environments, along with their categories and success rates out of 10 trials.

ID	Home	Task Description	Success /10	Task Category
1	1	Door closing: Brown Cabinet	10	Door closing
2	1	Drawer closing: Brown Drawer	10	Drawer closing
3	1	Drawer Opening: Brown Drawer	10	Drawer opening
4	1	Pick up: Plastic Plate	9	Misc object pickup
5	1	Pick up: Flowers	3	Misc object pickup
6	1	Pick and Place: Spices	6	6D pick & place
7	1	Pouring: translucent cup + marshmallows	10	Pouring
8	1	Air Fryer Opening	10	Air-fryer opening
9	1	Air Fryer Closing	10	Air-fryer closing
10	1	Knob Turning	8	Knob turning
11	1	Vertical Blinds Opening	2	Random
12	1	Horizontal Blinds Opening	10	Random
13	2	Sideways washing machine door	8	Door opening
14	2	Dresser drawer	8	Drawer opening
15	2	Placing a rag in laundry	7	6D pick & place
16	2	Picking and placing a keyring	9	6D pick & place
17	2	Pouring: transparent cup	5	Pouring
18	2	Trash pickup	9	Bag pickup
19	2	Toilet paper unloading	8	Random
20	2	Toaster button pressing	1	Random
21	3	Dishwasher drawer opening	8	Drawer opening
22	3	Cat massager pick and place (onto book)	7	6D pick & place
23	3	Rattatouille pick and place	5	6D pick & place
24	3	Air fryer opening	0	Air-fryer opening
25	3	Air fryer closing	10	Air-fryer closing
26	3	Chair pulling	10	Chair pulling
27	3	Light switch new demos	8	Light switch
28	3	Unplugging	10	Unplugging
29	3	Towel pickup	7	Towel pickup
30	3	Kettle switch	0	Random
31	3	Shower curtains	6	Random
32	4	Cabinet door closing	10	Door closing
33	4	Closet door opening	7	Door opening
34	4	Freezer door opening	9	Door opening
35	4	Dishwasher door opening	7	Door opening
36	4	Drawer closing	10	Drawer closing
37	4	Hammerhead shark pick and place	4	6D pick & place
38	4	Oil pouring	5	Pouring
39	4	Almonds pouring	6	Pouring
40	4	Chair pulling	8	Chair pulling
41	4	Book pulling	10	Pulling from shelf
42	4	Tissue pulling	5	Tissue pickup
43	4	Paper bag pickup	8	Bag pickup
44	5	Microwave Door Opening	7	Door opening
45	5	Drawer closing	10	Drawer closing
46	5	Drawer opening	10	Drawer opening
47	5	Chair pulling	10	Chair pulling
48	5	Towel pulling from the fridge	7	Towel pickup
49	5	DVD pulling	10	Pulling from shelf
50	5	Knob turning	5	Knob turning
51	5	Paper towel tube	5	Paper towel replacing
52	6	Door opening kitchen	10	Door opening
53	6	Door opening bathroom	7	Door opening
54	6	Drawer closing	10	Drawer closing
55	6	Mini drawer closing	10	Drawer closing

Continued on the next page

ID	Home	Task Description	Success /10	Task Category
56	6	Dishwasher drawer opening	8	Drawer opening
57	6	Lantern pick and place	9	6D pick & place
58	6	Chair pulling	10	Chair pulling
59	6	Table pulling	10	Chair pulling
60	6	Rag pull	9	Towel pickup
61	6	Book pulling	8	Pulling from shelf
62	6	Tissue pick up	10	Tissue pickup
63	6	Bag pick up	8	Bag pickup
64	6	Cushion lifting	10	Cushion flipping
65	7	Kitchen door closing	10	Door closing
66	7	Bathroom closet door opening	9	Door opening
67	7	Drawer closing black wardrobe	7	Drawer closing
68	7	Drawer closing white wardrobe	10	Drawer closing
69	7	Drawer closing desk	8	Drawer closing
70	7	Drawer closing table	8	Drawer closing
71	7	Chair pulling	9	Chair pulling
72	7	Dining table chair pulling	5	Chair pulling
73	7	Rag pulling	8	Towel pickup
74	7	Tissue paper pick up	10	Tissue pickup
75	7	Paper Towel pick up	10	Paper towel replacing
76	7	Trash pickup	8	Bag pickup
77	8	Door opening	8	Door opening
78	8	Air fryer open	9	Air-fryer opening
79	8	Air fryer close	10	Air-fryer closing
80	8	Chair pulling	10	Chair pulling
81	8	Unplugging	6	Unplugging
82	8	Toilet rag pulling	9	Towel pickup
83	8	Book pulling	8	Pulling from shelf
84	8	Codenames pulling	7	Pulling from shelf
85	8	Tissue pick up	7	Tissue pickup
86	8	Paper towel roll pickup	7	Paper towel replacing
87	8	Food bag pick up	8	Bag pickup
88	8	Cushion flip	10	Cushion flipping
89	8	Toilet flushing	9	Random
90	9	Door closing	10	Door closing
91	9	Door opening	7	Door opening
92	9	Bathroom drawer closing	10	Drawer closing
93	9	Kitchen drawer closing	10	Drawer closing
94	9	Kitchen drawer opening	6	Drawer opening
95	9	Hat pickup	9	Misc object pickup
96	9	Chair pulling	9	Chair pulling
97	9	Light switch	6	Light switch
98	9	Rag pulling	10	Towel pickup
99	9	Book pulling	7	Pulling from shelf
100	9	Paper bag pick up	10	Bag pickup
101	10	Door Closing	10	Door closing
102	10	Drawer Closing	10	Drawer closing
103	10	Air fryer opening	10	Air-fryer opening
104	10	Air fryer closing	10	Air-fryer closing
105	10	Light switch	8	Light switch
106	10	Hand towel (rag) pulling	7	Towel pickup
107	10	Book pulling	10	Pulling from shelf
108	10	Paper towel	9	Paper towel replacing
109	10	Cushion straightening	10	Cushion flipping

A. Turning on light switch



B. Shower curtain opening



C. Trash bag pickup



D. Plush keychain pick and place



E. Lantern pick and place



F. Window blinds opening



G. Oil pouring



H. Microwave door opening



Figure 10: A small subset of 8 robot rollouts from the 109 tasks that we tried in homes. A complete set of rollout videos can also be found at our website: <https://dobb-e.com/#videos>

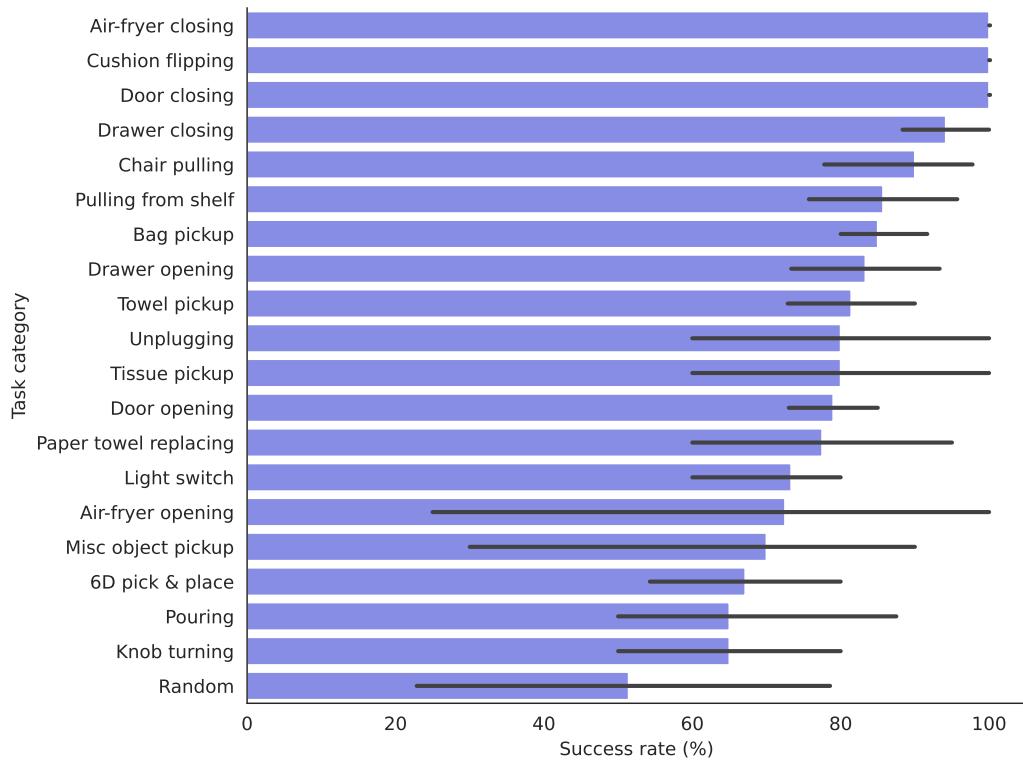


Figure 11: Success rate of our 20 different task groups, with the variance in each group's success rate shown in the error bar.

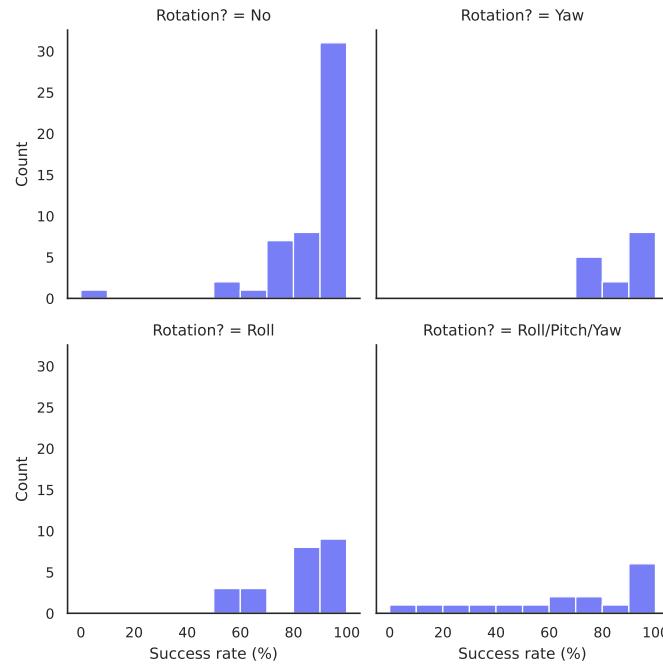


Figure 12: Success rate breakdown by type of actions needed to solve the task. The X-axis shows the number of successes out of 10 rollouts, and the Y-axis shows number of tasks with the corresponding number of success.

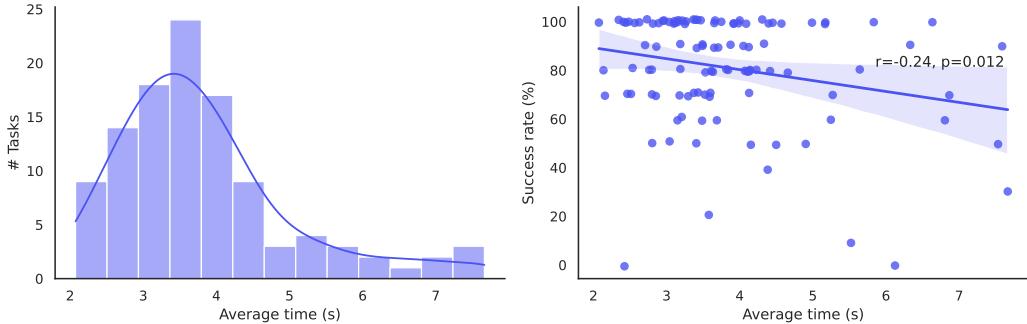


Figure 13: (a) Distribution of time (in seconds) taken to demonstrate a task on our experiment setup. The mean time taken to complete one demonstration is 3.82 seconds, and the median time taken is 3.49 seconds. (b) Correlation analysis between time taken to demonstrate a task and the success rate of the associated robot policy.

### 3.2 Understanding the Performance of Dobb-E

On a broad level, we cluster our tasks into 20 broad categories, 19 task specific and one for the miscellaneous tasks. There are clear patterns in how easy or difficult different tasks may be, compared to each other.

**Breakdown by Task Type** We can see from Figure 11 that Air Fryer Closing and Cushion Flipping are the task groups with the highest average success rate (100%) while the task group with the lowest success rate is 6D pick & place (56%). We found that 6D pick and place tasks generally fail because they generally require robot motion in a variety of axes: like translations and rotations at different axes at different parts of the trajectory, and we believe more data may alleviate the issue. We discuss the failure cases further in Section 3.3.

**Breakdown by Action Type** We can cluster the tasks into buckets by their difficulty as shown in Figure 12. We find that the type of movement affects the success rate of the tasks. Specifically, the distribution of success rates for tasks which do not require any wrist rotation is skewed much more positively compared to tasks where we need either yaw or roll, or a combination of yaw, pitch, and roll. Moreover, the distribution of successes for tasks which require 6D motion is the flattest, which shows that tasks requiring full 6D motions are harder compared to tasks where Dobb-E doesn't require full 6D motion.

**Correlation between demo time and difficulty** Here, we try to analyze the relationship between the difficulty of a task group when done by the robot, and the time required to complete the task by a human. To understand the relationship between these two variables related to a task, we perform a regression analysis between them.

We see from Figure 13 that there is a weak negative correlation ( $r = -0.24$ , with  $p = 0.012 < 0.05$ ) between the amount of time taken to complete a demo by the human demonstrator and how successful the robot is at completing the task. This analysis implies that while longer tasks may be harder for the robot to accomplish, there are other factors that contribute to making a task easy or difficult.

### 3.3 Failure Modes and Analysis

**Lighting and shadows** In many cases, the demos were collected in different lighting conditions than the policy execution. Generally, with enough ambient lighting, our policies succeeded regardless of day and night conditions. However, we found that if there was a strong shadow across the task space during execution that was not there during data collection, the policy may behave erratically.

The primary example of this is from Home 1 Air Fryer Opening (see Figure 14), where the strong shadow of the robot arm caused our policy to fail. Once we turned on an overhead light for even lighting, there were no more failures. However, this shadow issue is not consistent, as we can see in Figure 15, where the robot performs the Home 6 table pulling task successfully despite strong shadows.

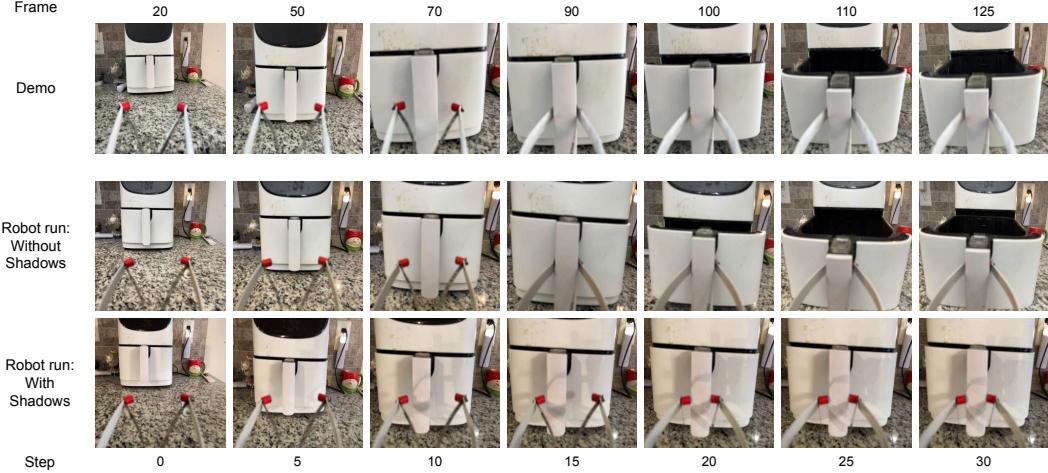


Figure 14: First-person POV rollouts of Home 1 Air Fryer Opening comparing (top row) the original demonstration environment, against robot performance in environments with (middle row) similar lighting, and (bottom row) altered lighting conditions with additional shadows.

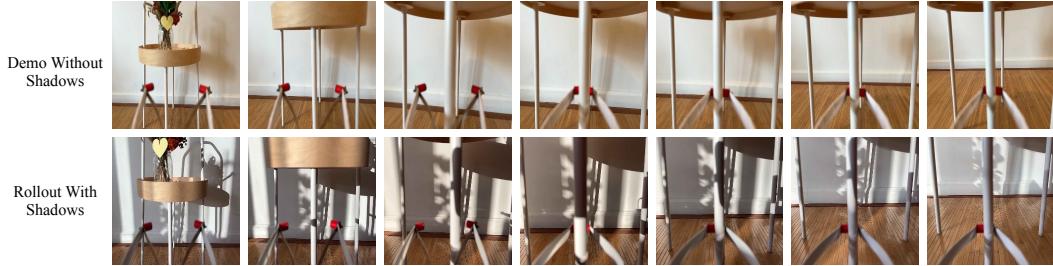


Figure 15: First person view from the iPhone from the (top row) Stick during demonstration collection and (bottom row) the robot camera during rollout. Even with strong shadows during rollout, the policy succeeds in pulling the table.



Figure 16: First person view from the iPhone from the (top row) Stick during demo collection and (bottom row) robot camera during rollout. The demonstrations were collected during early afternoon while rollouts happened at night; but because of the iPhone’s low light photography capabilities, the robot view is similar.

In many cases with lighting variations, the low-light photography capabilities of the iPhone helped us generalize across lighting conditions. For example, in Home 8 cushion straightening (Figure 16), we collected demos during the day and ran the robot during the night. However, from the robot perspective the difference in light levels is negligible.

**Sensor limitations** One of the limitations of our system is that we use a lidar-based depth sensor on the iPhone. Lidar systems are generally brittle at detecting and capturing the depth of shiny and reflective objects. As a result, around reflective surfaces we may get a lot of out-of-distribution values on our depth channel and our policies can struggle.

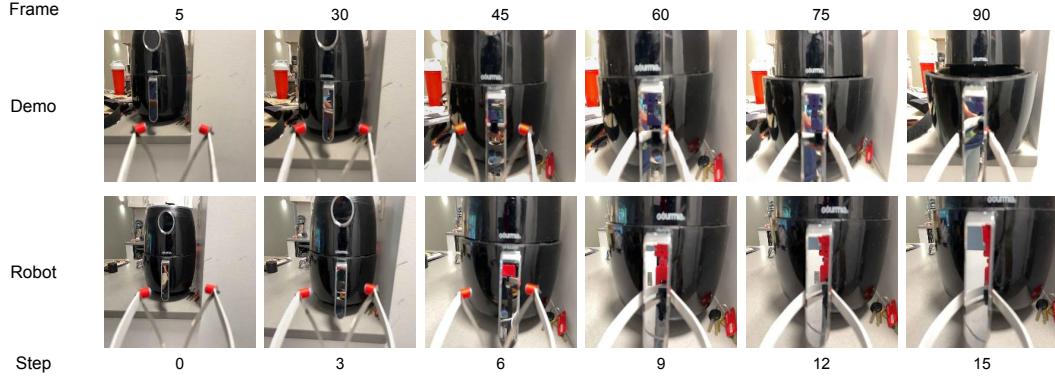


Figure 17: First-person POV rollouts of Home 3 Air Fryer Opening showcasing (top row) a demonstration of the task and (bottom row) robot execution.

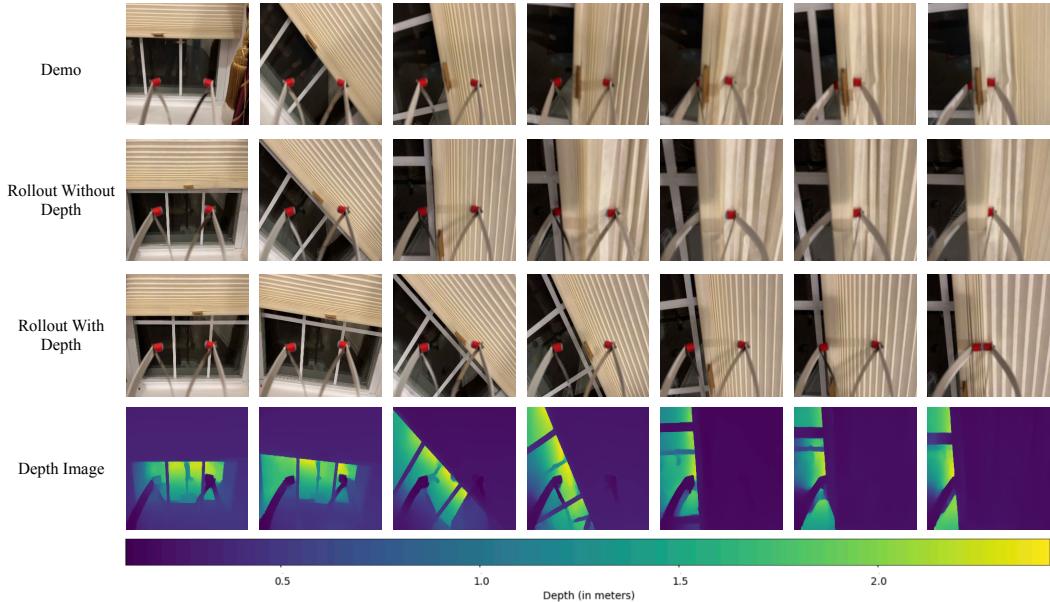


Figure 18: Opening an outward facing window blind (top row) both without depth (second row) and with depth (third row). The depth values (bottom row) for objects outside the window are high and noisy, which cause the depth-aware behavior model to go out of distribution.

A secondary problem with reflective surfaces like mirrors is that we collect demonstrations using the Stick but run the trained policies on the robot. In front of a mirror, the demonstration may actually end up recording the demo collector in the mirror. Then, once the policy is executed on the robot, the reflection on the mirror captures the robot instead of the demonstrator, and so the policy goes out-of-distribution and fails.

One of the primary examples of this is Home 3 Air Fryer Opening (Figure 17). There, the air fryer handle was shiny, and so had both bad depth and captured the demonstration collector reflection which was different from the robot reflection. As a result, we had 0/10 successes on this task.

Another example is Home 1 vertical window blinds opening, where the camera faced outwards in the dark and provided many out-of-distribution values for the depth (Figure 18). In this task, depth-free models performed better (10/10 successes) than depth-using models (2/10 successes) because of such values.



Figure 19: The robot pulling on a heavy door handle (top row) high up from the ground and (bottom row) closer to the ground. Since the robot is bottom heavy, the first case starts tipping the robot while the second case succeeds.

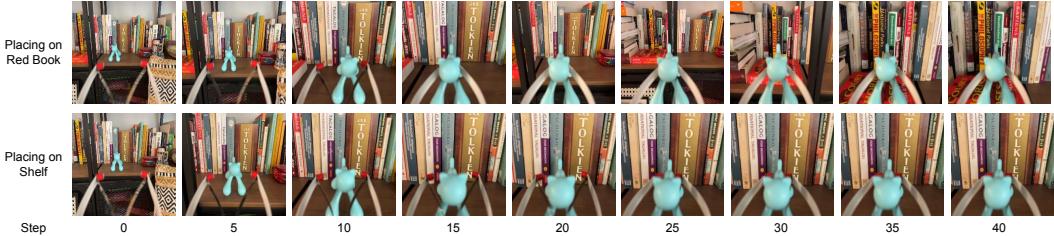


Figure 20: First-person POV rollouts of Home 3 Pick and Place comparing (top) a policy trained on demos where the object is picked and placed onto a red book on a different shelf and (bottom) a policy trained on demos where the object is picked and placed onto that same shelf without a red book. In the second case, since there is no clear signal for when to place the object, the BC policy keeps moving left and fails to complete the task.

**Robot hardware limitations** Our robot platform, Hello Robot Stretch RE1, was robust enough that we were able to run all the home experiments on a single robot with only minor repairs. However, there are certain hardware limitations that caused several of our tasks to fail.

The primary constraint we faced was the robot’s height limit. While the Stretch is tall, the manipulation space caps out at 1m, and thus a lot of tasks like light switch flicking or picking and placing from a high position are hard for the robot to do. Another challenge with the robot is that since the robot is tall and bottom-heavy, putting a lot of pulling or pushing force with the arm near the top of the robot would tilt the robot rather than moving the arm (Figure 19), which was discussed in [79]. Comparatively, the robot was much more successful at opening heavy doors and pulling heavy objects when they were closer to the ground than not, as shown in the same figure. A study of such comparative pulling forces needed can be found in [80, 81].

Knob turning, another low performing task, had 65% success rate because of the fine manipulation required: if the robot’s grasp is not perfectly centered on the knob, the robot may easily move the wrist without moving the knob properly.

**Temporal dependencies** Finally, while our policy only relies on the last observations, for a lot of tasks, being able to consider temporal dependency would give us a much more capable policy class. For example, for a lot of Pick and Place tasks, the camera view right after picking up an object and the view right before placing the object may look the same. In that case, a policy that is not aware of time or previous observations gets confused and can’t decide between moving forward and moving backwards. A clear example of this is in Home 3 Pick and Place onto shelf (Figure 20), where the policy is not able to place the object if the pick location and the place location (two shelf racks) look exactly the same, resulting in 0/10 successes. However, if the policy is trained to pick and place the exact same object on a different surface (here, a red book on the shelf rack), the model succeeds 7/10 times. A policy with temporal knowledge [82–84] could solve this issue.

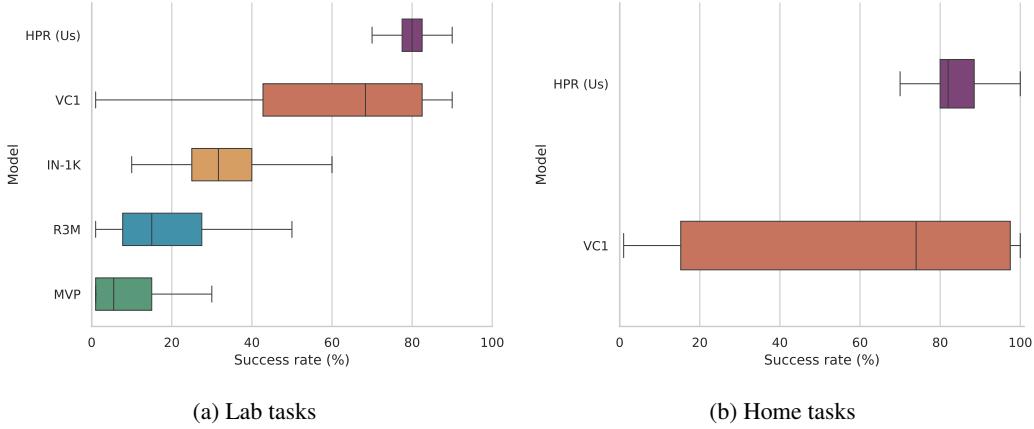


Figure 21: Comparison between different representation models at a set of tasks done in (a) our lab and (b) in a real home environment. As we can see, VC-1 is the representation model closest to ours in performance, however it has a high variance behavior where it either performs well or fails to complete the task entirely. The X-axis shows task completion rate distribution with the error bars showing the 95% confidence interval.

### 3.4 Ablations

We created a benchmark set of tasks in our lab, with a setup that closely resembles a home, to be able to easily run a set of ablation experiments for our framework. To compare various parts of our system, we compare them with alternate choices, and show the relative performance in different tasks. These ablation experiments evaluate different components of our system and how they contribute to our performance. The primary elements of our model that we ran ablations over are the visual representation, number of demonstrations required for our tasks, depth perception, expertise of the demonstrator, and the need for a parametric policy.

**Alternate visual representation models** Our alternate visual representation comparison is with other pretrained representation models such as MVP [7], R3M [8], VC1 [9], and a pretrained ImageNet-1k [5, 59] model. We compare them against our own pretrained models on the benchmark tasks, and compare the performances.

We see that in our benchmark environments, VC1 is the only representation that comes close to our trained representation. As a result, we ran some more experiments with VC1 representation in a household environment. As we can see, while VC1 is closer in performance to our model compared to IN-1K, R3M and MVP, it under-performs our model in household environments. However, VC-1 shows an interesting pattern of bimodal behavior: in each environment it either performs comparatively to HPR, or fails to complete the task entirely.

**Number of demonstrations required for tasks** While we perform all our tasks with 24 demonstrations each, different tasks may require different numbers of demonstrations. In this set of experiments, we show how models trained on different numbers of demonstrations compare to each other.

As we see in Figure 22, adding more demonstrations always improves the performance of our system. Moreover, we see that the performance of the model scales with the number of demonstrations until it saturates. This shows us that on the average case, if our model can somewhat solve a task, we can improve the performance of the system by simply adding more demonstrations.

**Depth Perception** In this work, we use depth information from the iPhone to give our model approximate knowledge of the 3D structure of the world. Comparing the models trained with and without depth in Figure 23, we can see that adding depth perception to the model helps it perform much better than the model with RGB-only input.

The failure modes for tasks without depth are generally concentrated around cases where the robot end-effector (and thus the camera) is very close to some featureless task object, for example a door or a drawer. Because such scenes do not have many features, it is hard for a purely visual imitation

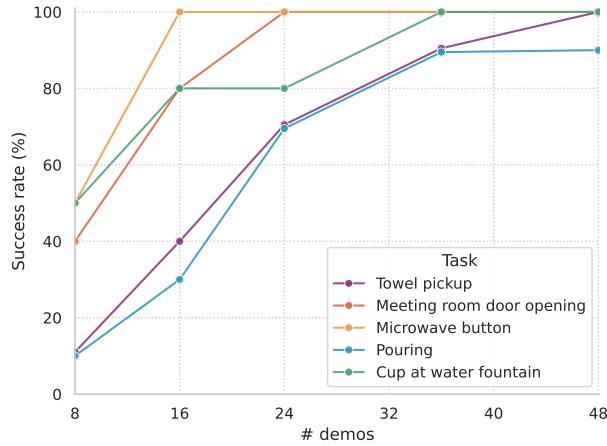


Figure 22: Success rates for a given number of demonstrations for five different tasks. We see how the success rate converges as the number of demonstrations increase.

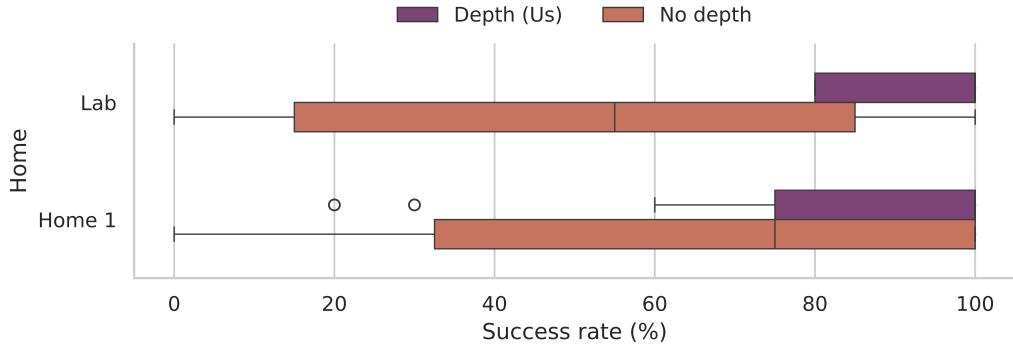


Figure 23: Barplot showing the distribution of task success rates in our two setups, one using depth and another not using depth. In most settings, using depth outperforms not using depth. However, there are some exceptional cases which are discussed in Section 3.3.2.

model without any depth information to know when exactly to close the gripper. On the other hand, the depth model can judge by the distance between the camera and the task surface when to open or close the gripper.

**Demonstrator Expertise** Over the course of our project, we gained experience of how to collect demonstrations with the Stick. A question still remains of how much expertise is needed to operate the Stick and collect workable demonstrations with it.

For this experiment, we have two novice demonstrators collect demonstrations for two tasks in our lab setup. In Task 1, our collected data gave 100% success, while in Task 2, our collected data gave 70% success. Novice collector 1 collected data for Task 1 first and Task 2 second, while collector 2 collected data for Task 2 first and Task 1 second. Collector 1's data had 10% success rate on Task 1, but had 70% success on Task 2. Collector 2's data had 0% success on Task 2 but 90% success on Task 1. From the data, we can see that while it may not be trivial initially to collect demonstrations and teach the robot new skills, with some practice both of our demonstrators were able to collect demonstrations that were sufficient.

**Odometry** In our system, we used the Stick odometry information based on the iPhone's odometry estimate. Previous demonstration collection systems in works like [19, 24] used structure-from-motion based visual odometry methods instead, like COLMAP [85] and OpenSfM [86]. In this



Figure 24: Open-loop rollouts from our demonstrations where the robot actions were extracted using (a) the odometry from iPhone and (b) OpenSfM respectively.

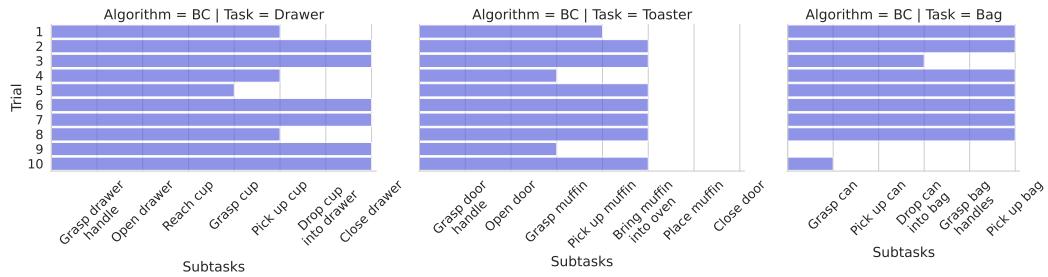


Figure 25: Analysis of our long-horizon tasks by subtasks. We see that Dobb-E can chain subtasks, although the errors can accumulate and make overall task success rate low.

section, we show the difference between the iPhone’s hardware-based and OpenSfM’s visual odometry methods, and compare the quality of the actions extracted from them.

As we can see from the Figure 24, OpenSfM-extracted actions are generally okay while the camera is far away from everything. However, it fails as soon as the camera gets very close to any surface and loses all visual features. The hardware odometry from the iPhone is much more robust, and thus the actions extracted from it are also reliable regardless of the camera view.

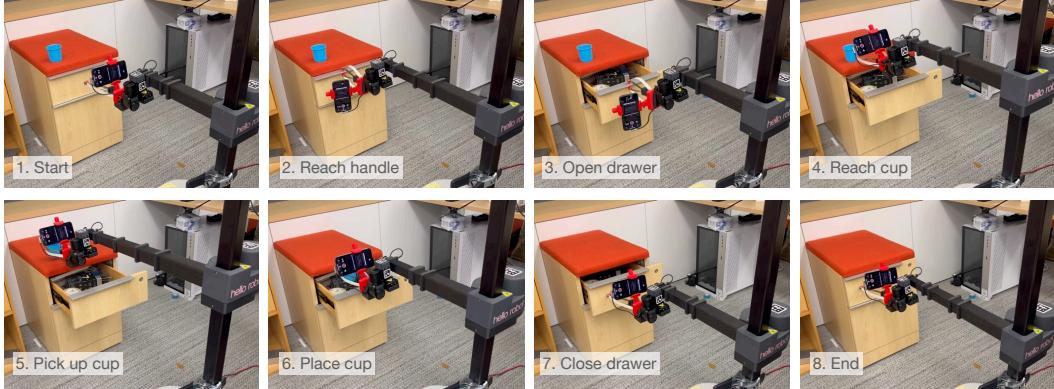
## 4 Open Problems and Request for Research

In this work we have presented an approach to scalable imitation learning that can be applied in household settings. However, there remains open problems that we must address before truly being able to bring robots to homes.

### 4.1 Scaling to Long Horizon Tasks

We primarily focused on short-horizon tasks in this work, but intuitively, our framework should be easily extensible to longer-horizon, multi-step tasks with algorithmic improvements. To validate this intuition, we train Dobb-E to perform some multi-step tasks in our lab.

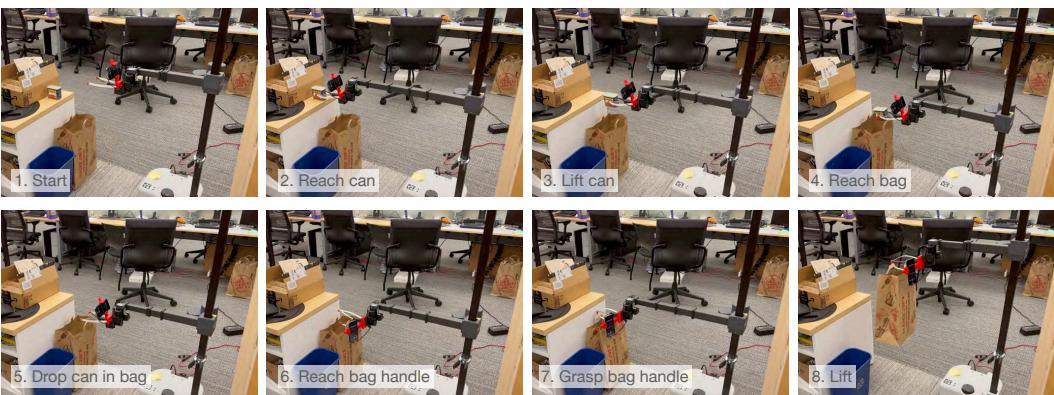
In Figures 26a, 26b, and 26c, we can see that Dobb-E can successfully perform multi-step, long horizon tasks like putting a cup in a drawer, placing a muffin in a toaster oven, or placing a can in a recycling bag and lifting it. However, because of the compound nature of these tasks, the failure cases also tend to compound with our simple methods, as seen in Figure 25. For example, in the muffin-in-toaster task, our model got 1 success out of 10 trials, and in the cup-in-drawer task, our model got 6 success out of 10 trials. In both cases, the sub-task causing primary failure was not letting go of the grasped object (cup or muffin). If we can improve on such particular subtasks, possibly using force-aware methods similar to [87], we believe Dobb-E can easily scale up to long-horizon tasks. Fast on-line adaptation on top of offline training [88, 89] has potential to improve such long horizon cases as well. In other cases, the robot was able to open the door but unable to disengage safely from the handle because some part of the robot gripper got stuck to the handle. This failure mode points to the need of better designed, less bare-boned robot grippers for household tasks.



(a) The robot opening a drawer, placing a cup inside of it, and closing it afterwards.



(b) The robot opening a toaster oven, placing a muffin inside of it, and closing it.



(c) The robot picking up a can, placing it in a bag, and then lifting it.

Figure 26: Dobb-E completing three temporally extended tasks each made up of five to seven subtasks.

## 4.2 Incorporating Memory

Another large challenge in our setup is the problem of robotic scene memory. With a single first person point of view on the Stick, the robot needs to either see or remember large parts of the scene to operate on it effectively. However, there is a dearth of algorithms that can act as standalone memory module for robots. The algorithms that currently exist, such as [90–97] also tend to have a rigid representation of the scene that is hard to change or edit on the fly, which will need to improve for real household deployments.

## 4.3 Improving Sensors and Sensory Representations

Most of current visual representation learning algorithms focus on learning from third-person views, since that is the dominant framework in Computer Vision. However, third person cameras often rely on camera calibration, which generally makes using large robot datasets and transferring data between robots difficult [55]. A closer focus on learning from first person cameras and eye-in-hand cameras would make sharing data from different environments, tasks, and robots much easier. Finally, one of the modality that our Stick is missing is having tactile and force sensors on the gripper. In deployment, we have observed the robot sometimes applies too much or too little force because our framework doesn't contain such sensors. Better integration of cheap sensors [98] with simple data collection tools like the Stick, or even more methods like learned visual contact force estimation [99, 100] could be crucial in such settings.

## 4.4 Robustifying Robot Hardware

A large limitation on any home robotics project is the availability of cheap and versatile robot platforms. While we are able to teach the Hello Robot Stretch a wide-variety of tasks, there were many more tasks that we could not attempt given the physical limitations of the robot: its height, maximum force output, or dexterous capabilities. Some of these tasks may be possible while teleoperating the robot directly rather than using the Stick, since the demonstrator can be creative and work around the limits. However, availability of various home-ready robotic platforms and further development of such demonstration tools would go a long way to accelerate the creation of household robot algorithms and frameworks.

## 5 Reproducibility and Call for Collaboration

To make progress in home robotics it is essential for research projects to contribute back to the pool of shared knowledge. To this end, we have open-sourced practically every piece of this project, including hardware designs, code, dataset, and models. Our primary source of documentation for getting started with Dobb-E can be found at <https://docs.dobb-e.com>.

- **Robot base:** Our project uses Hello Robot Stretch as a platform, which is similarly open sourced and commercially available on the market for US\$24,000 as of November 2023.
- **Hardware design:** We have shared our 3D-printable STL files for the gripper and robot attachment in the GitHub repo: <https://github.com/notmahi/dobb-e/tree/main/hardware>. We have also created some tutorial videos on putting the pieces together and shared them on our website. The reacher-grabber stick can be bought at online retailers, links to which are also shared on our website <https://dobb-e.com/#hardware>.
- **Dataset:** Our collected home dataset is shared on our website. We share two versions, a 814 MB version with the RGB videos and the actions, and an 77 GB version with RGB, depth, and the actions. They can be downloaded from our website, <https://dobb-e.com/#dataset>. At the same time, we share our dataset preprocessing code in GitHub <https://github.com/notmahi/dobb-e/tree/main/stick-data-collection> so that anyone can export their collected R3D files to the same format.
- **Pretrained model:** We have shared our visual pretraining code as well as checkpoints of our pretrained visual model in our GitHub <https://github.com/notmahi/dobb-e/tree/main/imitation-in-homes> and Huggingface Hub <https://huggingface.co/notmahi/dobb-e>. For this work, we also created a high efficiency video dataloader for robotic workload, which is also shared under the same GitHub repository.

- **Robot deployment:** We have shared our pretrained model fine-tuning code in <https://github.com/notmahi/dobb-e/tree/main/imitation-in-homes>, and the robot controller code in <https://github.com/notmahi/dobb-e/tree/main/robot-server>. We also shared a step-by-step guide to deploying this system in a household, as well as best practices that we found during our experiments, in a handbook under <https://docs.dobb-e.com>.

Beyond these shared resources, we are also happy to help other researchers set up this framework in their own labs or homes. We have set up a form on our website to schedule 30-minutes online meetings, and shared some available calendar slots where we would be available to meet online and help set up this system. We hope these steps would be beneficial for practitioners to quickly get started with our framework.

Finally, we believe that our work is an early step towards learned household robots, and thus can be improved in many possible ways. So, we welcome contributions to our repositories and our datasets, and invite researchers to contact us with their contributions. We would be happy to share such contributions with the world with proper credits given to the contributors.

## Acknowledgments

NYU authors are supported by grants from Amazon, Honda, and ONR award numbers N00014-21-1-2404 and N00014-21-1-2758. NMS is supported by the Apple Scholar in AI/ML Fellowship. LP is supported by the Packard Fellowship. Our utmost gratitude goes to our friends and colleagues who helped us by hosting our experiments in their homes, and those who helped us collect the pretraining data. We thank Binit Shah and Blaine Matulevich for support on the Hello Robot Platform and the NYU HPC team, especially Shenglong Wang, for compute support. We thank Jyo Pari and Anya Zorin for their work on earlier iterations of the Stick. We additionally thank Sandeep Menon and Steve Hai for his help in the early stages of data collection. We thank Paula Nina and Alexa Gross for their input on the designs and visuals. We thank Chris Paxton, Ken Goldberg, Aaron Edsinger, and Charlie Kemp for feedback on early versions of this work. Finally, we thank Zichen Jeff Cui, Siddhant Haldar, Ulyana Pieterberg, Ben Evans, and Darcy Tang for the valuable conversations that pushed this work forward.

## References

- [1] Steve Carper. *Robots in American popular culture*. McFarland, 2019.
- [2] Abhinav Gupta, Adithyavairavan Murali, Dhiraj Prakashchand Gandhi, and Lerrel Pinto. Robot learning in homes: Improving generalization and reducing dataset bias. *Advances in Neural Information Processing Systems*, 31:9094–9104, 2018.
- [3] Theophile Gervet, Soumith Chintala, Dhruv Batra, Jitendra Malik, and Devendra Singh Chaplot. Navigating to objects in the real world. *Science Robotics*, 8(79):eadf6991, 2023.
- [4] Charles C Kemp, Aaron Edsinger, Henry M Clever, and Blaine Matulevich. The design of stretch: A compact, lightweight mobile manipulator for indoor human environments. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 3150–3157. IEEE, 2022.
- [5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [6] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.
- [7] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022.
- [8] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *CoRL*, 2022.
- [9] Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Pieter Abbeel, Jitendra Malik, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? *arXiv preprint arXiv:2303.18240*, 2023.
- [10] Christopher G Atkeson and Stefan Schaal. Robot learning from demonstration. In *ICML*, volume 97, pages 12–20. Citeseer, 1997.
- [11] Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In *Conference on Robot Learning*, pages 879–893. PMLR, 2018.
- [12] Sridhar Pandian Arunachalam, Irmak Güzey, Soumith Chintala, and Lerrel Pinto. Holo-dex: Teaching dexterity with immersive mixed reality. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5962–5969. IEEE, 2023.
- [13] Hongjie Fang, Hao-Shu Fang, Yiming Wang, Jieji Ren, Jingjing Chen, Ruo Zhang, Weiming Wang, and Cewu Lu. Low-cost exoskeletons for learning whole-arm manipulation in the wild. *arXiv preprint arXiv:2309.14975*, 2023.
- [14] Fabian Falck, Kawin Larppichet, and Petar Kormushev. De vito: A dual-arm, high degree-of-freedom, lightweight, inexpensive, passive upper-limb exoskeleton for robot teleoperation. In *Towards Autonomous Robotic Systems: 20th Annual Conference, TAROS 2019, London, UK, July 3–5, 2019, Proceedings, Part I 20*, pages 78–89. Springer, 2019.
- [15] Liang Zhao, Tie Yang, Yang Yang, and Peng Yu. A wearable upper limb exoskeleton for intuitive teleoperation of anthropomorphic manipulators. *Machines*, 11(4):441, 2023.
- [16] Yasuhiro Ishiguro, Tasuku Makabe, Yuya Nagamatsu, Yuta Kojio, Kunio Kojima, Fumihiro Sugai, Yohei Kakiuchi, Kei Okada, and Masayuki Inaba. Bilateral humanoid teleoperation system using whole-body exoskeleton cockpit tablis. *IEEE Robotics and Automation Letters*, 5(4):6419–6426, 2020.
- [17] Shuran Song, Andy Zeng, Johnny Lee, and Thomas Funkhouser. Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations. *IEEE Robotics and Automation Letters*, 5(3):4978–4985, 2020.
- [18] Sarah Young, Dhiraj Gandhi, Shubham Tulsiani, Abhinav Gupta, Pieter Abbeel, and Lerrel Pinto. Visual imitation made easy. *arXiv e-prints*, pages arXiv–2008, 2020.
- [19] Jyothish Pari, Nur Muhammad Shafiullah, Sridhar Pandian Arunachalam, and Lerrel Pinto. The surprising effectiveness of representation learning for visual imitation, 2021.

- [20] Sridhar Pandian Arunachalam, Sneha Silwal, Ben Evans, and Lerrel Pinto. Dexterous imitation made easy: A learning-based framework for efficient dexterous manipulation. *arXiv preprint arXiv:2203.13251*, 2022.
- [21] Irmak Guzey, Ben Evans, Soumith Chintala, and Lerrel Pinto. Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play. *arXiv preprint arXiv:2303.12076*, 2023.
- [22] Irmak Guzey, Yinlong Dai, Ben Evans, Soumith Chintala, and Lerrel Pinto. See to touch: Learning tactile dexterity through visual incentives. *arXiv preprint arXiv:2309.12300*, 2023.
- [23] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.
- [24] Sarah Young, Jyothish Pari, Pieter Abbeel, and Lerrel Pinto. Playful interactions for representation learning. *arXiv preprint arXiv:2107.09046*, 2021.
- [25] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.
- [26] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. BC-z: Zero-shot task generalization with robotic imitation learning. In *5th Annual Conference on Robot Learning*, 2021.
- [27] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Junbo Wang, Haoyi Zhu, and Cewu Lu. RH20T: A robotic dataset for learning diverse skills in one-shot. In *RSS 2023 Workshop on Learning for Task and Motion Planning*, 2023.
- [28] Ajay Mandlekar, Jonathan Booher, Max Spero, Albert Tung, Anchit Gupta, Yuke Zhu, Animesh Garg, Silvio Savarese, and Li Fei-Fei. Scaling robot supervision to hundreds of hours with RoboTurk: Robotic manipulation dataset through human reasoning and dexterity. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1048–1055. IEEE, 2019.
- [29] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. In *Robotics: Science and Systems (RSS) XVIII*, 2022.
- [30] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale, 2023.
- [31] Yun Jiang, Stephen Moseson, and Ashutosh Saxena. Efficient grasping from RGBD images: Learning using a new rectangle representation. In *2011 IEEE International conference on robotics and automation*, pages 3304–3311. IEEE, 2011.
- [32] Lerrel Pinto and Abhinav Kumar Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3406–3413, 2015.
- [33] Daniel Kappler, Jeannette Bohg, and Stefan Schaal. Leveraging big data for grasp planning. In *ICRA*, pages 4304–4311, 2015.
- [34] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. Dex-Net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. In *Robotics: Science and Systems (RSS)*, 2017.
- [35] Amaury Depierre, Emmanuel Dellandréa, and Liming Chen. Jacquard: A large scale dataset for robotic grasp detection. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3511–3516. IEEE, 2018.
- [36] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International journal of robotics research*, 37(4-5):421–436, 2018.
- [37] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. QT-Opt:

- Scalable deep reinforcement learning for vision-based robotic manipulation. *arXiv preprint arXiv:1806.10293*, 2018.
- [38] Samarth Brahmbhatt, Cusuh Ham, Charles Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging, 04 2019.
  - [39] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Grasnet-1billion: a large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11444–11453, 2020.
  - [40] Clemens Eppner, Arsalan Mousavian, and Dieter Fox. ACRONYM: A large-scale grasp dataset based on simulation. In *2021 IEEE Int. Conf. on Robotics and Automation, ICRA*, 2020.
  - [41] Konstantinos Bousmalis, Alex Irpan, Paul Wohlhart, Yunfei Bai, Matthew Kelcey, Mrinal Kalakrishnan, Laura Downs, Julian Ibarz, Peter Pastor, Kurt Konolige, Sergey Levine, and Vincent Vanhoucke. Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In *ICRA*, pages 4243–4250, 2018.
  - [42] Xinghao Zhu, Ran Tian, Chenfeng Xu, Mingxiao Huo, Wei Zhan, Masayoshi Tomizuka, and Mingyu Ding. Fanuc manipulation: A dataset for learning-based manipulation with fanuc mate 200iD robot. <https://sites.google.com/berkeley.edu/fanuc-manipulation>, 2023.
  - [43] Kuan-Ting Yu, Maria Bauza, Nima Fazeli, and Alberto Rodriguez. More than a million ways to be pushed. a high-fidelity experimental dataset of planar pushing. In *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 30–37. IEEE, 2016.
  - [44] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2786–2793. IEEE, 2017.
  - [45] Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*, 2018.
  - [46] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. RoboNet: Large-scale multi-robot learning. In *Conference on Robot Learning (CoRL)*, volume 100, pages 885–897. PMLR, 2019.
  - [47] Dmitry Kalashnikov, Jacob Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. MT-Opt: Continuous multi-task robotic reinforcement learning at scale. *arXiv preprint arXiv:2104.08212*, 2021.
  - [48] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulakarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *arXiv preprint arXiv:2108.03298*, 2021.
  - [49] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Chormanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.
  - [50] Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*, 2023.
  - [51] Homanga Bharadhwaj, Jay Vakil, Mohit Sharma, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. RoboAgent: Towards sample efficient robot manipulation with semantic augmentations and action chunking. *arxiv*, 2023.
  - [52] Minho Heo, Youngwoon Lee, Doohyun Lee, and Joseph J. Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. In *Robotics: Science and Systems*, 2023.
  - [53] Pratyusha Sharma, Lekha Mohan, Lerrel Pinto, and Abhinav Gupta. Multiple interactions made easy (MIME): Large scale demonstrations data for imitation. In *Conference on robot learning*, pages 906–915. PMLR, 2018.

- [54] Serkan Cabi, Sergio Gómez Colmenarejo, Alexander Novikov, Ksenia Konyushkova, Scott Reed, Rae Jeong, Konrad Zolna, Yusuf Aytar, David Budden, Mel Vecerik, et al. Scaling data-driven robotics with reward sketching and batch reinforcement learning. *arXiv preprint arXiv:1909.12200*, 2019.
- [55] Homanga Bharadhwaj, Jay Vakil, Mohit Sharma, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. *arXiv preprint arXiv:2309.01918*, 2023.
- [56] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018.
- [57] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.
- [58] Kiran Somasundaram, Jing Dong, Huixuan Tang, Julian Straub, Mingfei Yan, Michael Goesele, Jakob Julian Engel, Renzo De Nardi, and Richard Newcombe. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023.
- [59] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.
- [60] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, pages 4171–4186, 2018.
- [61] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.
- [62] Maxime Oquab, Timothée Darzet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [63] David Brandfonbrener, Ofir Nachum, and Joan Bruna. Inverse dynamics pretraining learns good representations for multitask imitation. *arXiv preprint arXiv:2305.16985*, 2023.
- [64] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*, 2022.
- [65] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.
- [66] Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. *Robotics: Science and Systems (RSS)*, 2023.
- [67] Yao Mu, Shunyu Yao, Mingyu Ding, Ping Luo, and Chuang Gan. EC2: Emergent communication for embodied control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6704–6714, 2023.
- [68] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023.
- [69] Aviral Kumar, Anikait Singh, Frederik Ebert, Mitsuhiko Nakamoto, Yanlai Yang, Chelsea Finn, and Sergey Levine. Pre-training for robots: Offline rl enables learning new tasks from a handful of trials. *arXiv preprint arXiv:2210.05178*, 2022.

- [70] Alexander Herzog, Kanishka Rao, Karol Hausman, Yao Lu, Paul Wohlhart, Mengyuan Yan, Jessica Lin, Montserrat Gonzalez Arenas, Ted Xiao, Daniel Kappler, et al. Deep rl at scale: Sorting waste in office buildings with a fleet of mobile manipulators. *arXiv preprint arXiv:2305.03270*, 2023.
- [71] 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.
- [72] Joseph L Jones. Robots at the tipping point: the road to irobot roomba. *IEEE Robotics & Automation Magazine*, 13(1):76–78, 2006.
- [73] P Dempsey. Reviews-consumer technology. the teardown-amazon astro consumer robot. *Engineering & Technology*, 18(2):70–71, 2023.
- [74] Hai Nguyen and Charles C Kemp. Autonomously learning to visually detect where manipulation will succeed. *Autonomous Robots*, 36:137–152, 2014.
- [75] Tapomayukh Bhattacharjee, Joshua Wade, Yash Chitalia, and Charles C Kemp. Data-driven thermal recognition of contact with people and objects. In *2016 IEEE Haptics Symposium (HAPTICS)*, pages 297–304. IEEE, 2016.
- [76] Tapomayukh Bhattacharjee, Henry M Clever, Joshua Wade, and Charles C Kemp. Multimodal tactile perception of objects in a real home. *IEEE Robotics and Automation Letters*, 3(3):2523–2530, 2018.
- [77] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *Robotics: Science and Systems (RSS)*, 2022.
- [78] Dhruv Shah and Sergey Levine. Viking: Vision-based kilometer-scale navigation with geographic hints. *arXiv preprint arXiv:2202.11271*, 2022.
- [79] Charles C Kemp, Aaron Edsinger, Henry M Clever, and Blaine Matulevich. The design of stretch: A compact, lightweight mobile manipulator for indoor human environments. *arXiv preprint arXiv:2109.10892*, 2021.
- [80] Advait Jain, Hai Nguyen, Mrinal Rath, Jason Okerman, and Charles C Kemp. The complex structure of simple devices: A survey of trajectories and forces that open doors and drawers. In *2010 3rd IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics*, pages 184–190. IEEE, 2010.
- [81] Advait Jain and Charles C Kemp. Improving robot manipulation with data-driven object-centric models of everyday forces. *Autonomous Robots*, 35:143–159, 2013.
- [82] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.
- [83] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023.
- [84] Nur Muhammad Shafiullah, Zichen Cui, Ariuntuya Arty Altanzaya, and Lerrel Pinto. Behavior transformers: Cloning  $k$  modes with one stone. *Advances in neural information processing systems*, 35:22955–22968, 2022.
- [85] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.
- [86] Matthias Adorjan. *OpenSfM: A Collaborative Structure-From-Motion System*. PhD thesis, Wien, 2016.
- [87] Jeremy A Collins, Cody Houff, You Liang Tan, and Charles C Kemp. Forcesight: Text-guided mobile manipulation with visual-force goals. *arXiv preprint arXiv:2309.12312*, 2023.
- [88] Siddhant Haldar, Vaibhav Mathur, Denis Yarats, and Lerrel Pinto. Watch and match: Supercharging imitation with regularized optimal transport. In *Conference on Robot Learning*, pages 32–43. PMLR, 2023.
- [89] Siddhant Haldar, Jyothish Pari, Anant Rai, and Lerrel Pinto. Teach a robot to fish: Versatile imitation from one minute of demonstrations. *arXiv preprint arXiv:2303.01497*, 2023.

- [90] Nur Muhammad Mahi Shafiullah, Chris Paxton, Lerrel Pinto, Soumith Chintala, and Arthur Szlam. Clip-fields: Weakly supervised semantic fields for robotic memory. *arXiv preprint arXiv:2210.05663*, 2022.
- [91] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023.
- [92] Adam Rashid, Satvik Sharma, Chung Min Kim, Justin Kerr, Lawrence Yunliang Chen, Angjoo Kanazawa, and Ken Goldberg. Language embedded radiance fields for zero-shot task-oriented grasping. In *7th Annual Conference on Robot Learning*, 2023.
- [93] Yixuan Wang, Zhuoran Li, Mingtong Zhang, Katherine Driggs-Campbell, Jiajun Wu, Li Fei-Fei, and Yunzhu Li. D<sup>3</sup> fields: Dynamic 3d descriptor fields for zero-shot generalizable robotic manipulation. *arXiv preprint arXiv:2309.16118*, 2023.
- [94] William Shen, Ge Yang, Alan Yu, Jansen Wong, Leslie Pack Kaelbling, and Phillip Isola. Distilled feature fields enable few-shot language-guided manipulation. *arXiv preprint arXiv:2308.07931*, 2023.
- [95] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Obama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, et al. Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*, 2023.
- [96] Benjamin Bolte, Austin Wang, Jimmy Yang, Mustafa Mukadam, Mrinal Kalakrishnan, and Chris Paxton. Usa-net: Unified semantic and affordance representations for robot memory. *arXiv preprint arXiv:2304.12164*, 2023.
- [97] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10608–10615. IEEE, 2023.
- [98] Raunaq Bhirangi, Tess Hellebrekers, Carmel Majidi, and Abhinav Gupta. Reskin: versatile, replaceable, lasting tactile skins. *arXiv preprint arXiv:2111.00071*, 2021.
- [99] Patrick Grady, Jeremy A Collins, Samarth Brahmbhatt, Christopher D Twigg, Chengcheng Tang, James Hays, and Charles C Kemp. Visual pressure estimation and control for soft robotic grippers. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3628–3635. IEEE, 2022.
- [100] Jeremy A Collins, Cody Houff, Patrick Grady, and Charles C Kemp. Visual contact pressure estimation for grippers in the wild. *arXiv preprint arXiv:2303.07344*, 2023.