

# DemoDiffusion: One-Shot Human Imitation using pre-trained Diffusion Policy

Sungjae Park, Homanga Bharadhwaj, Shubham Tulsiani  
 Carnegie Mellon University  
 {sungjae2, hbharadh, stulsian}@andrew.cmu.edu

**Abstract:** We propose *DemoDiffusion*, a simple and scalable method for enabling robots to perform manipulation tasks in natural environments by imitating a single human demonstration. Our approach is based on two key insights. First, the hand motion in a human demonstration provides a useful prior for the robot’s end-effector trajectory, which we can convert into a rough open-loop robot motion trajectory via kinematic retargeting. Second, while this retargeted motion captures the overall structure of the task, it may not align well with plausible robot actions in-context. To address this, we leverage a pre-trained *generalist* diffusion policy to modify the trajectory, ensuring it both follows the human motion and remains within the distribution of plausible robot actions. Our approach avoids the need for online reinforcement learning or paired human-robot data, enabling robust adaptation to new tasks and scenes with minimal manual effort. Experiments in both simulation and real-world settings show that *DemoDiffusion* outperforms both the base policy and the retargeted trajectory, enabling the robot to succeed even on tasks where the pre-trained *generalist* policy fails entirely. Project page: <https://demodiffusion.github.io/>

**Keywords:** Diffusion Policy, One-shot Imitation, Human to Robot

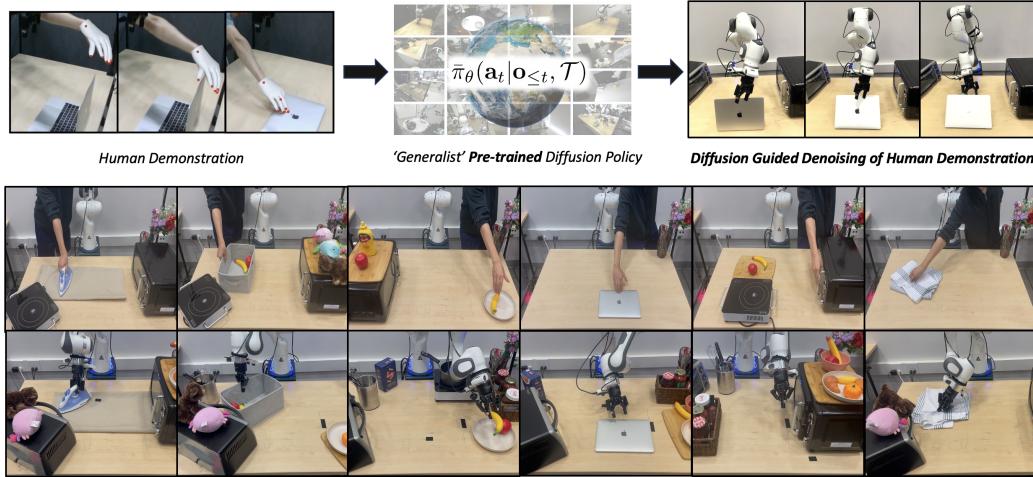


Figure 1: **Overview of DemoDiffusion.** We show how *generalist* pre-trained diffusion policies can be used for following a generic human demonstration showing a manipulation task during deployment. Our real-world manipulation results encompass a wide diversity of manipulation tasks involving everyday objects.

## 1 Introduction

How do we build robot manipulation systems that can be readily deployed in unstructured human environments? One possible answer is to learn ‘generalist’ policies that are capable of accomplishing

any generic task (specified via some description *e.g.*, language or image goal) in any environment. Indeed, there is a general optimism about this paradigm, reflected in several ongoing efforts to collect large-scale demonstration datasets to train such policies. While these efforts have led to impressive results in the form of unified policies capable of performing diverse tasks [1, 2, 3, 4, 5], these policies still struggle to perform meaningfully when deployed zero-shot to novel environments or asked to perform unseen tasks. Deployment thus often requires additional fine-tuning using task-and-scene-specific robot demonstrations [1], but this is a non-trivial overhead as collecting robot demonstrations in the real world can be time-consuming and beyond the expertise of an average user. In this work, we propose an alternate deployment mechanism – leveraging a pre-trained generalist policy to perform a task via imitating a single *human* demonstration.

We are of course not the first to consider this goal of allowing robots to imitate any given human demonstration. One common approach [6, 7, 8, 9] is to instantiate this as kinematic retargeting task and compute open-loop robot actions that maximize a manually defined similarity between the achieved robot end-effector configuration and observed human hand poses (*e.g.*, matching locations of fingertips). However, the human-robot embodiment mismatch makes it difficult for the retargeted actions to achieve precisely the same effects as the human ones, and the open-loop execution further makes the approach brittle to noise and scene variations. Another line of work for human imitation attempts to learn robot policies through online reinforcement learning [10, 11, 12, 13, 14], where the human demonstration helps define reward functions. While this can overcome the embodiment gap, such test-time online RL requires hours of online interaction and resets - making it difficult to adopt such methods for generic real-world manipulation tasks, especially in safety-critical scenarios. We instead seek to develop an approach for demonstration following that, akin to kinematic retargeting, can be deployed ‘one-shot’ *i.e.*, without any test-time training while still benefitting from generic learning-based priors for more precise closed-loop interaction.

Our approach builds on the insight that pre-trained diffusion policies can act as priors for robot action. Inspired by prior work in leveraging pre-trained diffusion models for image editing [15], we present *DemoDiffusion*—a formulation to utilize diffusion policy trained on robot interaction data for synthesizing coherent robot actions from a human demonstration. Specifically, we first perform kinematic retargeting by extracting human hand poses from the demonstration and obtain an open-loop robot action trajectory. While this trajectory is typically suboptimal due to embodiment differences and lack of closed-loop feedback, it serves as an effective initialization that can be improved via a diffusion policy. We do so by injecting gaussian noise and applying the pre-trained diffusion policy to iteratively denoise the trajectory conditioned on robot observations, yielding a refined, executable sequence of robot actions. *DemoDiffusion* thus enables the robot to use the pre-trained policy as a prior and adapt the human-derived trajectory to its own embodiment and environment in a closed-loop manner.

In summary, we propose *DemoDiffusion*: a framework for robotic manipulation that allows a robot to perform generic tasks in natural environments by following a human demonstration with guidance from a pre-trained diffusion policy. Importantly, our framework doesn’t require any robot demonstration of the target task in the target environment and doesn’t require any online interaction or fine-tuning. Our experiments across simulation and real-world environments show that *DemoDiffusion* surpasses the performance of both, the base policy and retargeted trajectory, even allowing the robot to perform tasks where pre-trained *generalist* policy completely fails.

## 2 Related Work

**Generalist Manipulation Policies.** There has been a growing trend in developing ‘*generalist*’ robotic policies that can perform multiple tasks based on a specified goal in the form of an image or a language instruction. Collecting large-scale robot interaction datasets [16, 17] combined with behavior cloning and vision-language pre-training [5, 2, 18, 1, 4, 3] is the predominant recipe for training such multi-task policies. However, since collecting large-scale robot interaction datasets in natural settings like homes and offices is challenging due to operational constraints, current best models still struggle to perform manipulation tasks zero-shot and thus are not yet deployable in-the-

wild. Our work enables using such *generalist* diffusion policies for following a human demonstration, thereby being able to perform tasks that the pre-trained policy might be unable to zero-shot.

**Robotic Manipulation with Non-Robot Datasets.** Instead of requiring large-scale robot interaction datasets, a growing body of works have begun utilizing human videos and large-scale web videos for robotics. A lot of these approaches have been enabled by recent advances in computer vision for representation learning [19, 20, 21], predicting tracks [22, 23], and reconstructing hand-object interactions from monocular videos [24, 25, 26, 27, 27]. Earlier works learned self-supervised visual representations from such non-robotic datasets [28, 29, 30] that can serve as the visual backbone of robotic policies [31, 32, 33, 34, 35, 36]. Recent works in this paradigm predict manipulation-relevant cues from web videos in the form of motion trajectories [37, 38] and object affordances [39, 40] and combine these predictive models with a limited amount of robot interaction data for training policies via conditional behavior cloning. Our work is orthogonal to these approaches in that instead of using human videos for training, we require a single human demonstration during deployment for guiding a pre-trained generalist policy to perform a new task.

**One-Shot Imitation from Robot Demonstration.** One-shot imitation aims to enable robots to perform a new task with guidance from a single demonstration. Prior works have explored this in the context of a robot demonstration provided at test time, via training demonstrateion-conditioned policies [41]. Recent works have developed learning-based visual servoing systems [42] and algorithms for identifying object invariances [43] to replicate the robot end-effector actions from the demonstration onto novel configurations of objects. However, requiring a robot demonstration for imitation is restrictive for ubiquitous deployment as end-users might find it challenging to tele-operate the robot to collect demonstrations in natural human environments.

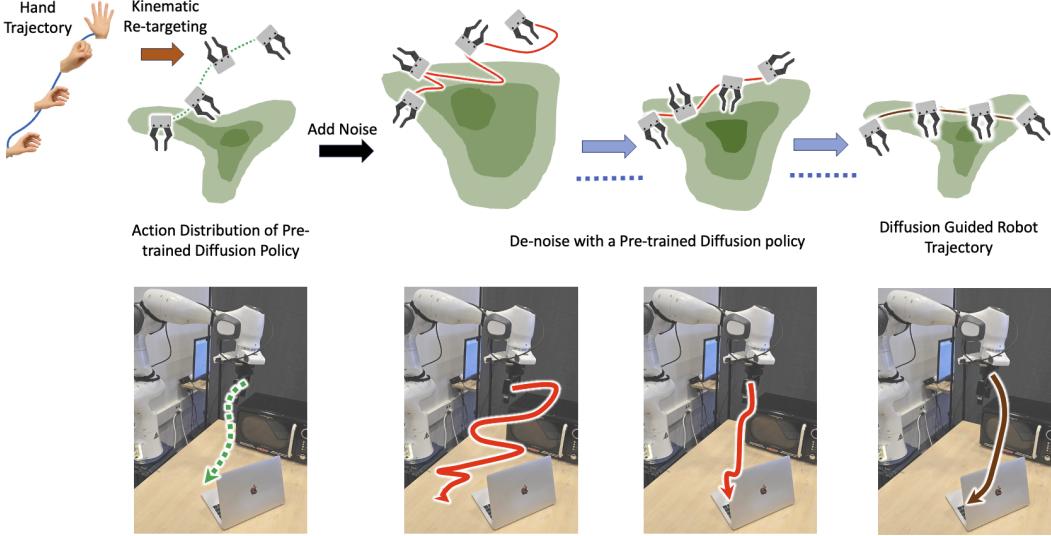
**One-Shot Imitation from Human Demonstration.** To imitate a human demonstration for robot manipulation, a straightforward approach is kinematic retargeting of the human hand pose to the robot end-effector pose per time-step, as demonstrated by some recent works [6, 7]. Although simple to implement, the human-robot embodiment mismatch typically introduces errors in the retargeting and the open-loop execution is brittle to object pose variations during deployment. To remedy this, another line of works seek to learn this retargeting implicitly by training a closed-loop policy from paired human-robot datasets [44, 45] enabling more flexible deployments. Since such paired datasets are difficult to collect, other works [10, 11, 46, 12] perform reinforcement learning with reward functions derived from video comparison. However, such methods require online interaction and resets, limiting their practicality. We propose an alternate paradigm: given the human demonstration, we refine retargeted human behavior using a pre-trained robot policy, avoiding brittle replays and enabling closed-loop control. This formulation doesn't rely on expensive paired human-robot data collection and also doesn't require cumbersome test-time fine-tuning via interaction.

### 3 Method

We target the problem of one-shot visual imitation. Given a human demonstration  $\mathbf{D}$  depicting a manipulation task with description  $\mathcal{T}$ , we want a robot manipulator to perform the same task. Unlike prior work that relies on test-time RL training [10] or paired human-robot datasets [44], we wish to enable such imitation ‘one-shot’ assuming access to a closed-loop diffusion policy  $\pi_\theta(\mathbf{a}_t | \mathbf{o}_{\leq t}, \mathcal{T})$  pre-trained on some broad robot interaction dataset  $\mathcal{D}_{robot}$ . Our assumption on the human demonstration  $\mathbf{D}$  is also very minimal: it can either be an RGBD video or a multi-view video of a human doing a task, such that 3D hand poses per-timestep can be reliably extracted from it.

#### 3.1 Overview

Our approach is based on two key insights. The first is that the trajectory of the hand pose in the human demonstration  $\mathbf{D}$  provides useful information of the approximate trajectory the robot end-effector should follow, and we can perform *kinematic retargeting* of the hand trajectory  $\{\mathbf{h}_t\}_{t=0}^T$  to an open-loop robot end-effector trajectory  $\{\hat{\mathbf{a}}_t\}_{t=0}^T$ . The second insight is that the kinematically retargeted robot trajectory has the correct form of motion, but these actions may not be very precise in the distribution of plausible robot actions given the current observation. A diffusion policy models



**Figure 2: Retargeted human hand trajectory to closed-loop robot action sequence, for the task  $\mathcal{T}$ : “close the laptop”.** The dotted line shows the trajectory of robot end-effector poses after kinematic retargeting. The olive contour plot depicts the distribution of trajectories from a pre-trained diffusion policy. Given a kinematic retargeting, we first perturb it with Gaussian noise and progressively remove the noise by simulating the reverse SDE with the diffusion policy. This process gradually projects a potentially unfeasible but approximately correct retargeting to the manifold of plausible robot actions that can perform real-world manipulation, in this case closing the laptop without missing the edge.

this likelihood, and we can use a pre-trained ‘*generalist*’ diffusion policy  $\bar{\pi}_\theta(\mathbf{a}_t | \mathbf{o}_{\leq t}, \mathcal{T})$  to refine the retargeted robot actions in a closed loop manner, thus inferring actions that, while still similar to the human demonstration, are more likely under the policy and better aligned with the robot embodiment (see Fig. 2). We describe both these steps in detail below.

### 3.2 Kinematic retargeting of Human Hand Trajectories to Robot End-Effector Poses

Given a human demonstration  $\mathbf{D}$  of a manipulation task  $\mathcal{T}$ , our first step is to extract the 3D hand pose trajectory  $\{\mathbf{h}_t\}_{t=0}^T$ . Each hand pose  $\mathbf{h}_t \in \mathbb{R}^{3 \times J}$  corresponds to the 3D locations of  $J$  keypoints (e.g., wrist and fingertips), which we estimate using a pre-trained monocular hand pose estimator applied to each video frame [47]. These keypoints encode the motion of the human performing the task and form the basis for retargeting to a robotic end-effector.

To translate this human motion into a robot-executable trajectory, we define a simple geometric mapping function  $f_{\text{retarget}} : \mathbb{R}^{3 \times J} \rightarrow \mathbb{R}^6$  that converts the human hand pose  $\mathbf{h}_t$  into a robot configuration  $\hat{\mathbf{a}}_t = f_{\text{retarget}}(\mathbf{h}_t)$ . The mapping aligns the wrist pose of the human to the robot end-effector pose. For a two-finger gripper, we use the distance between the thumb and the remaining fingers of the hand mesh to infer a binary robot grasp, and for a dexterous robotic hand we match the robot hand’s fingertip positions to those of a human hand using inverse kinematics. We set the initial configuration of the robot as the kinematically retargeted initial configuration from the human demonstration.

This kinematic retargeting procedure yields a full trajectory  $\{\hat{\mathbf{a}}_t\}_{t=0}^T$  in the robot’s configuration space, which can be executed as an open-loop policy. However, due to differences in morphology and embodiment between humans and robots, the absence of environment feedback, and the inaccuracy of the hand estimation module, this trajectory often leads to suboptimal or unstable behavior during open-loop execution. We thus treat this as the basis for searching for a plausible robot action trajectory via denoising with a (pre-trained) closed-loop diffusion policy.

### 3.3 Closed-Loop Denoising of Robot Actions with a Pre-Trained Diffusion Policy

A diffusion policy  $\bar{\pi}_\theta(\mathbf{a}_t | \mathbf{o}_{\leq t}, \mathcal{T})$  [48, 49] pre-trained on diverse offline robot interaction data reliably models the distribution of plausible robot actions  $\mathbf{a}_t$  given previous observations  $\mathbf{o}_{\leq t}$ . To predict

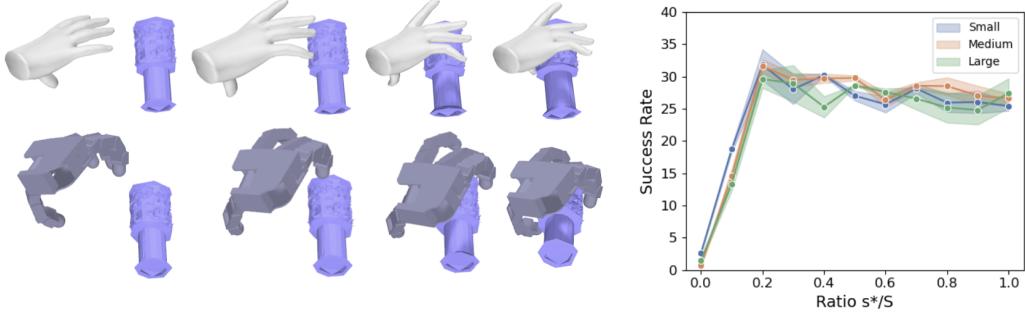


Figure 3: **Dexterous Grasping Results in simulation.** On the left, we show a human demonstration followed by a rollout with *DemoDiffusion* for dexterous grasping. On the right, we plot the average and standard deviation of success rates over 3 seeds for dexterous grasping, corresponding to different levels of the diffusion step  $s^*$ . Here,  $s^*/S = 0$  corresponds to kinematic retargeting and  $s^*/S = 1$  corresponds to the robot policy.

Table 1: **Simulated Dexterous Grasping Results.** The numbers show the average of success rates over 3 seeds. We use  $s^*/S = 0.2$ , where the robot policy uses  $S = 1000$ . *DemoDiffusion* consistently outperforms both baselines, across all groups.

Method	Small	Medium	Large	Average
Robot Policy	25.4	26.6	27.4	26.5
Kinematic Retargeting	2.6	0.7	1.5	1.6
<b><i>DemoDiffusion</i></b>	<b>31.8</b>	<b>31.6</b>	<b>29.6</b>	<b>31.0</b>

actions  $\mathbf{a}_t$ , diffusion policies start with Gaussian noise  $\tilde{\mathbf{a}}_t^{(S)} \sim \mathcal{N}(0, I)$  and iteratively denoise it at different diffusion steps  $s$ :

$$\tilde{\mathbf{a}}_t^{(s-1)} = \bar{\pi}_\theta \left( \tilde{\mathbf{a}}_t^{(s)}, \mathbf{o}_{\leq t} \right), \quad s = S, S-1, \dots, 0 \quad (1)$$

We use this to modify the kinematically retargeted trajectory  $\{\hat{\mathbf{a}}_t\}_{t=0}^T$  for obtaining robot actions that follow the high-level motion in the human trajectory and still lie within the distribution of plausible actions under  $\bar{\pi}_\theta$ . To do this, we modify this typical reverse diffusion process so that instead of starting with pure noise  $\tilde{\mathbf{a}}_t^{(S)}$  at step  $S$ , we start from an intermediate diffusion step  $s^*$  such that  $0 < s^* < S$ . We define  $\{\tilde{\mathbf{a}}_t^{(s^*)}\}_{t=0}^T$  to be a noisy version of the kinematically retargeted trajectory as follows:

$$\tilde{\mathbf{a}}_t^{(s^*)} = \sqrt{\alpha_{s^*}} \hat{\mathbf{a}}_t + \sqrt{1 - \alpha_{s^*}} \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2 \mathbf{I}) \quad (2)$$

Here  $\alpha$  corresponds to the diffusion schedule of the pre-trained policy. This procedure, inspired by SDEdit [15] which adopted a similar approach for image editing, relies on the assumption that  $\{\hat{\mathbf{a}}_t\}_{t=0}^T$  is an approximate version of the ideal action trajectory that should be executed by the robot, and it potentially lies outside the distribution of feasible actions under the pre-trained diffusion policy  $\bar{\pi}_\theta$ . The diffusion policy  $\bar{\pi}_\theta$  then performs iterative denoising steps, conditioned on the robot’s observations  $\mathbf{o}_{\leq t}$ , to refine this noisy trajectory into feasible robot actions, based on equation (2).

After  $s^*$  denoising steps, the final output  $\mathbf{a}_t = \tilde{\mathbf{a}}_t^{(0)}$  is deployed on the robot. Importantly, this process is carried out in a closed-loop manner: the policy uses real-time observations from the cameras in the scene to iteratively improve its predictions, thereby compensating for embodiment mismatch and external perturbations (e.g., object slippage or occlusion). The key hyperparameter in this process for *DemoDiffusion* is the diffusion step  $s^*$  that trades off between the faithfulness to the demonstration and the likelihood under the robot policy – in the limit  $s^* = S$  we recover a rollout from the base policy  $\bar{\pi}_\theta$  and in the limit  $s^* = 0$  we recover the kinematically retargeted trajectory  $\{\hat{\mathbf{a}}_t\}_{t=0}^T$ .

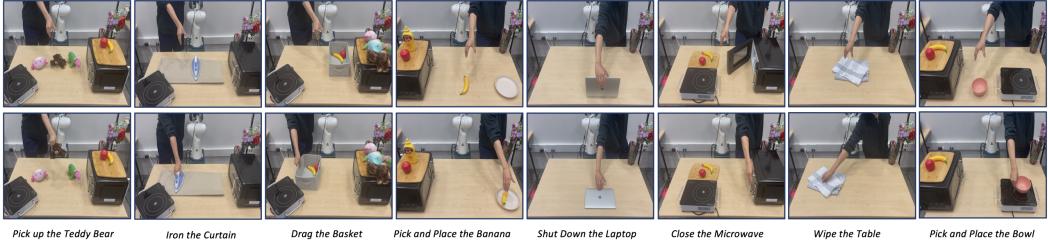


Figure 4: **Real-World Manipulation Tasks.** Each column corresponds to the respective human demonstrations for the real-world manipulation tasks, showing two frames from each demonstration.

## 4 Experiments

We evaluate our approach on dexterous grasping in simulation and across diverse real-world tasks comprising of prehensile and non-prehensile table-top manipulation. Through experiments, we aim to understand the following research questions:

- Can *DemoDiffusion* outperform pure kinematic retargeting from the human demonstration?
- Do the human demonstrations allow *DemoDiffusion* to perform new tasks where the pre-trained diffusion policy fails?
- How to effectively tradeoff between faithfulness to the human demonstration and performing the task reliably, with varying noise level  $s^*/S$ ?

### 4.1 Dexterous Grasping in Simulation

To verify our intuition, we first consider a simulation environment where the target task is restricted to picking up a generic object with a 16-DOF four-fingered Allegro hand. Specifically, we train a dexterous grasping policy across a small set of generic objects in simulation, and test our method on human hand grasping trajectories on a different set of objects. This serves as the pre-trained diffusion model for this experiment.

For training the robot policy, we collect total 985 grasping trajectories of Allegro hand over 58 training objects[50] (randomly sampled 26 and 32 objects from ShapeNet [51] and PartNet [52], respectively), resulting in 22.2 success rate. We use a variant of 3D Diffusion Policy [53], which takes 3D point tracks of an object instead of point clouds. At test time, we provide human grasping trajectories on a subset of Objvaerse dataset [54], which are unseen during training. The test set contains total 1220 objects, with 1 human grasping trajectory per object, provided by GraspXL [50]. Note that there is no error from hand estimation or the initial configuration of the object in this setup. We use ground-truth human hand 3D keypoints, and the object is located at the same location as the human demonstration.

**Results.** As in GraspXL [50], we group the objects based on their size, resulting in small, medium, and large objects. We compare *DemoDiffusion* with two baselines. The first is open-loop kinematic retargeting obtained from the human demonstration. The second is directly deploying the robot policy we trained.

Table. 1 shows the main results. Our method outperforms baselines, with more performance gain for small-sized objects. We hypothesize this roots from that larger objects are easier to grasp, supported by robot policy performance improving for larger sized objects. This experiment shows that even in a highly controlled setup where the robot environment is initialized identically to that of human demonstration, kinematic retargeting cannot solve the task, while slightly refining with a generalist grasping policy significantly improves the performance. Using kinematic retargeting to initiate the denoising process also helps the denoising process itself, outperforming the robot policy. Additionally, *DemoDiffusion* achieves higher inference speed, by the factor of  $S/s^*$ .

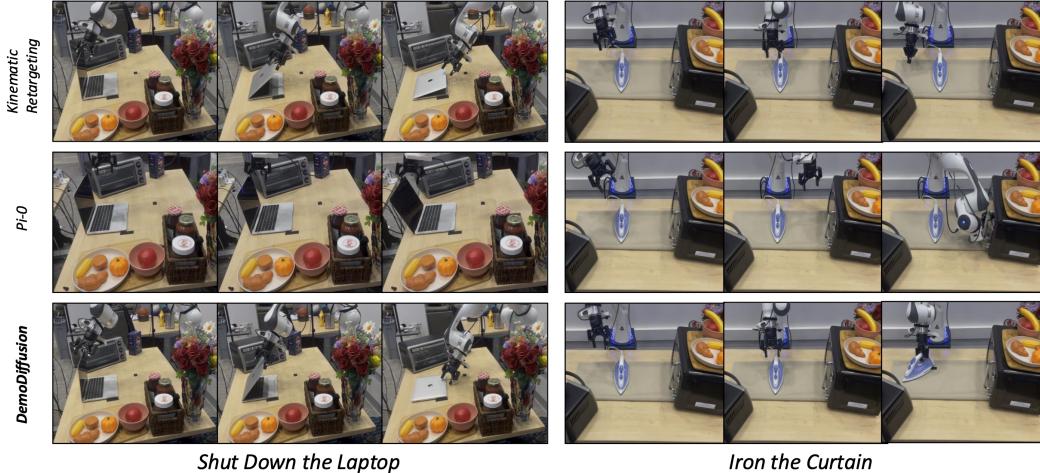


Figure 5: **Qualitative comparisons for real-robot manipulation.** We show rollouts from two tasks for *DemoDiffusion* and the baselines. The respective strips show progression of rollouts from left to right including the start frame, an intermediate frame, and the final frame. Frames are cropped for better visibility. Please refer to the website for detailed visualizations.

**Ablation studies.** We analyze the influence of  $s^*$  on *DemoDiffusion* in Fig. 3. Overall, we observe a consistent trend of performance gain from  $s^*/S = 1$  to  $s^*/S = 0.2$ , showing that *DemoDiffusion* is robust to the choice of hyperparameter.

## 4.2 Real-World Manipulation

For the real-world experiments, we use a Franka Emika Panda arm equipped with a two-finger gripper from Robotiq. We use the pre-trained diffusion policy  $\bar{\pi}_\theta(\mathbf{a}_t | \mathbf{o}_{\leq t}, \mathcal{T})$  called Pi-0 released by Physical Intelligence [1]. This policy takes as input a language task specification  $\mathcal{T}$ , and observations from two cameras in the scene, and outputs per-timestep joint velocity. This policy was trained on a large offline dataset of tele-operated robot demonstrations, and is a *generalist* policy that can perform a broad set of tasks. We do not fine-tune or adapt this policy in any way and directly use it for *DemoDiffusion*. We use an off-the-shelf monocular hand reconstruction model, Hamer [47] to obtain per-timestep hand mesh reconstructions from each of the external cameras given a human demonstration in the scene. Our experiments follow the protocol of a human first demonstrating an object manipulation in the scene, followed by robot execution after re-setting the scene.

**Workspace.** Fig. 6 describes our real-robot manipulation setup. The scene has 5 cameras including 4 Real-sense external cameras and 1 Zed-mini wrist camera mounted on the robot gripper. We calibrate the four external cameras and use them for triangulating the 3D hand pose so that we can obtain the position and orientation of a human hand in the world coordinates, with origin at the base of the robot. The pre-trained diffusion policy Pi-0 requires just two cameras (marked in purple) and the policy does not use any calibration information.

**Tasks.** We perform evaluations for 8 different manipulation tasks, the language description and the human demonstration for which are described in Fig. 4. Each of these tasks require a different manipulation strategy involving either prehensile or non-prehensile manipulation. We collect just



Figure 6: Workspace with 5 cameras. We use the four external cameras for triangulation to obtain the global pose of the hand mesh from a human demonstration. The pre-trained diffusion policy uses the two cameras marked in purple.

[47] to obtain per-timestep hand mesh reconstructions from each of the external cameras given a human demonstration in the scene. Our experiments follow the protocol of a human first demonstrating an object manipulation in the scene, followed by robot execution after re-setting the scene.

**Workspace.** Fig. 6 describes our real-robot manipulation setup. The scene has 5 cameras including 4 Real-sense external cameras and 1 Zed-mini wrist camera mounted on the robot gripper. We calibrate the four external cameras and use them for triangulating the 3D hand pose so that we can obtain the position and orientation of a human hand in the world coordinates, with origin at the base of the robot. The pre-trained diffusion policy Pi-0 requires just two cameras (marked in purple) and the policy does not use any calibration information.

**Tasks.** We perform evaluations for 8 different manipulation tasks, the language description and the human demonstration for which are described in Fig. 4. Each of these tasks require a different manipulation strategy involving either prehensile or non-prehensile manipulation. We collect just

Table 2: **Quantitative Results for Real-World Robot Manipulation.** Success rates (%) over 10 trials per task. Please refer to the Appendix for more comprehensive quantitative comparisons.

Method	Shut Laptop	Close Microwave	Drag Basket	Wipe Table	Iron Curtain	Pick up Bear	Pick Place Banana	Pick Place Bowl	Avg.
<i>Pi-0</i>	20	10	10	50	20	0	0	0	13.8
<i>Kinematic Retargeting</i>	10	80	<b>80</b>	0	70	40	70	70	52.5
<i>DemoDiffusion</i>	<b>60</b>	<b>90</b>	<b>80</b>	<b>100</b>	<b>90</b>	<b>60</b>	<b>90</b>	<b>100</b>	<b>83.8</b>

a single human demonstration per task but for robot execution, we make organic variations in the location of objects in the scene. Additionally, we add other objects in the scene as visual distractors, in order to see whether *DemoDiffusion* can generalize under such settings. All variations are consistent across the baselines and our method. Please refer to the Appendix for more explanations.

**Baselines.** For the real-world experiments we compare *DemoDiffusion* with two baselines. The first is open-loop kinematic retargeting obtained from the human demonstration. The second is directly deploying the Pi-0 model with the language instruction for the respective task. Since Pi-0 operates in the joint velocity space, for *DemoDiffusion* we first convert the kinematically retargeted trajectory to the joint velocity space with IK and then add noise to the trajectory, followed by denoising with Pi-0. We deploy *DemoDiffusion* with a heuristically selected noise level  $s^*/S$  depending on the task – preferring the kinematic trajectory with  $s^*/S = 0.2$  when Pi-0 has a zero success rate and  $s^*/S = 0.4$  otherwise. Note that in the limit of 0 noise added, we recover the first baseline, kinematic retargeting, and in the limit of maximum noise 1 we obtain the second baseline Pi-0.

**Qualitative Visualizations.** In Fig. 5 we show rollouts for two tasks across all the methods. We observe that Pi-0 is only able to perform general reaching behaviors but doesn’t follow the required manipulation trajectory necessary to perform these tasks, or confuse the target object to manipulate. On the other hand, we observe that (open-loop) kinematic retargeting from the human demonstration has a plausible motion of the robot end-effector, but it completely misses grasping the iron, and loses contact with the laptop before fully closing it. Compared to these, *DemoDiffusion* is able to reach the respective object and also follow the approximate motion of the retargeted trajectory while remaining in contact with the object in order to solve the tasks.

**Quantitative Comparisons.** In Table 2 we compare *DemoDiffusion* and the baselines across all the real-world manipulation tasks. We observe that *DemoDiffusion* either performs comparably or outperforms all the baselines in all the tasks. Interestingly, even for tasks like closing the laptop where both the baselines fail, *DemoDiffusion* is able to leverage the best of both the baselines (that represent extremes of our approach under the hyper-parameter  $s^*$ ) in order to achieve a non-zero success rate.

## 5 Discussion

We presented *DemoDiffusion*, a simple approach to leverage *generalist* diffusion models for imitating a human demonstration. By retargeting hand motion from the human demonstration into an approximate robot end-effector trajectory and modifying it through de-noising with a pre-trained diffusion policy, *DemoDiffusion* enables the robot to produce actions that are both aligned with the demonstrated intent and contextually plausible. Our method consistently outperforms both the base diffusion policy and the raw retargeted trajectory across diverse tasks in simulation and the real world, without requiring additional training, paired demonstrations, or reward functions. We envision that this simple approach can serve as a starting point for future efforts at human imitation across generic scenarios and perhaps also yield an improved exploration strategy for methods pursuing policy adaptation with online RL.

## Limitations

Our approach has several limitations. First, it assumes that the robot should act similarly to the human in order to successfully complete the task, which may not hold in scenarios requiring different strategies due to embodiment or environmental constraints. Second, while effective for one-shot imitation (and not requiring and training / test-time fine-tuning), the method does not produce a reusable task policy that can generalize across arbitrary variations of the task. Third, the quality of the retargeted trajectory is crucial: accurate 3D human motion capture is challenging and errors in retargeting can impact downstream performance. Additionally, the method implicitly assumes that the timing and speed of human and robot actions are aligned; extending the approach to allow for temporal alignment at test time is a promising direction of future work.

## Acknowledgements

We appreciate the helpful discussions with Yanbo Xu, Qitao Zhao, and Lucas Wu. We would also like to thank Kenny Shaw, Tony Tao, Jiahui Yang, Andrew Wang, Jason Liu, Hengkai Pan, Mohan Kumar Srirama for helping setting up the hardware. This work was supported by gift awards from CISCO and Google.

## References

- [1] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, et al.  $\pi\theta$ : A vision-language-action flow model for general robot control. In *RSS*, 2025.
- [2] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *CoRL*, 2023.
- [3] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *CoRL*, 2022.
- [4] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, et al. Openvla: An open-source vision-language-action model. In *CoRL*, 2024.
- [5] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. In *RSS*, 2023.
- [6] J. Li, Y. Zhu, Y. Xie, Z. Jiang, M. Seo, G. Pavlakos, and Y. Zhu. Okami: Teaching humanoid robots manipulation skills through single video imitation. In *CoRL*, 2024.
- [7] G. Papagiannis, N. Di Palo, P. Vitiello, and E. Johns. R+ x: Retrieval and execution from everyday human videos. In *ICRA*, 2025.
- [8] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. In *RSS*, 2024.
- [9] S. Chen, C. Wang, K. Nguyen, L. Fei-Fei, and C. K. Liu. Arcap: Collecting high-quality human demonstrations for robot learning with augmented reality feedback. In *ICRA*, 2025.
- [10] S. Bahl, A. Gupta, and D. Pathak. Human-to-robot imitation in the wild. In *RSS*, 2022.
- [11] I. Guzey, Y. Dai, G. Savva, R. Bhirangi, and L. Pinto. Bridging the human to robot dexterity gap through object-oriented rewards. In *ICRA*, 2025.

- [12] Y. Chen, C. Wang, Y. Yang, and C. K. Liu. Object-centric dexterous manipulation from human motion data. In *CoRL*, 2024.
- [13] X. Liu, J. Adalibieke, Q. Han, Y. Qin, and L. Yi. Dextrack: Towards generalizable neural tracking control for dexterous manipulation from human references. In *ICLR*, 2025.
- [14] X. Liu, K. Lyu, J. Zhang, T. Du, and L. Yi. Quasisim: Parameterized quasi-physical simulators for dexterous manipulations transfer. In *ECCV*, 2024.
- [15] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik. Reconstructing hands in 3d with transformers. In *ICLR*, 2022.
- [16] H.-S. Fang, H. Fang, Z. Tang, J. Liu, J. Wang, H. Zhu, and C. Lu. Rh20t: A robotic dataset for learning diverse skills in one-shot. In *ICRA*, 2024.
- [17] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *CoRL*, 2023.
- [18] 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. In *ICRA*, 2024.
- [19] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.
- [20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.
- [21] L. Zhu, G. Lin, J. Chen, X. Zhang, Z. Jin, Z. Wang, and L. Yu. Large images are gaussians: High-quality large image representation with levels of 2d gaussian splatting. In *AAAI*, 2025.
- [22] N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, and C. Rupprecht. Cotracker: It is better to track together. In *ECCV*, 2024.
- [23] C. Doersch, A. Gupta, L. Markeeva, A. Recasens, L. Smaira, Y. Aytar, J. Carreira, A. Zisserman, and Y. Yang. Tap-vid: A benchmark for tracking any point in a video. In *NeurIPS*, 2022.
- [24] Y. Rong, T. Shiratori, and H. Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *CVPR*, 2021.
- [25] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik. Reconstructing hands in 3d with transformers. In *CVPR*, 2024.
- [26] Y. Ye, A. Gupta, and S. Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *CVPR*, 2022.
- [27] Y. Ye, P. Hebbar, A. Gupta, and S. Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *ICCV*, 2023.
- [28] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018.
- [29] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

- [30] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *CVPR*, 2024.
- [31] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta. R3m: A universal visual representation for robot manipulation. In *CoRL*, 2022.
- [32] S. Parisi, A. Rajeswaran, S. Purushwalkam, and A. Gupta. The unsurprising effectiveness of pre-trained vision models for control. In *ICML*, 2022.
- [33] S. Karamcheti, S. Nair, A. S. Chen, T. Kollar, C. Finn, D. Sadigh, and P. Liang. Language-driven representation learning for robotics. In *RSS*, 2023.
- [34] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022.
- [35] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. In *ICLR*, 2023.
- [36] Y. J. Ma, V. Kumar, A. Zhang, O. Bastani, and D. Jayaraman. Liv: Language-image representations and rewards for robotic control. In *ICML*, 2023.
- [37] H. Bharadhwaj, R. Mottaghi, A. Gupta, and S. Tulsiani. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. In *ECCV*, 2024.
- [38] K. Shaw, S. Bahl, and D. Pathak. Videodex: Learning dexterity from internet videos. In *CoRL*, 2023.
- [39] M. K. Srirama, S. Dasari, S. Bahl, and A. Gupta. Hrp: Human affordances for robotic pre-training. In *RSS*, 2024.
- [40] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak. Affordances from human videos as a versatile representation for robotics. In *CVPR*, 2023.
- [41] S. Dasari and A. Gupta. Transformers for one-shot visual imitation. In *CoRL*, 2020.
- [42] E. Valassakis, G. Papagiannis, N. Di Palo, and E. Johns. Demonstrate once, imitate immediately (dome): Learning visual servoing for one-shot imitation learning. In *IROS*, 2022.
- [43] X. Zhang and A. Bouliarias. One-shot imitation learning with invariance matching for robotic manipulation. In *RSS*, 2024.
- [44] V. Jain, M. Attarian, N. J. Joshi, A. Wahid, D. Driess, Q. Vuong, P. R. Sanketi, P. Serbanet, S. Welker, C. Chan, et al. Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers. In *RSS*, 2024.
- [45] S. Park, S. Lee, M. Choi, J. Lee, J. Kim, J. Kim, and H. Joo. Learning to transfer human hand skills for robot manipulations. *arXiv preprint arXiv:2501.04169*, 2025.
- [46] T. G. W. Lum, O. Y. Lee, C. K. Liu, and J. Bohg. Crossing the human-robot embodiment gap with sim-to-real rl using one human demonstration. *arXiv preprint arXiv:2504.12609*, 2025.
- [47] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik. Reconstructing hands in 3d with transformers. In *CVPR*, 2024.
- [48] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *IJRR*, 2023.
- [49] M. Reuss, M. Li, X. Jia, and R. Lioutikov. Goal conditioned imitation learning using score-based diffusion policies. In *RSS*, 2023.

- [50] H. Zhang, S. Christen, Z. Fan, O. Hilliges, and J. Song. Graspxl: Generating grasping motions for diverse objects at scale. In *ECCV*, 2024.
- [51] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.
- [52] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *CVPR*, 2019.
- [53] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *RSS*, 2024.
- [54] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, 2023.
- [55] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. In *RSS*, 2024.
- [56] M. Lepert, J. Fang, and J. Bohg. Phantom: Training robots without robots using only human videos, 2025. URL <https://arxiv.org/abs/2503.00779>.
- [57] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. In *Transactions on Graphics*, 2017.

## A Real-World Manipulation

### A.1 Hardware Setup

We replicate DROID [55] setup for the hardware setup. Due to hardware availability, we use Realsense D455 instead of ZED 2 for the rear view camera.

### A.2 Task Setup Description

The overview of scenes for each task at test time is shown in Fig. 7. During data collection, we mark the target object location with a tape. During evaluation, we randomize the target object location within the tape, which is consistent across all methods. Overall, kinematic retargeting may fail due to small misalignment of object placement, inaccurate 3D hand keypoints, or camera extrinsics, etc. Pi-0 may fail if it cannot identify the object to interact with and manipulate other objects in the scene. It may also reach the target object but fail to manipulate in the desired way. Note that previous approaches [6, 7, 8, 9] use kinematic retargeting to transfer human demonstration to the robot.

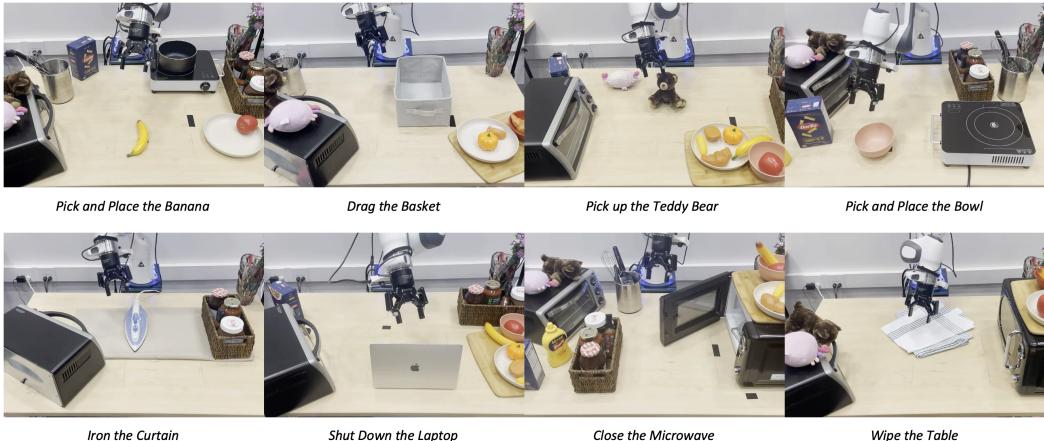


Figure 7: **Real-World Manipulation Tasks at Test Time.** Each image corresponds to an example of the scene for each task at test time. To test our method in a natural setup, we add other objects in the scene as visual distractors. For each rollout, we use the same set of distractors across methods.

**Pick and Place the Banana.** The task is to pick up the toy banana and place on the plate on the right side of the table. The episode is considered successful if the banana is successfully placed on the plate.

**Drag the Basket.** The task is to drag the basket to the right side of the table. The human demonstration grasps the edge of the basket and pulls it to the right. The episode is considered successful if the robot grasps the basket/ puts the gripper inside the basket and move it to any side of the table.

**Pick Up the Teddy Bear.** The task is to pick up the teddy bear located at the center of the table. The human demonstration grasps the neck of the bear and picks it up. The episode is considered successful if the robot picks up the bear by the end of the episode.

**Pick and Place the Bowl.** The task is to pick up the bowl and place on the stove on the right side of the table. The episode is considered successful if the bowl is successfully placed on the stove.

**Iron Curtain.** The task is to grasp the ironing machine located at the center of the table and slide it to the right side of the curtain. The episode is considered successful if the robot grasps/pushes the ironing machine to any part of the curtain.

**Shut Down the Laptop.** The task is to close the laptop located in front of the robot. The episode is considered successful if the angle of the laptop is smaller than 10 degrees at the end of the episode.

**Close the Microwave.** The task is to close the microwave located at the corner of the table. The episode is considered successful if the handle part touches the body part of the microwave at the end of the episode.

**Wipe the Table.** The task is to pinch/poke the given tissue and wipe the table. The episode is considered successful if the robot pinches/pokes the tissue and performs one stroke to any side of the table without missing it.

### A.3 Text Prompt.

We used the following text conditions for Pi-0 [1], which are also used for *DemoDiffusion*.

**Pick and Place the Banana:** *pick and place the banana on the plate*

**Drag the Basket:** *drag the basket to the right*

**Pick Up the Teddy Bear:** *pick up the bear doll*

**Pick and Place the Bowl:** *pick and place the bowl on the stove*

**Iron the Curtain:** *move the ironing machine to the right*

**Wipe the Table:** *wipe the table*

**Close the Microwave:** *close the microwave*

**Shut Down the Laptop:** *close the laptop*

### A.4 Inference Details

At each timestep  $t$ , we index the human demonstration at same timestep  $t$ , and compute the kinematically retargeted actions. As Pi-0 predicts a chunk of actions with length  $H = 10$ , we also use the action chunk computed from human demonstration from timestep  $t$  to  $t + H - 1$ . The episode length is set to be the same as the length(number of timesteps) of human demonstrations across all methods.

### A.5 Kinematic Retargeting Implementation Details

A human demonstration  $\mathbf{D}$  of a manipulation task  $\mathcal{T}$  consists of the 3D hand pose trajectory  $\{\mathbf{h}_t\}_{t=0}^T$ . Each hand pose  $\mathbf{h}_t \in \mathbb{R}^{3 \times J}$ , which corresponds to the 3D locations of  $J$  keypoints (e.g., wrist and fingertips). To kinematically retarget these 3D keypoints to a robotic end-effector, we use the human’s wrist position  $A$ , thumb tip position  $B$ , and the average of other fingertips  $C$ . Specifically,  $\vec{AB}$ , along with the cross product of  $\vec{AB}$  and  $\vec{AC}$  define the orientation of the end effector, and the human’s wrist position is used as the position of the end effector. As Pi-0 uses joint velocity as action space, the end effector pose is further converted into joint velocity using inverse kinematics. For the gripper, we use the norm of  $\vec{BC}$  as its width. To stabilize grasping, we set the gripper action to be fully closed if the norm of  $\vec{BC}$  is smaller than 80% of the gripper’s maximum width, similar to [56].

### A.6 Hyperparameters

To show our method’s applicability across diverse setups, we do not change the speed of the given human demonstration when performing kinematic retargeting. At inference, we use the following hyperparameters for Pi-0 and *DemoDiffusion*, with only difference on the number of denoising steps, as *DemoDiffusion* starts denoising from intermediate steps. Specifically, we use the following

hyperparameters. Depending on the noise level  $s^*/S$ , *DemoDiffusion* uses  $20 \times s^*/S$  denoising steps.

Table 3: Hyperparameters in Real-world.

Hyperparameter	Value
Open Loop Horizon	8
Predict Action Horizon	10
Denoising Steps at Inference (Pi-0)	20

## A.7 Additional Results

**Results with Different Hyperparameters.** We present the results of *DemoDiffusion* using two different hyperparameter settings,  $s^*/S = 0.2$  and  $s^*/S = 0.4$ . The results are summarized in Table 4. Overall, we observe a consistent trend based on the performance of the base diffusion policy (*Pi-0*):

- When the base policy *Pi-0* achieves a non-zero success rate,  $s^*/S = 0.4$  generally yields better performance. This is observed in tasks such as *shut laptop*, *close microwave*, *drag basket*, *wipe table*, and *iron curtain*.
- Conversely, when *Pi-0* has a zero success rate,  $s^*/S = 0.2$  tends to perform better. This is the case for tasks including *pick up bear*, *pick place banana*, and *pick place bowl*. We observe that while  $s^*/S = 0.4$  shows adaptive end effector motion, it tends to avoid activating the gripper at the correct timestep.

These findings suggest that a more effective base policy enables *DemoDiffusion* to benefit from a higher noise level, as it can leverage the base policy’s strengths more effectively. In contrast, when the base policy is less reliable, a lower noise level tends to show better performance. We believe this can serve as a guideline for choosing the noise level  $s^*/S$  for *DemoDiffusion*.

Table 4: Additional Quantitative Results for Real-World Robot Manipulation.

Method	Shut Laptop	Close Microwave	Drag Basket	Wipe Table	Iron Curtain	Pick up Bear	Pick Place Banana	Pick Place Bowl	Avg.
<i>Pi-0</i>	20	10	10	50	20	0	0	0	13.8
<i>Kinematic Retargeting</i>	10	80	<b>80</b>	0	70	40	70	70	52.5
<i>DemoDiffusionw</i> ( $s^*/S = 0.2$ )	40	80	70	60	70	<b>60</b>	<b>90</b>	<b>100</b>	<b>71.3</b>
<i>DemoDiffusionw</i> ( $s^*/S = 0.4$ )	<b>60</b>	<b>90</b>	<b>80</b>	<b>100</b>	<b>90</b>	40	20	0	60.0

**Zero-Shot Generalization to New Objects.** Lastly, we investigate whether *DemoDiffusion* can achieve zero-shot generalization to new objects within the same category. We use the same human demonstration and text prompt we previously used while evaluating on a new object within the same category. Results are shown in Fig. 8. Although we only provide human demonstration on a different object, we observe that *DemoDiffusion* can adapt its behavior to a new object.

## B Dexterous Grasping in Simulation

### B.1 Task Setup Description

In our simulation experiment, the task is to pick up the object. The rollout is considered success when the object’s z position is higher than  $0.10\text{ m}$  at the end of the episode. Similar to the real-world experiments, at each timestep of inference, we compute a kinematically retargeted action chunk computed from human demonstration at the same timestep.



Drag the Basket

Wipe the Table

Pick and Place the Bowl

**Figure 8: Zero-Shot Generalization to New Objects.** We evaluate *DemoDiffusion*’s generalization capabilities to new objects on 3 tasks - drag the basket, wipe the table, pick and place the bowl. While the objects have different colors and shapes from those of objects in the human demonstration, *DemoDiffusion* adjusts its behavior to these new objects. Please refer to the website for detailed visualizations.

**Training Data.** As we don’t have a *generalist* pre-trained diffusion policy in simulation, we first train a diffusion policy while constraining the target task to picking up generic objects. Specifically, we collected total 985 grasping trajectories of Allegro hand over 58 objects in the RaiSim simulator for training. The trajectories were generated by rolling out the expert Allegro RL policy with random grasping directions. The 58 objects are randomly sampled from ShapeNet [51] (26 objects) and PartNet [52] (32 objects). The training objects and expert Allegro RL policy are provided by GraspXL[50].

**Test Data.** At test time, we provide human grasping trajectories for a different set of objects. As it is challenging to provide human demonstration in simulation, we instead use MANO [57] hand RL policy to collect human demonstrations. The test set contains total 1220 objects, which is a subset of Objvaerse dataset [54]. The objects are divided in three groups based on its size: small scale  $s \in [3, 5]cm$  medium scale  $m \in [5, 7]cm$ , and large scale  $l \in [7, 9]cm$ . As the object meshes don’t contain material information, we assume fixed density, resulting in diverse mass. We additionally use the same friction coefficient for all objects. Again, test objects and expert MANO [57] RL policy are provided by GraspXL[50]. Additionally, we do not change the speed of the given human demonstration when performing kinematic retargeting.

## B.2 Kinematic Retargeting Implementation Details

Following [8], we use inverse kinematics to match the fingertip positions between the human hand and the robot hand, ignoring the pinky finger. The pose of the robot hand base is initialized to be the same as that of a human hand.

## B.3 Robot Policy Implementation Details

We train a hierarchical policy, where the high-level planner predicts the robot’s future goal state, and the low-level controller predicts the action given the desired future robot state. Both policies were trained using a variant of 3D Diffusion Policy [53], where we modify the point cloud encoder to take point tracks, for a richer representation of the object. Specifically, while the original point encoder takes object points at each timestep, the modified encoder takes a sequence of object points (object tracks) as input.

High-level planner and low-level controller use a subset of the following as input and output. All rotations are represented in Euler angles.

**Robot state** contains its wrist position in the world frame, rotation in the world frame, and absolute hand joint angles.

**Object state** contains 3D points of its surface. For a history of robot state, we assume 3D point tracks to be given(i.e. each 3D point across timesteps corresponds to the same point of the object).

**Robot goal state** contains its wrist position in the local frame(wrist frame at previous timestep), rotation in the local frame, and relative hand joint angles to the previous timestep.

**Action** contains the desired robot’s wrist position in the local frame(previous wrist frame), desired rotation in the local frame, and desired relative hand joint angles to the previous timestep. Note that the action and the robot’s goal state are not the same. For instance, the desired hand joint angles can penetrate the object surface, while the robot’s state cannot. Such a difference creates contact force, enabling a stable grasp.

**High-level Planner** The high-level planner takes in a history of robot states and object states, and predicts a sequence of robot goal states. At test-time, we use the kinematic retargeting result as an initial estimate for the robot’s goal state. We then apply *DemoDiffusion* to refine the robot goal state, and feed it to the low-level controller.

**Low-level Controller** The low-level controller takes in a history of robot states and a sequence of robot goal states, and predicts a sequence of robot actions. The goal state encoder has the same architecture to robot state encoder. Once it outputs the action, the action is used as a position target for a PD controller provided in the RaiSim simulator.

#### B.4 Hyperparameters

For policy training, we use the default hyperparameters provided in the 3D Diffusion Policy [53] with minimal modifications. For the simulation environment, we use the default hyperparameters provided by GraspXL [50]. Specifically, we use the following same parameters for both the high-level planner and low-level controller.

Table 5: Hyperparameters in Simulation.

Hyperparameter	Value
$H$	4
$N_{obs}$	2
$N_{act}$	1
$N_{goal}$	2
$N_{latency}$	0
Number of Object Points	1024
Train Timesteps	1000
Inference Timesteps	10
Episode Length	155
Wrist Position Pgain	100.0
Wrist Position Dgain	0.1
Wrist Rotation Pgain	100.0
Wrist Rotation Dgain	0.2
Finger Rotation Pgain	100.0
Finger Rotation Dgain	0.2