# Diffusion-PbD: Generalizable Robot Programming by Demonstration with Diffusion Features

Michael Murray, Entong Su, and Maya Cakmak

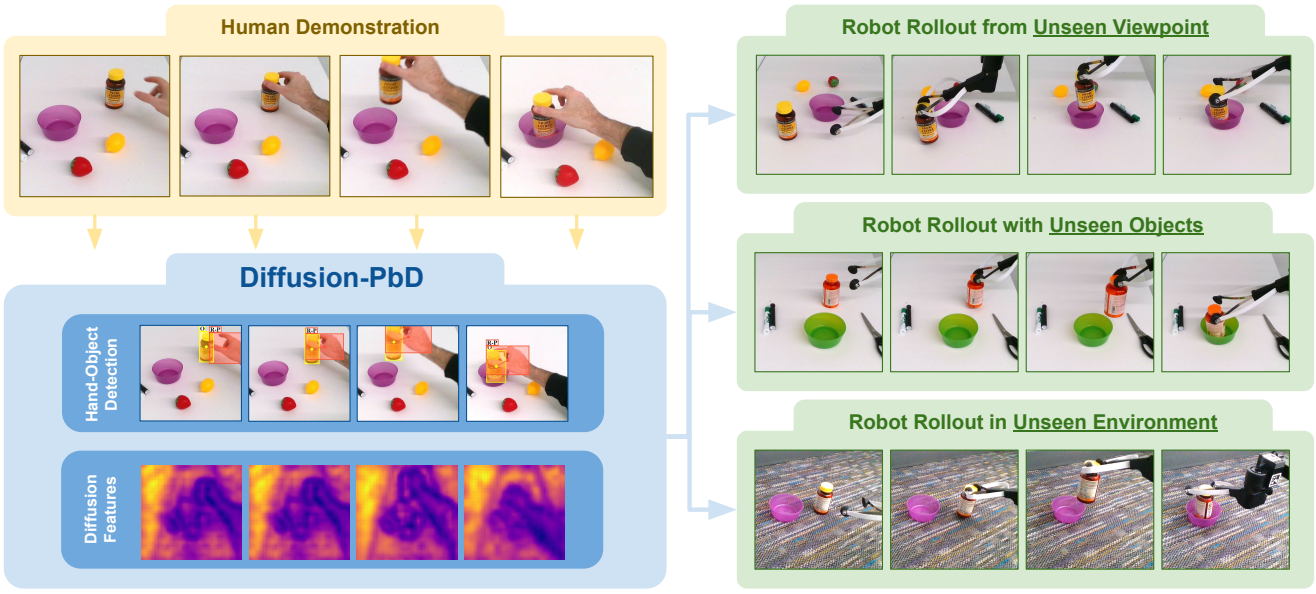


Fig. 1. Given just a single observed human demonstration, Diffusion-PbD can synthesize robot manipulation programs that adapt to unseen objects, unseen viewpoints, and unseen environments. This is done through the use of visual foundation models to both extract salient task structure in the form of waypoints and to transfer that structure to new scenes by identifying reference point correspondences.

***Abstract*— Programming by Demonstration (PbD) is an intuitive technique for programming robot manipulation skills by demonstrating the desired behavior. However, most existing approaches either require extensive demonstrations or fail to generalize beyond their initial demonstration conditions. We introduce Diffusion-PbD, a novel approach to PbD that enables users to synthesize generalizable robot manipulation skills from a single demonstration by utilizing the representations captured by pre-trained visual foundation models. At demonstration time, hand and object detection priors are used to extract waypoints from the human demonstrations anchored to reference points in the scene. At execution time, features from pre-trained diffusion models are leveraged to identify corresponding reference points in new observations. We validate this approach through a series of real-world robot experiments, showing that Diffusion-PbD is applicable to a wide range of manipulation tasks and has strong ability to generalize to unseen objects, camera viewpoints, and scenes. Code and supplementary videos can be found at `https://diffusion-pbd.github.io`**

## I. Introduction

General-purpose robots have the promise to automate tasks in many human-centric environments such as homes and workplaces. However, programming robots to robustly perform behaviors with every possible object in every possible environment is extremely challenging. Programming by Demonstration (PbD) is a popular approach that enables end-users to program new robot capabilities by simply demonstrating the desired behavior [1]. For robots deployed in human-centric environments, demonstration provides an intuitive way for end-users to teach robots new skills without having technical training or expertise in robotics. But this approach typically requires a large-scale and diverse set of demonstrations in order for the programmed capabilities to generalize to new environments and objects, which is not feasible for an end-user to provide. Ideally, an end-user could program robot capabilities by providing just a single demonstration of the desired behavior and those capabilities would generalize to new scenarios. For example, after demonstrating how to put a mug into a coffee machine, the robot should be able to repeat this task with other mugs even if they are visually distinct. Additionally, if the coffee machine and mugs are re-arranged or moved to an entirely different location the robot should still be able to perform the demonstrated task.

Humans possess a remarkable ability to learn tasks from a single demonstration and to apply the learned behaviors to new situations [2]–[5]. This is achieved in part by drawing on prior conceptual knowledge to infer the underlying structure

The authors are with the Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA 98195, USA.
{mmurr, ensu, mcakmak}@cs.washington.edu

of the task being demonstrated rather than directly mimicking the demonstrator's low-level actions [4]. For example, to learn new manipulation skills we primarily pay attention to interactions between the end-effector and objects rather than the relative motions within and between joints. By extracting the high level structure of the task rather than the low level actions, we are able to more easily transfer the task to new scenarios by identifying corresponding structure in new scenes.

Inspired by these insights, we propose a novel approach to PbD that enables programming generalizable robot manipulation skills from a single observed demonstration, illustrated in Figure 1. Our approach draws on the prior conceptual knowledge encoded by pre-trained web-scale foundation models to both extract the salient structure from an observed demonstration and to identify the corresponding structure in new scenes. In particular, we utilize features from pre-trained diffusion models. While diffusion models are primarily models for image synthesis, they have been shown to implicitly encode rich information about the structure of the scene, objects, and object parts within an image. We show that within the context of a PbD framework, such capability provides an elegant mechanism to generalize observed demonstrations to new scenarios. We study the performance of our method across 14 tasks on a real robotic manipulator and find that our approach is surprisingly effective at a wide range of manipulation skills, while utilizing only off-the-shelf models with no additional fine-tuning required. We thoroughly analyze the generalization capabilities of our approach and study the contribution of diffusion features as compared to popular alternatives.

## II. Related Work

### A. Programming by Demonstration

Programming by Demonstration, also referred to as Learning from Demonstration or Imitation Learning, has been the subject of four decades of robotics research [1], [6]. Approaches are often categorized based on the method of providing demonstrations and in contrast to methods that require moving the robot (e.g. through teleoperation [7]–[9], kinesthetic teaching [10]–[12], or spoken commands [13]–[15]), in this work we focus on programming by passive observation, where the robot is programmed by observing a human perform the desired behavior [16]–[20]. This is particularly easy and intuitive for the user, requiring almost no training to perform. However, learning generalizable skills from passive observation is especially challenging and approaches typically either heavily restrict the domain or require a large and diverse set of demonstrations in order to scale to scenarios outside of the demonstrated examples. Some works take a user-guided approach to the generalization problem, where the user provides additional information to adapt the learned skills to new scenes [21]–[24]. Another approach is to extract a reward function from the provided demonstration which can then be used to fine-tune the skill in novel scenes [18], [20], [25], but this requires additional training time to fine-tune the robot policy in new scenes. In this work, we propose to leverage large-scale visual foundation models off-the-shelf, with no additional fine-tuning, to extract robot manipulation skills from a single observed human demonstration and to apply those skills to new objects, viewpoints, and scenes.

### B. Diffusion Models for Robotics

Diffusion models [26] have made great breakthroughs in generative tasks such as image and video synthesis [27]–[31]. Within robotics, diffusion models have been trained to generate actions for manipulation [32]–[34], navigation [35], [36], and human-robot collaboration [37], [38]. Additionally, pre-trained image diffusion models have been utilized to generate images used for robot training [39]–[42] and planning [43], [44]. While diffusion models are primarily used for generative tasks, recent works show evidence that they implicitly encode rich information about the structure of objects and scenes in images [45]–[47]. Based on this insight, we propose to leverage features extracted from pre-trained generative image diffusion models within a PbD framework in order to find correspondences between structures observed in demonstration scenes and those observed in novel scenes.

## III. Diffusion PbD

We present Diffusion-PbD, a robot PbD framework for synthesizing generalizable robot manipulation programs using only a single passively observed human demonstration. Our approach utilizes pre-trained visual foundation models to both extract salient structure from the observed demonstration and to find corresponding structure in novel scenes. Specifically, we use pre-trained models with strong hand-object priors to extract waypoints relative to observation-centric reference points, then we utilize pre-trained diffusion features to find corresponding reference points in novel settings. In the following sub-sections we first formalize the problem setting, then we provide a high-level overview of the approach, and finally we describe each phase of the approach.

### A. Problem Formulation

In this work, we consider PbD for robotic manipulation tasks. Let $\mathcal{A}$ be the set of robot actions, and $\mathcal{S}$ the set of world states. We assume access to a human demonstration $D = \langle d_0, d_1, \ldots, d_{T_D} \rangle$ where each demonstration frame $d_t$ is an RGB-D image at time $t$. Given a demonstration $D$ and an initial state $s \in \mathcal{S}$, the goal is to generate an execution $\xi = \langle a_0, a_1, \ldots, a_{T_\xi} \rangle$, where $a_t \in \mathcal{A}$ is an action taken by the robot at time $t$. The initial state $s$ is defined by the environment layout, the poses and states of all objects, and the pose and state of the robot. The robot does not directly have access to the initial state $s$, but only to an initial observation $o$. The initial observation $o = (I, K, D)$ includes an RGB-D camera image $I$, the robot's proprioceptive state $K$, and the demonstration $D$. The task is considered successful if the goal-conditions corresponding to the demonstration $D$ are true at the end of execution.

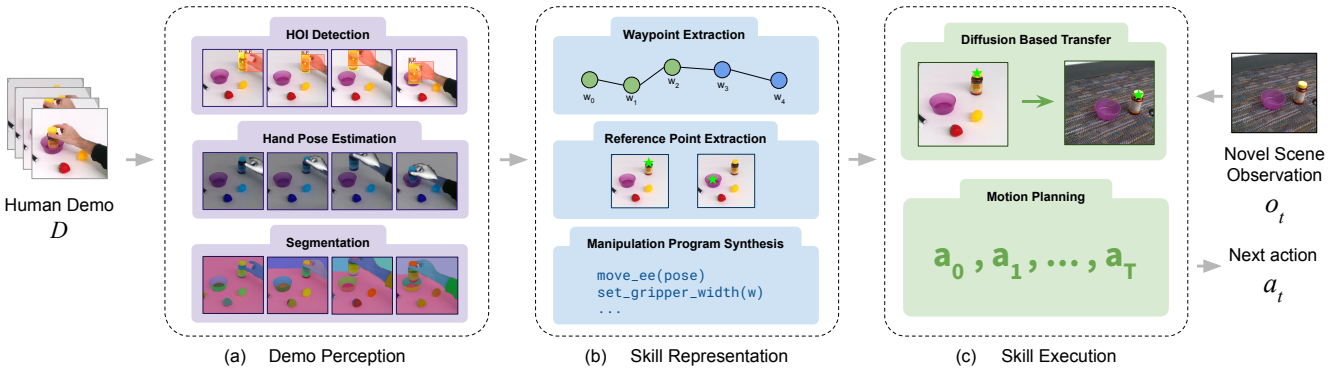


Fig. 3. Diffusion-PbD composes a mixture of pre-trained web-scale foundation models to both extract salient structure from demonstration videos and to transfer that structure to new scenes. Diffusion-PbD is composed of three main phases: (1) human and object detection, (2) waypoint extraction, (3) skill execution. In the first phase, we pre-process the demonstration frames by detecting human hands and their interactions with objects in the scene. Next, we map these detections to waypoints and robot gripper configurations. We anchor the waypoints relative to observation-centric reference points. This representation allow us to map the skill to new scenes by finding corresponding reference points in the new observations.

### B. Overview

Our PbD method composes a mixture of pre-trained web-scale foundation models to both extract salient structure from demonstration videos and to transfer that structure to new scenes. Our method has three main phases (see Figure 3): (1) demo perception, (2) skill representation, (3) skill execution. In the first phase, we pre-process the demonstration frames by detecting human hands and their interactions with objects in the scene. Next, we map those detections to waypoints and robot gripper configurations. We represent waypoints relative to reference points in the observation, which allows us to map the skill to new scenes by finding corresponding reference points in the new observations.

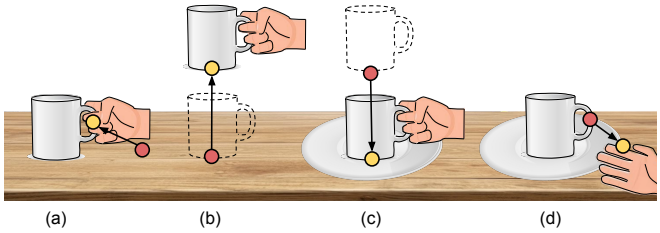


Fig. 4. An illustration of the conditions used to identify key contact frames for an example where a cup is lifted by the handle and moved onto a plate. Key frames are extracted when (a) contact is made between the hand and the target object, (b) contact is broken between the target object and the environment (c) contact is made between the target object and the environment or (d) contact is broken between the hand and the target object.

### C. Demo Perception

For robot manipulation tasks, timesteps when the end-effector interacts with objects in the environment are particularly important. We process the demonstration $D$ to extract information about the hands in the scene and the objects that they contact using 100DOH [48], a hand-object interaction model that has been pre-trained on 100K images extracted from a large-scale (131+ days) video dataset of humans interacting with objects. We use 100DOH to extract, for each demonstration frame $d_t$, a hand bounding box $b_t^h$ and a boolean contact variable $c_t$ indicating whether the hand is in contact with an object or not. For every frame where the hand is in contact with an object we additionally extract an object bounding box $b_t^o$. While bounding boxes give us the rough position of hands and objects in the scene, we look to obtain fine grained masks using Segment Anything Model (SAM) [49]. For each hand bounding box $b_t^h$ and object bounding box $b_t^o$ we prompt SAM to produce a hand mask $m_t^h$ and object mask $m_t^o$. 3D perception of the scene is crucial for manipulation tasks, so we additionally produce a point cloud $C_t$ for each demonstration frame $d_t$ using the RGB-D image and camera intrinsics. To properly imitate the grasp and interaction on a robot, the pose of the hand is important as well. We employ Mediapipe [50] for this purpose, and detect the human hand pose $p_t$ represented as 21 landmarks following the topology in [51]. The two landmarks on the thumb and another two on the index finger are used to represent a parallel jaw gripper. We define $r_t$ as the rotation of this gripper model and $g_t$ as the distance between fingers.

### D. Skill Representation

Inferring the hand-object interactions from the demonstration is useful, but ultimately we want to extract waypoints that can be executed by the robot. To accomplish this, we first identify and extract contiguous *contact sequences*, or clusters of timestamps where $c_t$ is true. Because hand-object interaction detection can be noisy, we filter any sequences that span less than three timestamps, leaving only those that indicate sustained contact. We extract this set of contact sequences $\Sigma = \langle \sigma_0, \sigma_1, \ldots, \sigma_{L_\Sigma} \rangle$ where each contact sequence is initially represented as a start timestamp and end timestamp $\sigma = (t_{start}, t_{end})$. We additionally compute a pre-contact timestamp $t_{pre}$ by backing up from the start of contact $t_{start}$ until the hand mask $m_t^h$ no longer overlaps with the object mask $m_t^o$ to obtain timestamp $t_{pre}$. For each contact sequence in $\Sigma$, we extract a set of waypoints where each waypoint $w_i = (P_i, g_i, t_i)$ is made up of a 6-DOF pose $P_i$, gripper width $g_i$, and timestamp $t_i$. We first define a waypoint at the start of interaction, $w_{start}$ using

the hand pose landmarks at the start of contact $p_{t_{start}}$. The two pose landmarks on the thumb and another two on the index finger are used to represent a parallel jaw gripper. These points are lifted into 3D using the depth map and averaged to obtain our contact point, which is combined with the gripper rotation $r_{t_{start}}$ to obtain our contact pose $P_{start}$ and waypoint $w_{start} = (P_{start}, g_{t_{start}})$. We additionally compute a pre-contact waypoint $w_{pre}$, the points from the thumb and index finger landmarks at $t_{pre}$ are again lifted into 3D, averaged, and combined with $r_{t_{pre}}$ to produce pose $P_{pre}$ and waypoint $w_{pre} = (P_{pre}, g_{t_{pre}})$. As illustrated in Figure 4, we identify additional waypoints centered around timesteps where contact is made or broken between the target object and the environment. Finally, we define a waypoint at the end of interaction $w_{end}$ by repeating this process with the pose landmarks at the end of contact $p_{t_{end}}$. At the end of this process each contact sequence $\sigma$ is represented as a set of waypoints $\sigma = \langle w_0, w_1, \ldots, w_{L_\sigma} \rangle$.

The set of contact sequences $\Sigma$ contains waypoints to reproduce skills in the current scene, but we desire to reproduce skills in novel scenes, including those with novel viewpoints, object configurations, and objects. In this work, we aim to leverage the features from pre-trained image diffusion models for the purpose of re-identifying key waypoints in new scenes. To that end, we extract waypoints relative to *observation-centric* reference points. To obtain reference points, we look for key frames where contact is made between the hand and the target object or between the target object and the environment. To obtain a reference point for a key frame where contact is made between the hand and target object we extract a 3D point from the pose at the start of contact $P_{start}$. We project the 3D point onto the image at timestep $t_{pre}$ to obtain a 2D reference point in image space. To obtain a reference point for a key frame where contact is made between the target object and the environment we average the points in contact to obtain a 3D point and project the resulting point on the the image at timestep $t_{pre}$ to obtain a 2D reference point. After identifying reference points, we recompute the pose in all waypoints using relative translation from the nearest preceding reference point.

### E. Skill Execution

To apply skills to new scenes we first map our reference points to the novel observations using the popular Stable Diffusion (SD) [52] image foundation model. SD has been pre-trained on billions of images and the intermediate-layer features of the model have been shown to implicitly encode rich information about the structure of objects and scenes in an image. In this work, we propose to utilize these features within a PbD framework for robust reference point generalization to unseen viewpoints, objects, and scenes as illustrated in Figure 5. For each contact sequence in $\Sigma$, we use SD to extract the diffusion features of our reference demonstration frame $d_{t_{pre}}$ and the first observation image in the new scene $I$. The features are generated by adding noise to the images, feeding the images through the network of SD, and extracting the intermediate layer activations. For more details we refer the reader to [45]. Through this process we obtain two diffusion feature maps $F_{ref}$ and $F_{target}$. For every waypoint $w_i$ in $\sigma$, we compare the cosine similarity of the two features maps and identify the point in $F_{target}$ that is most similar to the reference point in $F_{ref}$. This point is then lifted into 3D using the depth map from $I$ to produce a 3D point in the new scene $\hat{P}_i$ and new waypoint $\hat{w}_i = (\hat{P}_i, g_i)$. Ultimately we obtain a set of waypoints for the new scene $\hat{\sigma} = (\hat{w}_0, \hat{w}_1, \ldots, \hat{w}_{L_\sigma})$. We convert each contact sequence to a manipulation program for execution on the robot wherein the end-effector motion and gripper state are commanded according to the waypoints in $\hat{\sigma}$. To generate the motion between waypoints, we use a collision-free motion planner to generate a trajectory of robot actions for reaching the next desired waypoint goal. Specifically, we use the GPU accelerated motion generation library cuRobo [53]. After successfully reaching every waypoint goal, this process is repeated for every contact sequence in $\Sigma$.

## IV. Experiments

To evaluate our approach, we conduct a series of real world experiments across 5 indoor environments as illustrated in Figure 6. In our experiments we seek to answer the following research questions: 1) Is Diffusion-PbD practical for a wide range of robot manipulation tasks? 2) How effective is Diffusion-PbD at applying demonstrated manipulation tasks to new viewpoints, objects, and scenes? 3) To what extent do diffusion features contribute to the effectiveness?

### A. Hardware and Environments

We use the Stretch RE2 robot [54] for our experiments. The robot's mobile base, arm lift, and telescoping arm are moved in conjunction to reach 6-DOF target waypoints. The robot's end effector is a parallel-jaw gripper with rubber fingertips. An Intel RealSense D435i RGB-D camera is mounted to the frame which is used both to record demonstrations and to provide observations during execution. One of the authors initialized scenes and categorized tasks as success or failed based on the criteria in Section IV-B.

### B. Evaluation Tasks

We evaluate our approach using 14 different real world manipulation tasks. We design our evaluation tasks to cover a wide range of contact-rich manipulation behaviors involving prehensile and non-prehensile motions. The tasks range from rearranging objects, to multi-step extraction from cluttered scenes, to tool use, to manipulation of deformable and articulated objects. Below, we describe each of the tasks and how success is defined for each task.

*1) Pick-and-place:* In this task, the robot picks up a bottle by its top and places it into a bowl. The task is successful if the robot grasps from the top of the bottle and the bottle is contained inside of the bowl at the end of execution.

*2) Bookshelf extraction:* In this task, the robot is required to do both non-prehensile and prehensile motions to successfully extract a slender object from a bookshelf. The target object is densely packed into the shelf, so the robot must first

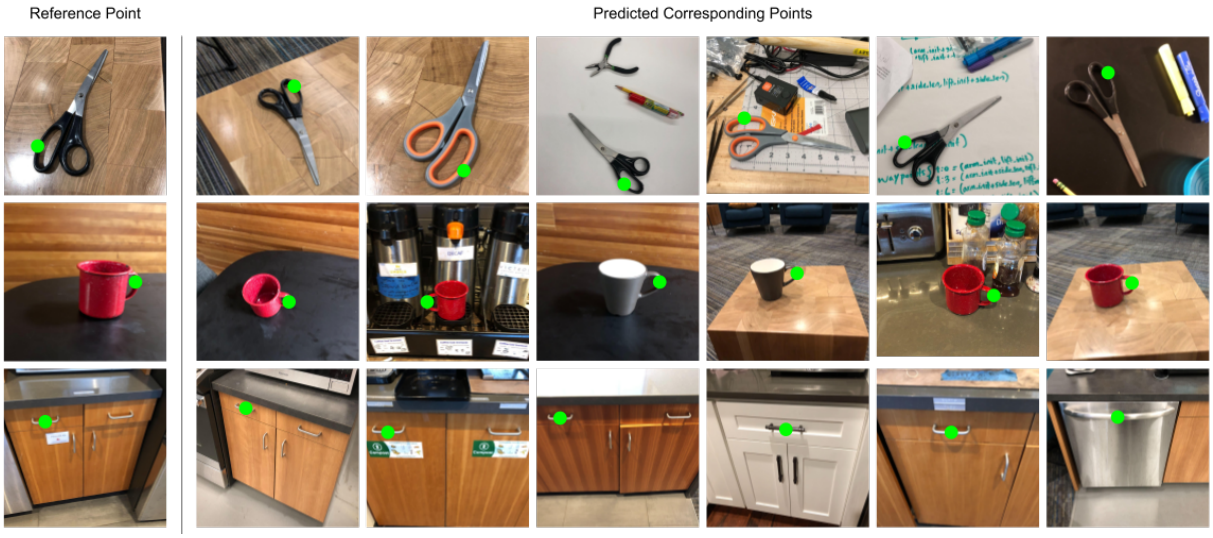


Fig. 5. In Diffusion-PbD, features from a pre-trained Stable Diffusion [52] image model are utilized to transfer demonstrated contact points to new scenes. The examples in this figure show the effectiveness of this method at finding corresponding points in novel viewpoints, objects, and scenes. The reference points on the left are extracted from human demonstrations, and the corresponding points on the right are predicted through the use of diffusion features.

tip the object with a pushing motion from the top before grasping and extracting the object. The task is successful when the robot extracts the target object without displacing any other objects from the shelf.

*3) Occluded pick:* For this task, a target object is occluded by another object in the initial scene. The robot must first push the occluding object out of the way using non-prehensile motion and then extract the target object. The task is successful if the object is extracted by the robot.

*4) Occluded place:* For this task, the robot must use non-prehensile motion to push an object out of the way to make room for the target object on a surface. The target object is then picked and placed onto the surface. The task is successful when the target object rests on the target surface.

*5) Open drawer:* In this task, the robot is required to open a drawer. This requires a precise grasp of the drawer handle and careful imitation of the demonstrated trajectory to open the drawer. The task is successful if the drawer is open at the end of execution.

*6) Close drawer:* In this task, the robot is required to close a drawer. The task is successful if the drawer is closed at the end of execution.

*7) Stack blocks:* This task demonstrates a manipulation program with a multi-step horizon. The robot must stack a set of three colored blocks in the same order as the demonstration. The task is successful when the blocks are stacked in a stable column following the order given by the demonstration.

*8) Unstack blocks:* Another multi-step horizon task, the robot must unstack a set of three colored blocks. The task is successful when none of the blocks are stacked.

*9) Clear table into drawer:* Our longest horizon task where the robot must first open a drawer by the handle, then pick objects one by one off of a counter and place them into the drawer, and finally close the drawer. The task is successful when the drawer is closed with all items from the counter top contained inside.

*10) Unplug charger:* In this task the robot must grasp and pull a laptop charger to remove it from a power outlet socket. The task is successful when the charger is removed from the socket.

*11) Assemble bento:* In this task the robot must pick and place food items into a bento box, putting the items into the same sections of the box as the demonstrator.

*12) Push chair:* In this task the robot must perform a non-prehensile motion to push a chair into a table.

*13) Clean whiteboard:* This task demonstrates a manipulation program with tool use. The robot must first grasp a cloth, then follow the demonstrated trajectory to clean a marking off of a whiteboard using the cloth. The task is successful when the whiteboard is cleaned.

*14) Fold towel:* This task demonstrates a manipulation program with deformable objects. The robot must first grasp the corner of a towel, then follow the demonstrated trajectory to fold the towel. The task is successful when the towel is folded.

## V. Results

To demonstrate the ability of Diffusion-PbD to synthesize a wide variety of robot manipulation skills, we perform experiments on a set of 14 tasks, ranging from pick-and-place, to tool use, to manipulation of deformable and articulated objects. The results are summarized in Table I. For each task, we report results averaged across 15 trials, with a new viewpoint and human demonstration used for each trial. Diffusion-PbD is able to complete all 14 tasks with an average success rate of 81.3%.

**Generalization to Unseen Scenarios:** To understand the ability of Diffusion-PbD to apply demonstrated manipulation tasks to new viewpoints, objects, and scenes we perform a

Fig. 6. We evaluate Diffusion-PbD using a Stretch RE2 robot to perform 14 real world manipulation tasks across 5 visually distinct environments. We show that this approach is effective for single-shot imitation of a wide range of manipulation tasks and generalizes to novel viewpoints, objects, and scenes.

deeper analysis with a representative subset of the manipulation tasks: pick-and-place, open drawer, fold towel, and clear counter into drawer. For each task, we perform additional trials across three novel scenarios. First, we perform 15 trials from unseen viewpoints. Then we perform 15 trials with unseen objects: for the pick-and-place task bottles distinct in appearance and size are used, for the drawer task we use a visually distinct drawer, and for the towel folding task we use towels of varying colors and sizes. Finally, we perform 15 trials from an entirely unseen environment. For each evaluation scenario, we report the average across the 15 trials. The results summarized in Table II show the strong generalization ability of Diffusion-PbD.

**Contribution of Diffusion Features:** To evaluate the major design decision of using features from SD within our framework, we conduct additional experiments using two other pre-trained feature spaces commonly used for correspondence matching in similar robotic applications: CLIP [55], and DINOv2 [56]. The results in Table I and Table II highlight the importance of this design decision as features from SD enable a higher success rate on a range of manipulation tasks and across a range of previously unseen scenarios. Figure 7 qualitatively illustrates this advantage, with examples of the strong correspondence matching enabled by the SD features.

**Failure Analysis:** While the results show that this approach can be practical for all 14 benchmarked manipulation tasks, there is still room for improvement. To better understand the failure cases, we analyze the failures in Figure 8, finding that the most common source of failure happens during demo perception due to inaccuracies in hand-object detection models, highlighting detection improvements as an important area for future work.

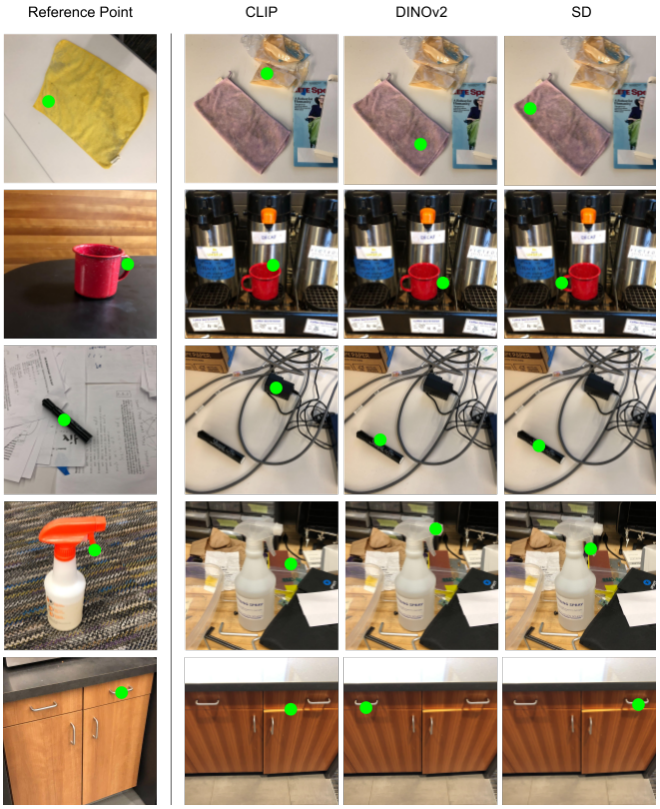


Fig. 7. Qualitative comparison of point correspondences using features from Stable Diffusion [52], DINOv2 [56], and CLIP [55] for scenes with various visual distinctions from the reference and various amounts of clutter.

## VI. Conclusion

We propose Diffusion-PbD, a novel method for robot PbD that can synthesize generalizable robot manipulation programs from observing a single human demonstration. Our method utilizes pre-trained image foundation models off-the-shelf to both extract salient structure from human demonstrations and to transfer that structure to novel scenes. We perform an evaluation on a Stretch RE2 robot and demonstrate the ability of our approach to synthesize robot manipulation programs for a wide-range of different manipulation tasks. Our analysis shows that Diffusion-PbD is effective at generalizing demonstrated skills to unseen viewpoints, objects, and scenes, and highlights the utility of diffusion features for robot PbD.

Despite the promising results, Diffusion-PbD has multiple important limitations. First, our approach relies on sampling a set of waypoints from the provided demonstration. While this representation allows our approach to imitate a wide variety of manipulation tasks, some tasks may need a more

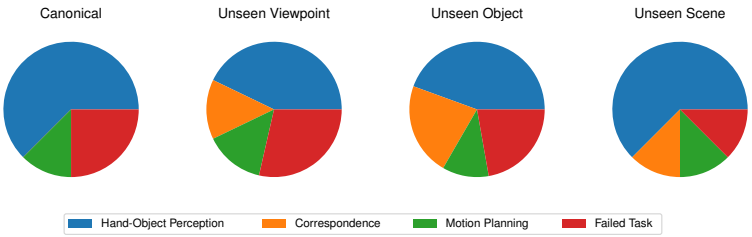


Fig. 8. The distribution of failures for pick-and-place, open drawer, and fold towel tasks across unseen viewpoints, unseen objects, and unseen scenes. Executions can fail due to errors in hand-object perception, errors in correspondence matching, failure to motion plan, or failure to meet the task requirements.

| Task | Success Rate | | |
|---|---|---|---|
| | CLIP | DINOv2 | SD |
| Pick-and-place | 0.86 | 0.86 | **0.93** |
| Bookshelf pick | 0.40 | 0.53 | **0.67** |
| Occluded pick | 0.40 | 0.60 | **0.80** |
| Occluded place | 0.33 | 0.80 | **0.87** |
| Open drawer | 0.53 | 0.73 | **0.80** |
| Close drawer | 0.40 | **0.73** | **0.73** |
| Clear counter into drawer | 0.33 | 0.66 | **0.73** |
| Stack blocks | 0.53 | **0.80** | **0.80** |
| Unstack blocks | 0.40 | 0.80 | **0.87** |
| Unplug charger | 0.66 | **0.93** | **0.93** |
| Assemble bento | 0.66 | 0.66 | **0.80** |
| Push chair | 0.86 | 0.80 | **0.93** |
| Clean whiteboard | 0.66 | **0.73** | **0.73** |
| Fold towel | 0.60 | 0.73 | **0.80** |

TABLE I

WE PRESENT A SET OF EVALUATIONS ON 14 REAL WORLD TASKS. FOR EACH TASK THE ROBOT MUST IMITATE A HUMAN DEMONSTRATION.

| Task | Eval Scenario | Success Rate | | |
|---|---|---|---|---|
| | | CLIP | DINOv2 | SD |
| Pick-and-place | Canonical | 0.86 | 0.86 | **0.93** |
| | Unseen Viewpoint | 0.66 | 0.73 | **0.80** |
| | Unseen Objects | 0.33 | **0.73** | **0.73** |
| | Unseen Scene | 0.66 | **0.86** | **0.86** |
| Open drawer | Canonical | 0.60 | 0.73 | **0.80** |
| | Unseen Viewpoint | 0.40 | 0.66 | **0.80** |
| | Unseen Objects | 0.53 | 0.66 | **0.73** |
| | Unseen Scene | 0.40 | 0.73 | **0.80** |
| Fold towel | Canonical | 0.60 | 0.73 | **0.80** |
| | Unseen Viewpoint | 0.53 | 0.53 | **0.73** |
| | Unseen Objects | 0.33 | 0.66 | **0.80** |
| | Unseen Scene | 0.33 | **0.73** | **0.73** |
| Clear counter | Canonical | 0.33 | 0.66 | **0.73** |
| | Unseen Viewpoint | 0.26 | 0.66 | **0.73** |
| | Unseen Objects | 0.26 | 0.53 | **0.66** |
| | Unseen Scene | 0.13 | 0.53 | **0.66** |

TABLE II

TO STUDY THE ROBUSTNESS OF DIFFUSION-PBD, WE EVALUATE A REPRESENTATIVE SUBSET OF TASKS ON UNSEEN VIEWPOINTS, UNSEEN OBJECTS, AND UNSEEN SCENES.

densely sampled set of waypoints or alternative trajectory representations for finer grained manipulation which is an exciting direction for future work. Our approach uses open-loop execution of actions and could be extended to use dynamic motion generation in order to handle dynamic disturbances or changes to the environment during execution. Additionally, our approach assumes that viable reference points are visible in new scenes, and future work should explore strategies for handling missing or occluded objects.

## ACKNOWLEDGMENT

This research is funded by the UW + Amazon Science Hub. We would like to thank Michael Wolf and Sylvia Dai from Amazon for their support. We would like to thank Yi Li, Markus Grotz, and Abhishek Gupta for helpful discussions.

## REFERENCES

[1] A. Billard, S. Calinon, R. Dillmann, and S. Schaal, "Survey: Robot programming by demonstration," Springrer, Tech. Rep., 2008.

[2] S. A. Al-Abood, K. Davids, and S. J. Bennett, "Specificity of task constraints and effects of visual demonstrations and verbal instructions in directing learners' search during skill acquisition," *Journal of motor behavior*, vol. 33, no. 3, pp. 295–305, 2001.

[3] H. Bekkering, A. WohlschlaÈger, and M. Gattis, "Imitation of gestures in children is goal-directed," *The Quarterly Journal of Experimental Psychology Section A*, vol. 53, no. 1, pp. 153–164, 2000.

[4] N. J. Hodges, A. M. Williams, S. J. Hayes, and G. Breslin, "What is modelled during observational learning?" *Journal of sports sciences*, vol. 25, no. 5, pp. 531–545, 2007.

[5] R. Ramsey, D. M. Kaplan, and E. S. Cross, "Watch and learn: the cognitive neuroscience of learning from others' actions," *Trends in Neurosciences*, vol. 44, no. 6, pp. 478–491, 2021.

[6] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and autonomous systems*, vol. 57, no. 5, pp. 469–483, 2009.

[7] W. Si, N. Wang, and C. Yang, "A review on manipulation skill acquisition through teleoperation-based learning from demonstration," *Cognitive Computation and Systems*, vol. 3, no. 1, pp. 1–16, 2021.

[8] Y. Zhu, A. Joshi, P. Stone, and Y. Zhu, "Viola: Imitation learning for vision-based manipulation with object proposal priors," *6th Annual Conference on Robot Learning (CoRL)*, 2022.

[9] Y. Zhu, Z. Jiang, P. Stone, and Y. Zhu, "Learning generalizable manipulation policies with object-centric 3d representations," in *7th Annual Conference on Robot Learning*, 2023.

[10] T. Lozano-Perez, "Robot programming," *Proceedings of the IEEE*, vol. 71, no. 7, pp. 821–841, 1983.

[11] M. Hersch, F. Guenter, S. Calinon, and A. Billard, "Dynamical system modulation for robot learning via kinesthetic demonstrations," *IEEE Transactions on Robotics*, vol. 24, no. 6, pp. 1463–1467, 2008.

[12] B. Akgun, M. Cakmak, J. W. Yoo, and A. L. Thomaz, "Trajectories and keyframes for kinesthetic teaching: A human-robot interaction perspective," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012, pp. 391–398.

[13] S. Wermter, M. Elshaw, C. Weber, C. Panchev, and H. Erwin, "Towards integrating learning by demonstration and learning by instruction in a multimodal robot," in *Proceedings of the IROS-2003 Workshop on Robot Learning by Demonstration*, 2003, pp. 72–79.

[14] M. Forbes, R. P. Rao, L. Zettlemoyer, and M. Cakmak, "Robot programming by demonstration with situated spatial language understanding," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 2014–2020.

[15] S. Stepputtis, J. Campbell, M. Phielipp, S. Lee, C. Baral, and H. Ben Amor, "Language-conditioned imitation learning for robot manipulation tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 139–13 150, 2020.

[16] Y. Liu, A. Gupta, P. Abbeel, and S. Levine, "Imitation from observation: Learning to imitate behaviors from raw video via context translation," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1118–1125.

[17] J. Huang, D. Fox, and M. Cakmak, "Synthesizing robot manipulation programs from a single observed human demonstration," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 4585–4592.

[18] S. Bahl, A. Gupta, and D. Pathak, "Human-to-robot imitation in the wild," *arXiv preprint arXiv:2207.09450*, 2022.

[19] M. Xu, Z. Xu, C. Chi, M. Veloso, and S. Song, "Xskill: Cross embodiment skill discovery," in *Conference on Robot Learning*. PMLR, 2023, pp. 3536–3555.

[20] S. A. Sontakke, J. Zhang, S. M. Arnold, K. Pertsch, E. Bıyık, D. Sadigh, C. Finn, and L. Itti, "Roboclip: one demonstration is enough to learn robot policies," *arXiv preprint arXiv:2310.07899*, 2023.

[21] S. Calinon and A. Billard, "Active teaching in robot programming by demonstration," in *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2007, pp. 702–707.

[22] M. Forbes, M. Chung, M. Cakmak, and R. Rao, "Robot programming by demonstration with crowdsourced action fixes," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 2, 2014, pp. 67–76.

[23] S. Alexandrova, M. Cakmak, K. Hsiao, and L. Takayama, "Robot programming by demonstration with interactive action visualizations." in *Robotics: science and systems*, 2014, pp. 1–9.

[24] S. Elliott, R. Toris, and M. Cakmak, "Efficient programming of manipulation tasks by demonstration and adaptation," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2017, pp. 1146–1153.

[25] Y. Zhou, Y. Aytar, and K. Bousmalis, "Manipulator-independent representations for visual imitation," *arXiv preprint arXiv:2103.09016*, 2021.

[26] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265.

[27] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[28] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[29] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[30] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet *et al.*, "Imagen video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[31] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7346–7356.

[32] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *arXiv preprint arXiv:2303.04137*, 2023.

[33] U. A. Mishra, S. Xue, Y. Chen, and D. Xu, "Generative skill chaining: Long-horizon skill planning with diffusion models," in *Conference on Robot Learning*. PMLR, 2023, pp. 2905–2925.

[34] Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, C. Xu, J. Luo, T. Kreiman, Y. Tan, D. Sadigh, C. Finn, and S. Levine, "Octo: An open-source generalist robot policy," https://octo-models.github.io, 2023.

[35] A. Sridhar, D. Shah, C. Glossop, and S. Levine, "Nomad: Goal masked diffusion policies for navigation and exploration," *arXiv preprint arXiv:2310.07896*, 2023.

[36] C.-F. Yang, H. Xu, T.-L. Wu, X. Gao, K.-W. Chang, and F. Gao, "Planning as in-painting: A diffusion-based embodied task planning framework for environments under uncertainty," *arXiv preprint arXiv:2312.01097*, 2023.

[37] T. Yoneda, L. Sun, B. Stadie, G. Yang, and M. Walter, "To the noise and back: Diffusion for shared autonomy," *arXiv preprint arXiv:2302.12244*, 2023.

[38] E. Ng, Z. Liu, and M. Kennedy, "Diffusion co-policy for synergistic human-robot collaborative tasks," *IEEE Robotics and Automation Letters*, 2023.

[39] T. Yu, T. Xiao, A. Stone, J. Tompson, A. Brohan, S. Wang, J. Singh, C. Tan, J. Peralta, B. Ichter *et al.*, "Scaling robot learning with semantically imagined experience," *arXiv preprint arXiv:2302.11550*, 2023.

[40] Z. Chen, S. Kiami, A. Gupta, and V. Kumar, "Genaug: Retargeting behaviors to unseen situations via generative augmentation," *arXiv preprint arXiv:2302.06671*, 2023.

[41] A. D. Vuong, M. N. Vu, H. Le, B. Huang, B. Huynh, T. Vo, A. Kugi, and A. Nguyen, "Grasp-anything: Large-scale grasp dataset from foundation models," *arXiv preprint arXiv:2309.09818*, 2023.

[42] H. Bharadhwaj, J. Vakil, M. Sharma, A. Gupta, S. Tulsiani, and V. Kumar, "Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking," *arXiv preprint arXiv:2309.01918*, 2023.

[43] I. Kapelyukh, V. Vosylius, and E. Johns, "Dall-e-bot: Introducing web-scale diffusion models to robotics," *IEEE Robotics and Automation Letters*, 2023.

[44] G. Zhai, X. Cai, D. Huang, Y. Di, F. Manhardt, F. Tombari, N. Navab, and B. Busam, "Sg-bot: Object rearrangement via coarse-to-fine robotic imagination on scene graphs," *arXiv preprint arXiv:2309.12188*, 2023.

[45] L. Tang, M. Jia, Q. Wang, C. P. Phoo, and B. Hariharan, "Emergent correspondence from image diffusion," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[46] G. Luo, L. Dunlap, D. H. Park, A. Holynski, and T. Darrell, "Diffusion hyperfeatures: Searching through time and space for semantic correspondence," in *Advances in Neural Information Processing Systems*, 2023.

[47] G. Zhan, C. Zheng, W. Xie, and A. Zisserman, "What does stable diffusion know about the 3d scene?" in *arXiv:2310.06836*, 2023.

[48] D. Shan, J. Geng, M. Shu, and D. F. Fouhey, "Understanding human hands in contact at internet scale," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9869–9878.

[49] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," *arXiv:2304.02643*, 2023.

[50] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee *et al.*, "Mediapipe: A framework for perceiving and processing reality," in *Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR)*, vol. 2019, 2019.

[51] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1145–1153.

[52] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2021.

[53] B. Sundaralingam, S. K. S. Hari, A. Fishman, C. Garrett, K. V. Wyk, V. Blukis, A. Millane, H. Oleynikova, A. Handa, F. Ramos, N. Ratliff, and D. Fox, "curobo: Parallelized collision-free minimum-jerk robot motion generation," 2023.

[54] C. C. Kemp, A. Edsinger, H. M. Clever, and B. Matulevich, "The design of stretch: A compact, lightweight mobile manipulator for indoor human environments," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 3150–3157.

[55] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[56] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2023.