

Phantom: Training Robots Without Robots Using Only Human Videos

Marion Lepert¹, Jiaying Fang¹, Jeannette Bohg¹

¹Stanford University

Collect human videos in many scenes

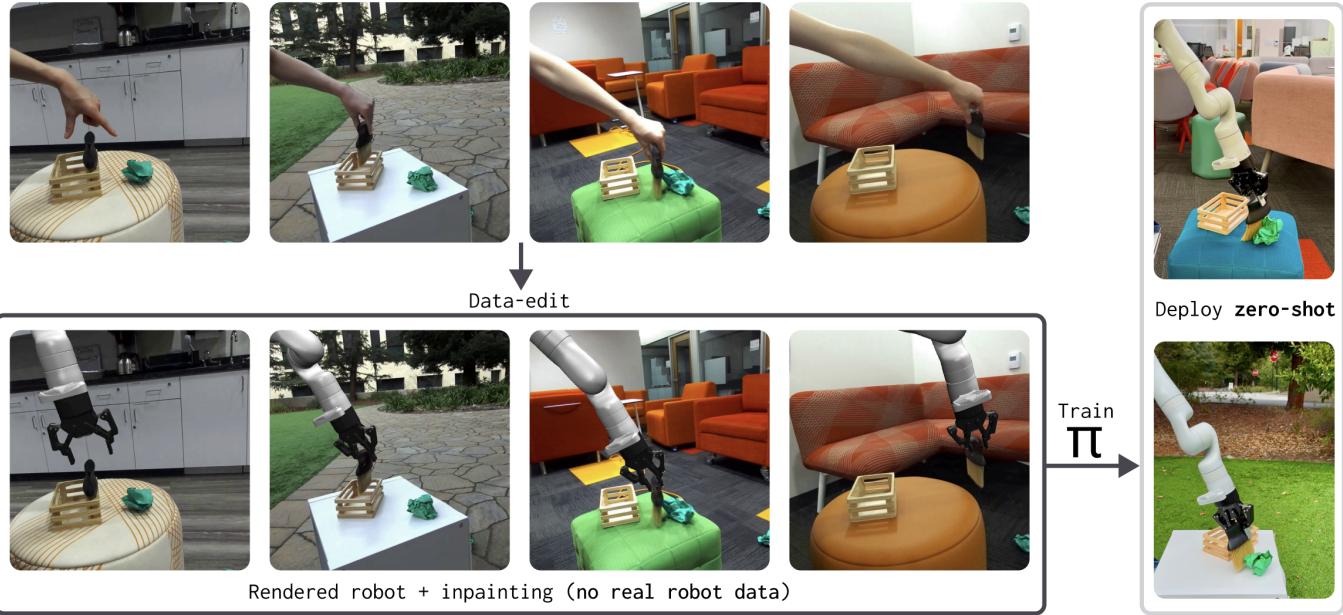


Fig. 1: **Overview of learning from human videos.** Our method enables training robot policies without collecting any robot data. We first collect human video demonstrations in diverse environments and use inpainting to remove the human hand. A rendered robot is then inserted into the scene using the estimated hand pose. The resulting augmented dataset is used to train an imitation learning policy, which is deployed zero-shot on a real robot.

Abstract—Scaling robotics data collection is critical to advancing general-purpose robots. Current approaches often rely on teleoperated demonstrations which are difficult to scale. We propose a novel data collection method that eliminates the need for robotics hardware by leveraging human video demonstrations. By training imitation learning policies on this human data, our approach enables zero-shot deployment on robots without collecting any robot-specific data. To bridge the embodiment gap between human and robot appearances, we utilize a data editing approach on the input observations that aligns the image distributions between training data on humans and test data on robots. Our method significantly reduces the cost of diverse data collection by allowing anyone with an RGBD camera to contribute. We demonstrate that our approach works in diverse, unseen environments and on varied tasks. Videos are available at <https://phantom-human-videos.github.io>.

I. INTRODUCTION

Data scarcity remains a key challenge in advancing robotics research. While large-scale data collection efforts are gaining momentum, even the largest robotics datasets [1, 7] are significantly smaller than those used to train generalist models in natural language processing and computer vision. These

efforts are constrained by the slow and costly process of collecting data with robotics hardware. Moreover, increasing data quantity alone is insufficient—diversity in the data is equally critical [13]. Moving physical robots to many new environments to collect enough diversity to train a generalist robot policy remains a formidable challenge.

We propose an approach that does not require any robotics hardware, and instead relies exclusively on collecting human video demonstrations. While collecting robot demonstrations is slow, requires expensive hardware, and poses logistical challenges for achieving diverse scenes, collecting human video demonstrations is fast, cheap, and scalable. Our method converts human videos into data-edited “robot” demonstrations by extracting actions using a hand pose estimator and replacing the human arm with a rendered robot. We then train an imitation learning policy on these “robot” demonstrations and deploy our policy zero-shot on a robot in a new scene, without the need to collect any robot data. We demonstrate that our method works on six tasks, including one demonstrating generalization to new scenes.

Learning from human videos presents significant chal-

lenges: these videos lack action labels, and humans look very different from robots. Prior methods trying to leverage human video demonstrations typically rely on co-training with robot data or reinforcement learning. Put simply, such approaches fail to extract sufficient learning signals from human data alone, necessitating robot data to bridge the gap.

Recent data-editing techniques have shown impressive success in cross-embodiment robot-to-robot policy transfer. However, these methods rely on precise proprioception and action labels, and until now, have not been effectively adapted to the more challenging human-to-robot setting. Our method does exactly this, leveraging a simple yet effective data-editing strategy for human-to-robot transfer. That such a straightforward method works and generalizes across diverse environments underscores a striking and perhaps unexpected insight: human demonstrations alone, when subject to simple data editing, can be directly leveraged for training a robot policy.

Our method dramatically reduces the cost of scaling data collection across diverse scenes compared to robot teleoperation, while remaining easy to implement. Data collectors can collect demos with their own hands, improving ergonomics and avoiding the challenges of carrying bulky hardware required by other methods [11], [32]. Additionally, our data is robot agnostic (see Fig. 9), meaning that it can be used on many different robots. Although our data is inherently less precise than that collected via teleoperation, we demonstrate that it can still achieve a high success rate across a wide range of tasks using only human video demonstrations. By trading some precision for scalability, our method eliminates the dependency on robot hardware and the logistical challenges of moving it to multiple locations. This enables anyone with access to an RGBD camera, anywhere in the world, to contribute to data collection for robotics.

To summarize, our main contribution is demonstrating that data-editing-based cross-embodiment learning techniques are adaptable to human-to-robot transfer, which unlocks their utility for collecting larger and more diverse datasets than can be achieved using traditional robot teleoperation.

II. RELATED WORKS

A. Data Collection Methods

Data-driven robotics relies on expert demonstrations to train effective policies. Prior methods have used a variety of interfaces to collect data including using a 3D SpaceMouse [46], VR or AR controllers [19, 17], leader-follower devices such as ALOHA [45] and GELLO [39], and smartphones [38, 25]. These methods rely on controlling a real robot to collect data, limiting the scale and diversity of data that can be collected with them. Other approaches like the UMI gripper [11] and the DOBB-E gripper [32] use a portable, hand-held gripper to collect data anywhere in the world without having to move an actuated robot. However, these methods require data collectors to carry bulky hardware, creating barriers to adoption. In contrast, our method allows data collectors to use their own hands and requires only an RGBD camera.

B. Learning from Human Videos

Many works have explored leveraging human videos to improve robot policies. A prominent line of research focuses on using diverse in-the-wild videos (e.g., YouTube) to improve generalization. Common strategies include pre-training visual representations [20, 29, 40, 26], learning reward functions [33, 8, 24], and predicting object motion [6, 4]. Despite their potential, these methods struggle to overcome the wide embodiment gap between humans and robots and still rely on extensive robot data.

Alternative approaches leverage curated human video demonstrations, which simplify the problem by ensuring that the videos explicitly show task-relevant behaviors. While these videos must be manually collected, they are faster to gather compared to robot teleoperation data. Some works use human video demonstrations to learn motion priors [3, 36, 42, 5]. MimicPlay [36] trained a high-level planner using human videos alongside a plan-guided imitation learning policy trained with robot demonstrations. Other methods [17, 16] rely on paired human video and robot data to bridge the embodiment gap.

Object-centric approaches have also been explored as an alternative direction [15, 14, 2]. These methods typically estimate and track target object poses from video demonstrations, and learn a robot policy conditioned on the extracted object trajectories [47]. However, they require identifying objects of interest and estimating rigid-body transformations which makes it hard to apply them to scenarios with deformables or multiple objects. Flow-based methods [41, 27, 37, 43, 31] address some of these limitations by tracking trajectories of points instead of rigid body transformations. Wen et al. [37], Ren et al. [31] track points on the human embodiment, which provides information on the general direction the robot should move in, but because humans and robots move in different ways, these methods still require robot data to refine the motion. Conversely, Xu et al. [43] only tracks flow on the manipulated object, but relies on object detection and simulation environments to refine robot motions. The approach in [27] exclusively uses human data but is limited to open-loop execution. In contrast, as summarized in Table I, our method is closed-loop, not bottlenecked by an object detector, and works equally well on rigid, deformable, and multiple objects.

C. Data Editing for Cross-Embodiment Learning

While many works focus on human-to-robot transfer, robot-to-robot cross-embodiment learning is also gaining attention as large-scale robotics datasets increasingly incorporate diverse data sources. Vision-based policies face significant challenges with cross-embodiment learning due to distribution shifts caused by the varying appearances of different embodiments. To address this, several methods propose data-editing strategies to mitigate these shifts. RoviAug [9] uses inpainting during training to remove the source embodiment from images and overlays a virtual rendering of the target embodiment in the same pose. At test time, the policy is deployed directly on the target embodiment. Shadow [21] replaces both the

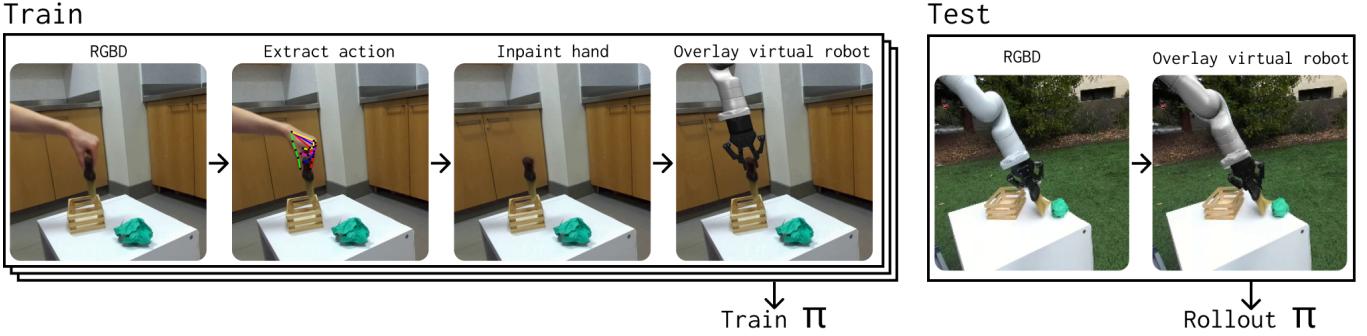


Fig. 2: **Overview of our data-editing pipeline for learning robot policies from human videos.** During training, we first estimate the hand pose in each frame of a human video demonstration and convert it into a robot action. We then remove the human hand using inpainting and overlay a virtual robot in its place. The resulting augmented dataset is used to train an imitation learning policy, π . At test time, we overlay a virtual robot on real robot observations to ensure visual consistency, enabling direct deployment of the learned policy on a real robot.

	No Robot Data	Deformable Objects	Closed-loop
WHIRL [3]	✗	✓	✓
Track2Act [6]	✗	✗	✓
HOPMan [5]	✗	✓	✓
Mimicplay [36]	✗	✓	✓
Xskill [42]	✗	✓	✓
ORION [47]	✓	✗	✓
Im2Flow2Act [43]	✗*	✓	✓
MotionTracks [31]	✗	✓	✓
AR2-D2 [12]	✗	✓	✓
R+x [27]	✓	✓	✗
Ours	✓	✓	✓

TABLE I: Comparison between our method and other related works. **No Robot Data:** the method does not require robot data in policy training. **✗*** indicates that the method relies on simulation data which is limited by the need to create simulation environments that are representative of real world interactions. **Deformable Objects:** the method is demonstrated to work on deformable objects. **Closed-loop:** the method is closed-loop.

source and target robots with composite segmentation masks at train and test time, ensuring a close match between the input data distributions. Other methods, such as EgoMimic [18] and AR2-D2 [12], adapt data-editing techniques for human-to-robot transfer. EgoMimic masks out each embodiment and overlays a red line along each arm, while AR2-D2 employs the same inpainting and virtual overlay strategy as [9]. However, both methods still rely on co-training with robot data to bridge the human-to-robot embodiment gap.

III. APPROACH

A. Problem Setup

We assume access to a dataset $\mathcal{D}_{\text{human}} = \{\tau_h^i\}_{i=1}^N$ of N human (h) video demonstrations τ_h^i of a manipulation task. Each demonstration consists of a sequence of images $\{I_{h,t}\}_{t=1}^T$ captured from a third-person viewpoint using an RGBD camera. The demonstration is performed using a pinch grasp with the thumb and index finger.

Our goal is to use only these human video demonstrations to train a closed-loop policy using imitation learning that can

be deployed zero-shot in an out-of-distribution scene on a target robot (r) for which no data has ever been collected. To do so, we use a data-editing strategy to convert our dataset $\mathcal{D}_{\text{human}}$ into $\mathcal{D}_{\text{robot}} = \{\tau_r^i\}_{i=1}^N$. Our objective is to convert each frame of a human demonstration into a corresponding robot observation-action pair: $I_{h,t} \rightarrow (I_{r,t}, a_{r,t})$. Importantly, the goal of data-editing is for each $I_{r,t}$ to be drawn from the same distribution as images at test time on the target robot. Then, we can simply train our imitation learning policy on $\mathcal{D}_{\text{robot}}$ and deploy it on test-time robot observations.

Each robot action $a_{r,t}$ consists of the position and orientation of the end-effector and the opening width of the gripper:

$$a_{r,t} = (\mathbf{p}_t, \mathbf{R}_t, g_t) \quad (1)$$

where:

- $\mathbf{p}_t \in \mathbb{R}^3$ is the Cartesian position of the end-effector.
- $\mathbf{R}_t \in \mathbb{R}^6$ represents the orientation of the end-effector using a 6D continuous rotation representation.
- $g_t \in [0, 1]$ is the normalized opening width of the gripper, where 0 corresponds to fully closed and 1 to fully open.

While the scenes at train and test time do not need to match, we assume that the height and angle of the camera used to collect videos is similar to that of the camera used to deploy the policy. This requirement could be alleviated by collecting more data from a wide range of angles and heights, as done in [19], but this amount of data collection is outside the scope of this paper. We also assume that the extrinsics of the camera used to deploy the policy on the robot are known.

B. Action Labeling of Human Videos

Since the human videos lack explicit action information, we first address how to go from a frame of the human video $I_{h,t+1}$ to the corresponding robot action for the previous frame $a_{r,t} = (\mathbf{p}_t, \mathbf{R}_t, g_t)$.

First, we estimate the hand's pose at each timestep. We apply HaMeR [28] to each frame $I_{h,t}$ to obtain a 3D hand pose estimate. HaMeR predicts 21 keypoints, $\hat{\mathbf{X}}_t \in \mathbb{R}^{21 \times 3}$, corresponding to anatomical landmarks, along with a dense set of 778 vertices, $\hat{\mathbf{V}}_t \in \mathbb{R}^{778 \times 3}$, representing the hand mesh.

While HaMeR accurately captures hand shape, it struggles to estimate the absolute 3D pose due to its reliance on a monocular image. To refine this estimate, we incorporate depth data. First, we segment the hand in the RGB image using SAM2 [30], yielding a segmentation mask M_t . Using M_t and the corresponding depth image D_t , we extract a partial hand point cloud, \mathbf{P}_t . We then align the HaMeR-predicted mesh $\hat{\mathbf{V}}_t$ with \mathbf{P}_t via Iterative Closest Point (ICP) registration, obtaining the optimal rigid transformation $\mathbf{T}_t \in SE(3)$ such that $\mathbf{P}_t \approx \mathbf{V}_t = \mathbf{T}_t \hat{\mathbf{V}}_t$ (see Fig. 3). Since $\hat{\mathbf{V}}_t$ and $\hat{\mathbf{X}}_t$ are internally consistent, we can apply \mathbf{T}_t to the predicted keypoints to refine their positions: $\mathbf{X}_t = \mathbf{T}_t \hat{\mathbf{X}}_t$.

HaMeR also struggles with keypoints that are occluded in the RGB image—an issue exacerbated during grasping. Since HaMeR models all hand joints as ball joints, it often predicts unrealistic finger configurations under occlusion. To address this, we constrain the last two joints of the thumb and index fingers to a single degree of freedom, limiting their movement to anatomically feasible ranges. This ensures more accurate finger pose estimation when occlusions occur.

We use the refined keypoints \mathbf{X}_t to define a target action for our policy, visualized in Fig. 3:

- The target position, \mathbf{p}_t , is set as the midpoint between the keypoints at the tips of the thumb, $\mathbf{x}_t^{\text{thumb, tip}}$, and index finger, $\mathbf{x}_t^{\text{index, tip}}$.
- For the target orientation, \mathbf{R}_t , we fit a plane through all the keypoints of the thumb $\mathbf{x}_t^{\text{thumb}}$ and index finger $\mathbf{x}_t^{\text{index}}$ and compute a principal axis by fitting a vector through the keypoints of the thumb. \mathbf{R}_t is then defined using the normal of this plane and the fitted vector.
- The gripper opening g_t is computed as the distance between the keypoints corresponding to the fingertips of the thumb and index finger, $\mathbf{x}_t^{\text{thumb, tip}}$ and $\mathbf{x}_t^{\text{index, tip}}$. To mitigate slippage during grasping, we enforce a threshold, setting the bottom 20th percentile of predicted gripper distances in a single trajectory to fully closed.

HaMeR predicts keypoints in the camera’s reference frame, meaning that p_t and \mathbf{R}_t are also expressed in this frame. We convert them into the robot’s frame using the known camera extrinsics of our target setup to obtain the final robot action $a_{r,t}$.

C. Bridging the Visual Observation Gap

Human arms and hands appear visually distinct from robot arms and grippers. A vision-based policy trained solely on human demonstrations struggles to generalize to a robot embodiment. To address this, we adapt the data-editing scheme from Rovi-Aug [9] to the human-to-robot transfer setting to compute $I_{h,t} \rightarrow I_{r,t}$. The edited images are used to train an imitation learning policy, which is then deployed on the target robot.

1) *Data-editing at train time*: Each frame in the training dataset contains an image of a human arm performing a task. To replace the human embodiment with a robot, we first segment out the pixels corresponding to the human arm using SAM2 [30], and then remove the segmented arm via inpainting

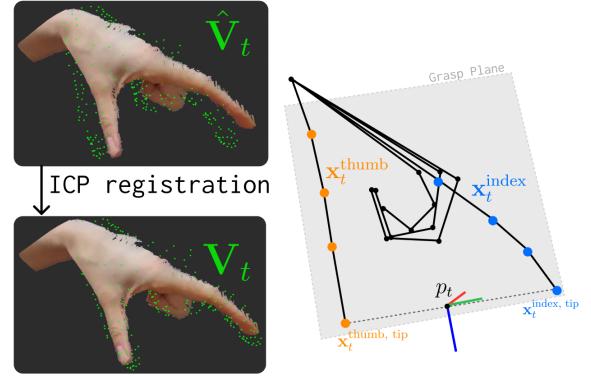


Fig. 3: **Left**: We use HaMeR to estimate the pose of the hand at each timestep. To refine the HaMeR predicted mesh points $\hat{\mathbf{V}}_t$ shown in green, we use ICP registration to align them with the partial point cloud of the hand, \mathbf{P}_t to obtain \mathbf{V}_t . **Right**: After aligning the HaMeR keypoints with the hand point cloud, we calculate the target position \mathbf{p}_t as the midpoint between the tips of the thumb and index finger and the target orientation by fitting a plane through the points of the thumb and index fingers.

using E2FGVI [22]. Next, we render a virtual model of the target robot with its end effector in the corresponding pose, obtained from Section III-B (i.e., its end effector pose at $(\mathbf{p}_t, \mathbf{R}_t, g_t)$). Given the known camera extrinsics, we synthesize an image of the robot from the appropriate viewpoint and overlay it onto the original image. To ensure realistic occlusions, we use depth data to determine which parts of the overlaid robot should be masked by objects in the environment. The final result is an image that closely resembles a real robot completing the task, as illustrated in Figure 2.

2) *Data-editing at inference time*: At inference time, each observation image contains a real robot arm. However, training images feature a rendered robot arm, which may have slight discrepancies in color and texture. To minimize domain shift, we overlay a virtual robot arm onto the real robot in each observation image, ensuring consistency between train and test distributions. An alternative approach, as proposed in Rovi-Aug, is to introduce color variations in the overlays during training to make the policy robust to these shifts. However, since this strategy has already been explored in prior work, we opt for the simpler inference-time overlay approach.

IV. RESULTS

We evaluate our method across a range of tasks that highlight the versatility of our method. To demonstrate that our method works across different robots, we present results on both a Franka and a Kinova robot. For imitation learning, we use Diffusion Policy [10]. Virtual robot renderings are generated using Mujoco [34] with models from Mujoco Menagerie [44].

A. Comparison of Data Editing Methods

To the best of our knowledge, no existing work has trained closed-loop imitation learning policies using only human video demonstrations that can manipulate rigid objects, deformable

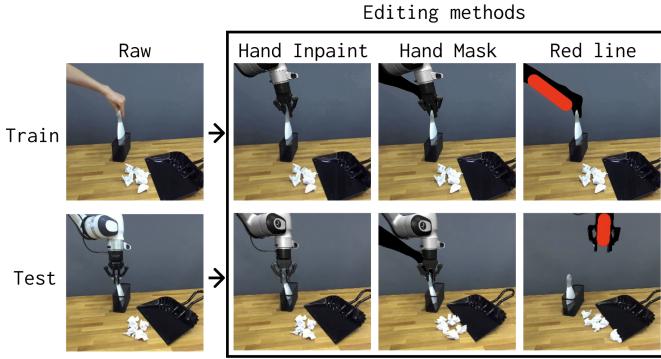


Fig. 4: **The different data-editing strategies we compare for human-to-robot transfer.** We evaluate three data-editing approaches: (1) **Hand Inpaint**, where the human hand is removed via inpainting and replaced with a rendered robot; (2) **Hand Mask**, where the human hand is blacked out during training, and a rendered robot is overlaid on top. At test time, a black mask of a human arm is added to match the training distribution; and (3) **Red Line**, where the human arm is blacked out and replaced with a red line during training, and at test time, the robot arm is blacked out and similarly overlaid with a red line. Both Hand Inpaint and Hand Mask achieve high success rates, but Hand Inpaint produces more realistic images and allows for faster rollouts.

objects, and groups of objects. Therefore, we focus our experiments on identifying the most effective data-editing strategy for human-to-robot policy transfer. We evaluate the following approaches (see Fig. 4):

1) **Hand Inpaint:** we adapt the data-editing strategy from Rovi-Aug [9], which was developed for the simpler robot-to-robot setting, to the human-to-robot setting. During training, the human arm is segmented out and replaced with inpainting. An image of the target robot is synthesized using a virtual model from the appropriate viewpoint and overlaid onto the original image. At test time, a rendered robot arm is overlaid onto the real robot arm to minimize domain shift. See Section III-C for more details.

2) **Hand Mask:** We adapt the data-editing strategy from Shadow [21]. This method was also developed for the simpler robot-to-robot setting. During training, the human arm is masked out, and a virtual robot in the same pose is overlaid. While the original method overlays a black mask of the robot, we use an RGB image for a cleaner comparison with Hand Inpaint. At test time, a hand mask generated by a trained diffusion model is applied, and a virtual robot is overlaid on the real robot. See Appendix-B for more details.

3) **Red Line:** EgoMimic [18] proposes a data-editing approach for learning robot policies from egocentric human videos. While we do not directly compare with their full method, as it requires robot data, we evaluate their data-editing strategy. The human arm is masked out in black during training and overlaid with a red line along its length. At test time, the robot arm is similarly blacked out, with a red line overlaid in the same manner.

4) **Vanilla:** We also compare to a baseline that does not modify the train or test images in any way.

B. In-distribution Scene

We start by evaluating how well our method can transfer a policy trained exclusively on data-edited human video demonstrations in a single scene to a robot in the same scene. This evaluates how well our method bridges the physical and visual embodiment gap between human and robot without the added complexity of testing scene generalization. We fix a camera in a scene, collect human video demonstrations, train a policy on data-edited videos, and deploy our trained policy using the same camera. We evaluate our method on five tasks that highlight the diversity of skills our method can learn. For each task we collect between 250-350 demonstrations (see Appendix-A).

- **Pick and Place Book:** The robot must pick up a book and place it inside a wooden container.
- **Stack Cups:** The robot must stack the green cup inside the purple cup. Precise alignment is critical, as the cups differ in diameter by only 1.5 cm.
- **Sweep Trash:** The robot must pick up a sweeper and sweep six pieces of trash into a dustpan. This task involves coordinated multi-object manipulation, requiring the robot to control the sweeper while simultaneously managing the movement of multiple loose objects. Additionally, the pieces of trash exhibit unpredictable dynamics, necessitating continuous adaptation based on real-time feedback.
- **Tie Rope:** The robot must tie a simplified cleat hitch, a sailing knot that follows a figure-eight ∞ pattern. This task is challenging due to the precise manipulation required of a highly deformable object.
- **Rotate Box:** The robot must rotate a box 90 degrees onto a new face in a controlled fashion (simply knocking it over is not valid).

	Pick/ Place Book	Stack Cups	Tie Rope	Rotate Box
Hand Inpaint	0.92	0.72	0.64	0.72
Hand Mask	0.92	0.52	0.60	0.76
Red Line	0.0	0.0	0.0	0.0
Vanilla	0.0	0.0	0.0	0.0

TABLE II: **In-distribution scene results:** Both Hand Inpaint and Hand Mask achieve high success rates across all tasks, with Hand Inpaint performing the best overall. The Red Line strategy fails to achieve success on any task, as does the Vanilla baseline. 25 rollouts per evaluation.

Hand Inpaint and Hand Mask achieve high success rates across all tasks. However, Hand Mask takes on average 73% longer to rollout due to having to run an additional diffusion model at test time to generate the hand masks. The Red Line data-editing strategy fails to complete any tasks, indicating that it does not adequately bridge the visual embodiment gap between humans and robots.

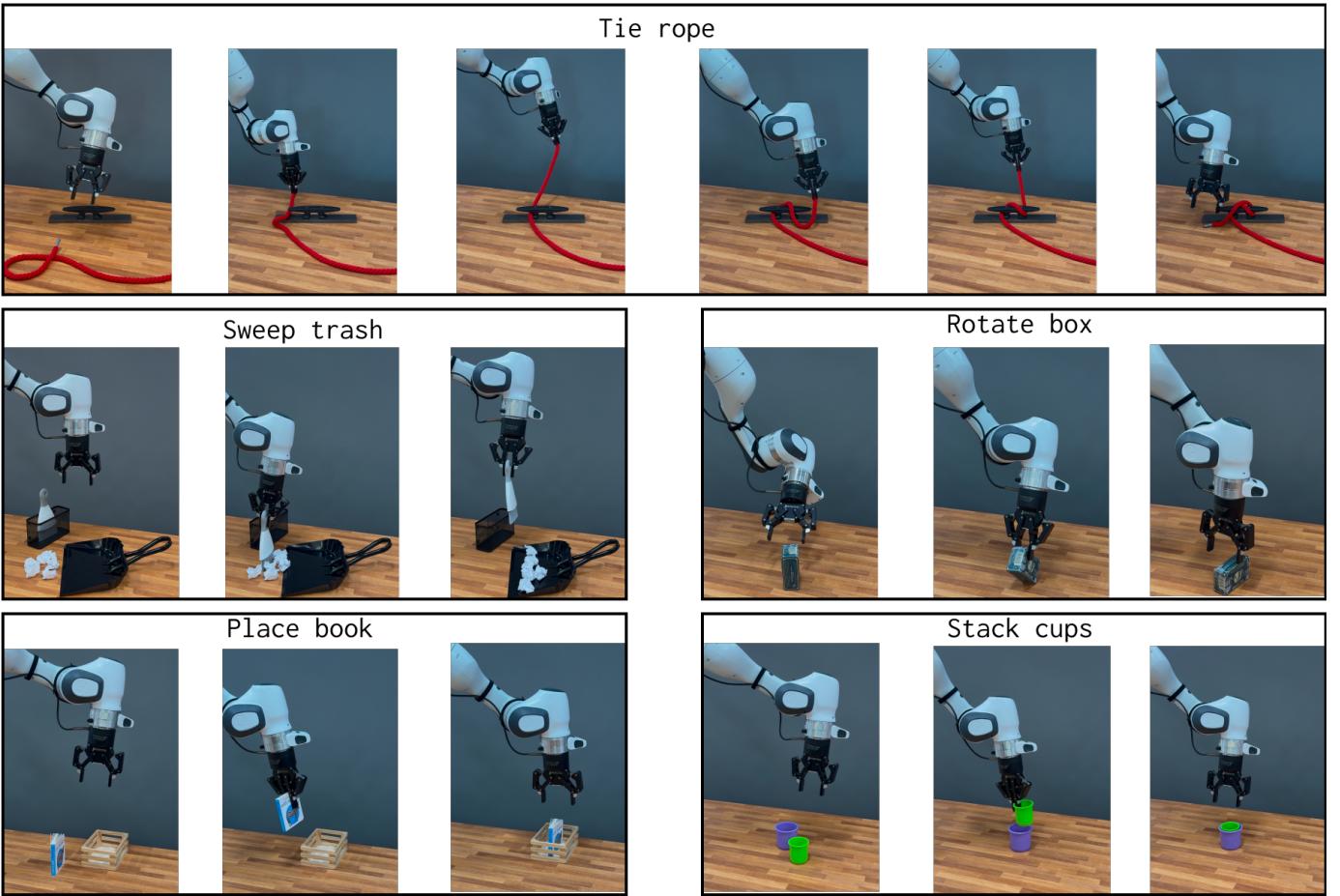


Fig. 5: The five tasks used to evaluate our method in an in-distribution scene on a Franka robot.

	Grasp Brush	Sweep > 0	Sweep > 2	Sweep > 4
Hand Inpaint	0.88	0.80	0.72	0.40
Hand Mask	0.75	0.75	0.72	0.68
Red Line	0.0	0.0	0.0	0.0
Vanilla	0.0	0.0	0.0	0.0

TABLE III: **In distribution scene results — Sweep Trash:** We evaluate success at multiple levels of completion: Grasp Brush measures whether the robot successfully picks up the brush, while Sweep > 0, Sweep > 2, and Sweep > 4 indicate the number of pieces swept into the dustpan. Hand Inpaint and Hand Mask perform comparably, with Hand Mask performing better at the final level. Red Line fails entirely. 25 rollouts per evaluation.

C. Out-of-distribution Scenes

Next, we evaluate how well our method generalizes to new, unseen environments. To do this, we collect human video demonstrations of a sweeping task across diverse scenes (see Fig. 1 for examples). To complete the task, the robot must grasp the sweeper and sweep the green piece of trash off the surface. We collect 950 human video demonstrations across a wide range of indoor and outdoor scenes (see Appendix-A).

We assess generalization in three out-of-distribution (OOD) scenarios (see Fig 6):

- **Outdoor Lawn:** No training data was collected in this environment. However, the white box used as a sweeping surface has appeared in other scenes. Rollouts take place with dynamic background variations, including moving cars and passersby.
- **Indoor Lounge:** A completely new indoor setting with furniture that was not present in any training scenes. The sweeping surface remains the familiar white box.
- **Indoor Lounge + New Surface:** The same indoor lounge as above, but with an unseen blue surface replacing the white box. See Fig. 10 for details on surfaces encountered during training.

Hand Inpaint achieves high success rates across all three OOD environments. Its best performance is in the indoor lounge, which aligns with expectations, as 80% of the training data was collected indoors. When evaluated on an unseen surface, performance drops by 20%, likely due to the limited diversity of training surfaces (four, see Appendix-A.)

Overall, Hand Inpaint and Hand Mask perform comparably both for in-distribution scenes and out-of-distribution scenes. Red Line and the Vanilla baseline were never able to complete any of the tasks due to the significant visual differences between human and robot embodiments. Since, Hand Inpaint

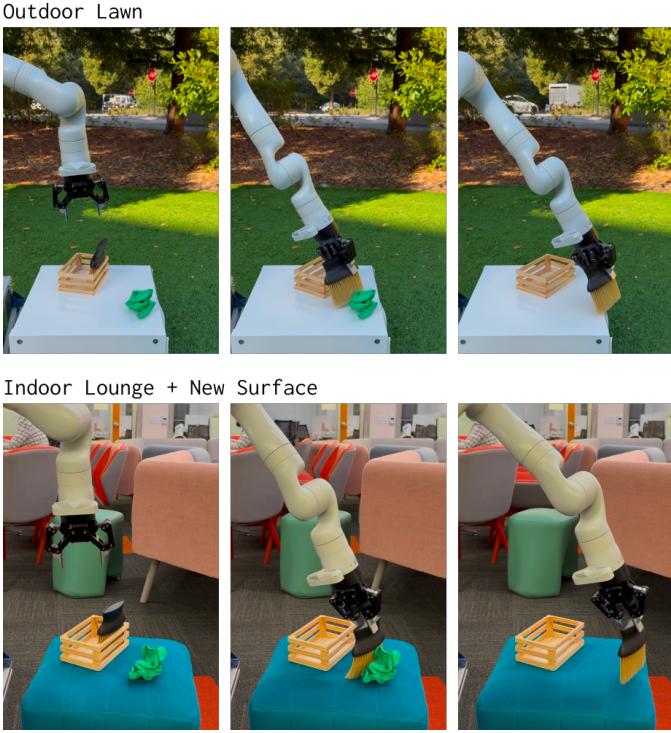


Fig. 6: The out-of-distribution evaluation scenes used to evaluate the sweeping task on the Kinova robot.

	Outdoor lawn	Indoor lounge	Indoor lounge + OOD surface
Hand Inpaint	0.72	0.84	0.64
Hand Mask	0.52	0.76	0.68

TABLE IV: **Out-of-distribution (OOD) scene results:** This table shows the success rates of policies trained on human video demonstrations and tested in three unseen environments: Outdoor Lawn, Indoor Lounge, and Indoor Lounge with an OOD Surface. Hand Inpaint achieves the highest success rates across all settings. Hand Mask performs comparably but is worse in the outdoor lawn setting, perhaps due to weather variations during rollouts. 25 evals per rollout.

is on average 73% faster to rollout than Hand Mask, we use Hand Inpaint for the remaining experiments.

D. Evaluating the Need for High-quality In-painting

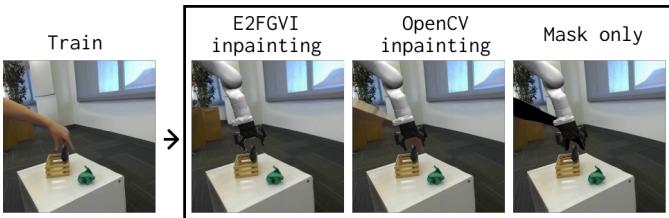


Fig. 7: The three inpainting strategies we compare.

At train time, our preferred Hand Inpaint method relies on inpainting to remove the human arm from the human video and replace it with a realistic background. We test how important

this inpainting is by comparing three variations (see Fig. 7) on the sweeping task in the unseen Indoor Lounge scene:

- High-quality inpainting (E2FGVI [22]): A state-of-the-art video inpainting method.
- Low-quality inpainting (OpenCV inpaint): The OpenCV inpainting function removes the arm but leaves visible artifacts.
- No-inpainting (Mask Only): The human arm is simply masked out during training. Unlike Hand Mask, no hand mask is overlaid at test time.

High-quality inpainting with E2FGVI yields the best performance, achieving a 84% success rate. However, low-quality inpainting performs surprisingly well, with a 76% success rate. This suggests that there is enough variation in the low-quality painting at train time for the model to become agnostic to the artifacts. In contrast, using no-inpainting noticeably degrades performance. The mask-only approach results in a 24 percentage point performance drop. These results suggest that while high-quality inpainting is ideal, our method still performs well with primitive inpainting. Additionally, the high performance of the mask-only approach relative to the 0% success rate of the Red Line method in the easier in-distribution experiments implies that including an overlay of the robot at training time is essential.

Indoor Lounge	
E2FGVI inpaint	0.84
OpenCV inpaint	0.76
Mask only	0.60

TABLE V: **Comparison of in-painting methods:** Using the highest quality inpainting method E2FGVI [22] achieves the highest success rate, but the very primitive OpenCV inpainting function also does remarkably well. Using no inpainting at all leads to a 24 percentage point drop in performance. 25 rollouts per evaluation.

E. Comparing Human vs. Robot Data

While our approach significantly reduces the cost of scaling data collection across diverse environments, it introduces a tradeoff: human video demonstrations provide scalability at the expense of some precision, due to uncertainty in hand pose estimation from RGBD videos. To investigate how much precision is lost, we compare policies trained on teleoperated robot demonstrations (collected using an Oculus controller) against those trained on human video demonstrations for the Kinova sweeping task. All data is collected and evaluated in the same scene.

Our results show that a policy trained on 50 teleoperated demos achieves a 52% success rate, whereas a policy trained on 50 human demos achieves only 44%, indicating a small drop in precision. Despite the lower per-demo precision, increasing the number of human demonstrations to 300 enables the policy to match the success rate of 100 teleoperated demonstrations. This suggests that by leveraging the ease

of collecting large-scale human data, our approach can help overcome the downsides of reduced precision on some tasks.

# of demos	Robot only	Human only
25	0.16	—
50	0.52	0.44
100	0.88	0.64
300	—	0.84

TABLE VI: **Robot demonstrations vs. human video demonstration**
Policies trained on teleoperated data exhibit higher per-demo success rates. However, increasing the number of human demonstrations to 300 allows the policy to match the performance of 100 teleoperated demonstrations. 25 rollouts per evaluation.

F. Evaluating the Benefits of Co-training with Diverse Human Data

While we have already shown in previous experiments that our policy can be deployed zero-shot without any robot data, we also investigate the benefits of co-training with robot data given that there already exists considerable amounts of robot data. To do this, we collect 100 teleoperated demonstrations in a single scene on the Kinova robot using an Oculus controller for the sweeping task. In each observation image from these demonstrations, we overlay a virtual Kinova on the real robot to align the visual distributions between our datasets. Next we co-train our robot data with our larger scale human videos dataset consisting of 950 demonstrations in many different scenes. We see that while the robot-only policy performs well in-distribution, its success rate drops to zero in a new scene. Co-training with human videos from diverse scenes, however, increases the performance to 80%.

	In-distribution Scene	Out-of-distribution Scene
Robot	0.88	0.0
Robot + Human	—	0.80

TABLE VII: **Co-training with diverse human data.** Evaluating the benefits of co-training with diverse human data. 25 rollouts per evaluation.

G. Evaluating the Need for Test Setup Camera Extrinsic during Training.

To generate a robot overlay on a human demonstration video, we need the camera extrinsics of the deployment setup. This constraint makes rapid deployment challenging, as it requires knowing the deployment setup before policy training. We address this limitation using data augmentation, converting each human demonstration into N distinct data-edited videos. Each augmented video features a robot overlay with a randomly positioned base relative to the camera, as shown in Fig. 8.

We evaluate this approach on the Pick and Place Book task, using $N = 5$ during training. In each augmented video, the virtual robot's base is randomly shifted by up to 20 cm along

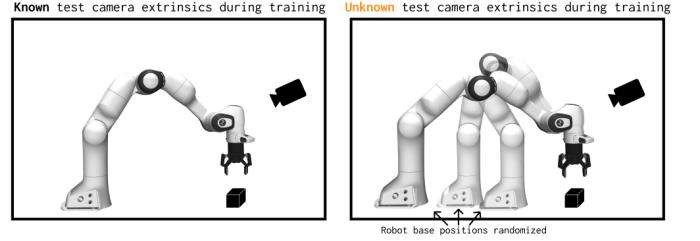


Fig. 8: **Camera extrinsics augmentation.** **Left:** The camera extrinsics of the test setup are known during training and used to generate a virtual robot overlay on the training images. **Right:** The camera extrinsics of the test setup are unknown during training, so random base positions are used to generate the virtual robot overlays.

the x-axis. At test time, we evaluate on a single unseen robot base position.

	Pick/ Place Book
Known test camera extrinsics during training	0.92
Unknown test camera extrinsics during training	0.96

TABLE VIII: **Camera extrinsics augmentation.** Success rates for the Pick and Place Book task when the policy is trained with and without prior knowledge of the test setup's camera extrinsics. 25 rollouts per evaluation.

Our policy trained with randomized camera extrinsics during training achieved a 96% success rate, matching the 92% success rate of a policy trained with known extrinsics. This suggests that data augmentation can mitigate the need for prior knowledge of the test setup's camera extrinsics, enabling more flexible deployment.

V. LIMITATIONS

- The performance of our approach is limited by the performance of existing hand pose estimators since it relies on them to obtain the target actions from a demonstration video. Because hand pose estimators currently still struggle with occlusions, our method does too. However, this also means that our method will get better with time as hand pose estimators improve.
- Our method only works when the robot can follow the same strategy as the human to complete the task. As a result, our policy may lead the robot to collide with the environment even though the human hand does not. Additionally, differences in the surface properties of a human fingertip and robot gripper may lead to different object motions.
- We limit our demonstrations to pinch grasps because our robots are limited to using parallel jaw grippers - a limitation that is shared with virtually all large-scale robot data collection efforts [19, 1, 7].
- We only assess quasi-static tasks, as we do not address the latency mismatch between a human demonstration and a trained policy rolled out on real hardware.

VI. CONCLUSION

We present a method for training robot policies without collecting any robot data, using only human video demonstrations. Our approach successfully transfers policies across a diverse set of tasks, including deformable object manipulation and multiple objects manipulation. Furthermore, we demonstrate zero-shot deployment in novel scenes, showing that human video demonstrations and robot rollouts do not need to occur in the same environment. This flexibility makes our method highly scalable and accessible. By enabling anyone with an RGBD camera to collect meaningful training data anywhere, we lower the barrier to large-scale robot learning and broaden the potential for real-world deployment.

Lastly, a promising direction in robotics is training autoregressive generalist policies on large-scale robotics datasets [1, 20, 7]. Other interesting methods using human video demonstrations introduce complexities that are not amenable to such architectures [43, 36, 3, 47]. In contrast, our simple data-editing approach generates observation-action pairs of robots performing tasks, which can be easily integrated into datasets used to train these generalist policies — a promising avenue for future work.

ACKNOWLEDGMENTS

This work was supported by the NSF through grant number #2327974 as well as Intrinsic. We thank Jimmy Wu for help with hardware, and Claire Chen, Priya Sundaresan, and Juntao Ren for their help throughout the project.

REFERENCES

- [1] Abby O’Neill et al. Open x-embodiment: Robotic learning datasets and rt-x models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903, 2024. doi: 10.1109/ICRA57147.2024.10611477.
- [2] Arpit Bahety, Priyanka Mandikal, Ben Abbatematteo, and Roberto Martín-Martín. Screwmimic: Bimanual imitation from human videos with screw space projection. *arXiv preprint arXiv:2405.03666*, 2024.
- [3] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *arXiv preprint arXiv:2207.09450*, 2022.
- [4] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023.
- [5] Homanga Bharadhwaj, Abhinav Gupta, Vikash Kumar, and Shubham Tulsiani. Towards generalizable zero-shot manipulation via translating human interaction plans. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6904–6911. IEEE, 2024.
- [6] Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. In *European Conference on Computer Vision*, pages 306–324. Springer, 2025.
- [7] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0 : A vision-language-action flow model for general robot control, 2024. URL <https://arxiv.org/abs/2410.24164>.
- [8] Annie S Chen, Suraj Nair, and Chelsea Finn. Learning generalizable robotic reward functions from “in-the-wild” human videos. *arXiv preprint arXiv:2103.16817*, 2021.
- [9] Lawrence Yunliang Chen, Chenfeng Xu, Karthik Dharmarajan, Muhammad Zubair Irshad, Richard Cheng, Kurt Keutzer, Masayoshi Tomizuka, Quan Vuong, and Ken Goldberg. Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning. In *Conference on Robot Learning (CoRL)*, Munich, Germany, 2024.
- [10] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.
- [11] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024.
- [12] Jiafei Duan, Yi Ru Wang, Mohit Shridhar, Dieter Fox, and Ranjay Krishna. Ar2-d2: Training a robot without a robot. *arXiv preprint arXiv:2306.13818*, 2023.
- [13] Jensen Gao, Annie Xie, Ted Xiao, Chelsea Finn, and Dorsa Sadigh. Efficient data collection for robotic manipulation via compositional generalization. *arXiv preprint arXiv:2403.05110*, 2024.
- [14] Nick Heppert, Max Argus, Tim Welschehold, Thomas Brox, and Abhinav Valada. Ditto: Demonstration imitation by trajectory transformation. *arXiv preprint arXiv:2403.15203*, 2024.
- [15] Cheng-Chun Hsu, Bowen Wen, Jie Xu, Yashraj Narang, Xiaolong Wang, Yuke Zhu, Joydeep Biswas, and Stan Birchfield. Spot: Se (3) pose trajectory diffusion for object-centric manipulation. *arXiv preprint arXiv:2411.00965*, 2024.
- [16] Vidhi Jain, Maria Attarian, Nikhil J Joshi, Ayzaan Wahid, Danny Driess, Quan Vuong, Pannag R Sanketi, Pierre Sermanet, Stefan Welker, Christine Chan, et al. Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers. *arXiv preprint arXiv:2403.12943*, 2024.
- [17] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic

- imitation learning. In *Conference on Robot Learning*, pages 991–1002. PMLR, 2022.
- [18] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. *arXiv preprint arXiv:2410.24221*, 2024.
- [19] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srivama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.
- [20] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.
- [21] Marion Lepert, Ria Doshi, and Jeannette Bohg. Shadow: Leveraging segmentation masks for cross-embodiment policy transfer. In *8th Annual Conference on Robot Learning*.
- [22] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [23] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.
- [24] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.
- [25] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. *arXiv preprint arXiv:2108.03298*, 2021.
- [26] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.
- [27] Georgios Papagiannis, Norman Di Palo, Pietro Vitiello, and Edward Johns. R+ x: Retrieval and execution from everyday human videos. *arXiv preprint arXiv:2407.12957*, 2024.
- [28] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024.
- [29] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*, pages 416–426. PMLR, 2023.
- [30] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädl, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL <https://arxiv.org/abs/2408.00714>.
- [31] Juntao Ren, Priya Sundaresan, Dorsa Sadigh, Sanjiban Choudhury, and Jeannette Bohg. Motion tracks: A unified representation for human-robot transfer in few-shot imitation learning. *arXiv preprint arXiv:2501.06994*, 2025.
- [32] Nur Muhammad Mahi Shafiullah, Anant Rai, Haritheja Etukuru, Yiqian Liu, Ishan Misra, Soumith Chintala, and Lerrel Pinto. On bringing robots home. *arXiv preprint arXiv:2311.16098*, 2023.
- [33] Lin Shao, Toki Migimatsu, Qiang Zhang, Karen Yang, and Jeannette Bohg. Concept2robot: Learning manipulation concepts from instructions and human demonstrations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2020.
- [34] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.
- [35] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930, June 2023.
- [36] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. *arXiv preprint arXiv:2302.12422*, 2023.
- [37] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. *arXiv preprint arXiv:2401.00025*, 2023.
- [38] Jimmy Wu, William Chong, Robert Holmberg, Aaditya Prasad, Yihuai Gao, Oussama Khatib, Shuran Song, Szymon Rusinkiewicz, and Jeannette Bohg. Tidybot++: An open-source holonomic mobile manipulator for robot learning. In *Conference on Robot Learning*, 2024.
- [39] Philipp Wu, Yide Shentu, Zhongke Yi, Xingyu Lin, and Pieter Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12156–12163. IEEE, 2024.
- [40] Annie Xie, Lisa Lee, Ted Xiao, and Chelsea Finn.

- Decomposing the generalization gap in imitation learning for visual robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3153–3160. IEEE, 2024.
- [41] Haoyu Xiong, Quanzhou Li, Yun-Chun Chen, Homanga Bharadhwaj, Samarth Sinha, and Animesh Garg. Learning by watching: Physical imitation of manipulation skills from human videos. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7827–7834. IEEE, 2021.
 - [42] Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela Veloso, and Shuran Song. Xskill: Cross embodiment skill discovery. In *Conference on Robot Learning*, pages 3536–3555. PMLR, 2023.
 - [43] Mengda Xu, Zhenjia Xu, Yinghao Xu, Cheng Chi, Gordon Wetzstein, Manuela Veloso, and Shuran Song. Flow as the cross-domain manipulation interface. *arXiv preprint arXiv:2407.15208*, 2024.
 - [44] Kevin Zakka, Yuval Tassa, and MuJoCo Menagerie Contributors. MuJoCo Menagerie: A collection of high-quality simulation models for MuJoCo, 2022. URL http://github.com/google-deepmind/mujoco_menagerie.
 - [45] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.
 - [46] Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors. In *Conference on Robot Learning*, pages 1199–1210. PMLR, 2023.
 - [47] Yifeng Zhu, Arisrei Lim, Peter Stone, and Yuke Zhu. Vision-based manipulation from single human video with open-world object graphs. *arXiv preprint arXiv:2405.20321*, 2024.

APPENDIX

A. Data collection

The details of the human video demonstration datasets collected for all tasks are shown in Table IX.

	Task	Number of demos	Max Horizon
Panda tasks	Zebra	313	228
	Stack	268	148
	Sweep	279	407
	Cleat	307	516
Kinova task	Soap	283	175
	Sweep	950	168

TABLE IX: Human video demonstration datasets details. **Number of demos**: number of human video demonstrations used in training, **Max Horizon**: the maximum number of steps in the human video demonstration dataset.



Fig. 10: Left: surfaces used during human video data collection for the kinova sweep task. Right: Unseen surface used to evaluate policy.

B. Hand Mask Data Editing Method

We describe in detail how we adapt the data-editing strategy from Shadow [21] to the human-to-robot setting.

1) *Data-editing at Train Time*: We segment out the pixels corresponding to the hand and set them to black. We extract the hand pose using the strategy described in Section III-B and overlay an RGB rendering of the target robot in this pose using the known camera extrinsics.

2) *Data-editing at Inference Time*: To ensure that the train and test time images match closely, we overlay a black segmentation mask of the human arm and hand in the same pose as the robot. However, unlike in Shadow where the authors transferred policies between two robots, we do not have access to a realistic virtual model of a human arm and hand. Instead we train a diffusion model to predict the segmentation mask of the hand given a 6-DOF pose. This diffusion model is trained from scratch using the hand masks and corresponding target poses from our training dataset. At test time, we overlay the segmentation mask predicted by the diffusion model onto our image, and feed this edited image into our policy.

3) *Training the Hand Mask Diffusion Model*: We train the diffusion model for generating hand masks using the hyperparameters in Table X. Due to high compute requirements, we generate 64×64 images and upscale them to the desired resolution using a super-resolution model [23]. To improve temporal consistency, the Hand Mask Diffusion Model generates hand masks for both time t and $t + 1$ given a robot pose at time t . Additionally, we apply attention injection [35] at test time, using the model’s output at time t as the reference image for time $t + 1$.

Forward Diffusion Timesteps	Sampling Steps	ImgRes	Batch	Lr
1000	50	64	32	1e-4

TABLE X: Hyperparameters used to train the Hand Mask diffusion model. **Forward Diffusion Timesteps**: the number of forward process steps at train time. **Sampling Steps**: the number of sampling steps used during inference.

4) *Data Augmentation*: Because the angle of the human arm can vary for the same 6-DOF pose, we need to make our policy agnostic to this variation. In addition, we find that the diffusion model we use to render the human segmentation mask does not output temporally consistent angles of the arm across frames. To address this problem, we augment our training data to include a second segmentation mask of the arm that is randomly shifted by a few pixels at each timestep. Importantly, the segmentation mask of the robot is not shifted, ensuring that the model can rely on the segmentation mask of the robot to localize the embodiment without relying on the hand mask.

We evaluate the impact of this augmentation on the Pick and Place Book and Stack Cups tasks. As shown in Table XI, incorporating the augmented hand mask achieves comparable performance in the easier Pick and Place Book task, but increases the success rate for the harder Stack Cups task by 12 percentage points. This suggests that the augmentation helps mitigate inconsistencies in the human arm mask generation, leading to more robust policy learning.

	Pick/ Place Book	Stack Cups
Hand Mask	0.92	0.52
Hand Mask (no data aug)	0.88	0.40

TABLE XI: **Effect of data augmentation on the Hand Mask strategy**. Adding random shifts to the hand mask during training improves performance in Stack Cups, where success increases from 40% to 52%. This suggests that the augmentation helps the policy generalize despite inconsistencies in the human arm segmentation. 25 rollouts per evaluation.

C. Policy Training Details

Table XII describes the hyperparameters used to train diffusion policy. All experiments done on the Kinova robot use the same set of hyperparameters. This includes the diverse scene experiments, inpainting quality experiments, and robot vs.

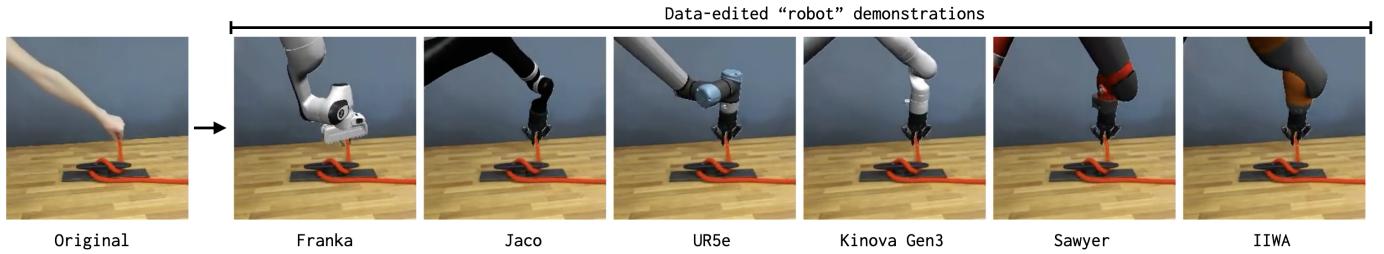


Fig. 9: Our method is robot agnostic. Each human video can be converted into a robot demonstration for any robot capable of completing the task.

human video data comparison experiments. We used the same set of hyperparameters for all data-editing methods across each task. All tasks in the paper use the DDIM scheduler with 100 training steps and 10 inference steps.

Robot	Task	To	Ta	ImgRes	Batch	Lr	Aug
Panda	Zebra	2	8	240	200	1e-4	Yes
	Stack	2	8	240	200	1e-4	Yes
	Sweep	2	8	240	200	1e-4	Yes
	Cleat	2	8	240	200	1e-4	Yes*
	Soap	2	8	240	200	1e-4	Yes
Kinova	Sweep	2	8	240	256	1e-4	Yes

TABLE XII: Hyperparameters for Diffusion policy training. **To:** observation horizon, **Ta:** action horizon. **Aug:** Image augmentations (RandomCrop, RandomRotation, ColorJitter) used during training. Only ColorJitter was used for the Cleat task to avoid cropping out grasps of the rope that frequently occur on the edge of the image.

D. Detailed Task Descriptions

The variation in the placement of objects for each task is visualized in Fig. 11.

- **Pick and Place Book:** The robot must pick up a book and place it inside a wooden container. The book’s initial position is randomly sampled within the outlined 30 cm x 35 cm rectangular region. Additionally, its orientation can vary by ± 45 degrees.
- **Rotate Box:** The robot must rotate a box 90 degrees onto a new face in a controlled fashion (simply knocking it over is not valid). The box’s initial position is randomly sampled within the outlined 40 cm x 35 cm rectangular region.
- **Stack Cups:** The robot must stack the green cup inside the purple cup. Precise alignment is critical, as the cups differ in diameter by only 1.5 cm. The cups’ initial positions are randomly sampled within the outlined 40 cm x 35 cm rectangular region.
- **Tie Rope:** The robot must tie a simplified cleat hitch, a sailing knot that follows a figure-eight ∞ pattern. This task is challenging due to the precise manipulation required of a highly deformable object. The position of the cleat is randomly sampled within the outlined 30cm x 45cm white rectangular region and rotated by ± 30 degrees.

- **Franka Sweep Trash:** The robot must pick up the sweeper and sweep six pieces of trash into a dustpan. This task involves coordinated multi-object manipulation, requiring the robot to control the sweeper while simultaneously managing the movement of multiple loose objects. Additionally, the pieces of trash exhibit unpredictable dynamics, necessitating continuous adaptation based on real-time feedback. The position of the six pieces of trash is randomly sampled within the 25cm x 35cm rectangular region outlined in white. The position of the sweeper is randomly varied by 5cm x 4cm laterally and rotated by ± 5 degrees. The position of the dustpan is randomly varied by 15cm x 15cm laterally and rotated by ± 20 degrees.
- **Kinova Sweep Trash:** The robot must pick up the sweeper and sweep the green piece of trash off the table. The position of the sweeper is randomly sampled within the 33cm x 35cm rectangular region outlined in yellow. The position of the green piece of trash is randomly sampled within the 15cm x 33cm region outlined in orange.

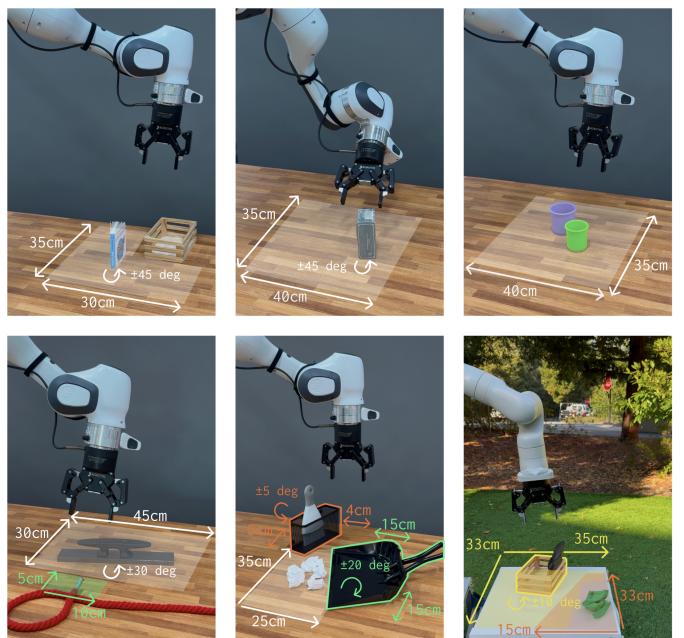


Fig. 11: Variation in object placement during evaluations of each task.