

Imitation Learning with Precisely Labeled Human Demonstrations

Yilong Song

Abstract—Within the imitation learning paradigm, training generalist robots requires large-scale datasets obtainable only through diverse curation. Due to the relative ease to collect, human demonstrations constitute a valuable addition when incorporated appropriately. However, existing methods utilizing human demonstrations face challenges in inferring precise actions, ameliorating embodiment gaps, and fusing with frontier generalist robot training pipelines. In this work, building on prior studies that demonstrate the viability of using hand-held grippers for efficient data collection, we leverage the user’s control over the gripper’s appearance—specifically by assigning it a unique, easily segmentable color—to enable simple and reliable application of the RANSAC and ICP registration method for precise end-effector pose estimation. We show in simulation that precisely labeled human demonstrations on their own allow policies to reach on average 88.1% of the performance of using robot demonstrations, and boost policy performance when combined with robot demonstrations, despite the inherent embodiment gap.

I. INTRODUCTION

Eyeing the success of large language models (LLMs), we are increasingly interested in exploring end-to-end vision-language-action (VLA) models as a path to generalist robots. If the LLM’s success can be attributed to two factors: expressive architectures and diverse high quality datasets, the VLA’s current obstacle lies in obtaining the latter. Robot demonstrations, the most straightforward and proven form of training data, are expensive to collect; Brohan et al. [2022], for instance, used 17 months to collect 130K trajectories, but still fell short in matching the data quantity LLMs use to achieve generality. Methods for collecting and using alternative categories of data are thus explored out of necessity.

Among these alternative categories, human demonstrations are of special importance due to its direct relevance to robot tasks. One category of prior work that seeks to leverage internet-scale human videos pretrains with a self-supervised objective a model that learns a discrete latent action representation from unstructured videos, then maps latent actions to true actions with small amounts of labeled video [Schmidt and Jiang, 2024, Ye et al., 2024]. This class of methods, which we call latent-action methods, is advantageous in its easy setup (involving no hardware usage), wide applicability (to almost any video), scalability (which is the result of the former two properties), and effectiveness in learning representations of motions [Bjorck et al., 2025], but as the tradeoff struggles with higher dimensional continuous action spaces and fine-grained motion planning, and also suffers from shortages of actual large-scale high-quality training data despite inherent scalability [Ye et al., 2024]. Another category of work finds a middle ground between teleoperated robot data and latent-action data in terms of scalability, control over video content, and label precision by proposing data collection methods that circumvent robot usage [Song et al., 2020, Young et al., 2020,

Shafiullah et al., 2023, Chi et al., 2024a, Lepert et al., 2025]. Our work falls within this category.

All methods utilizing human demonstrations need to address two issues: the lack of action labels, and the presence of an embodiment gap. Our work relies on the observation that, when granting two assumptions: first the appearance of the end-effector stays constant between training and testing (i.e. the same gripper is used hand-held and mounted on the robot), and second an embodiment invariant wrist camera view is provided, a visuomotor policy model trained on *precisely labeled* human demonstrations generalizes zero-shot to the robot embodiment with respectable performance. On this basis, we

- 1) devise a method for reliable, precise action labeling, and
- 2) conduct experiments in simulation quantifying policy performance when trained both solely on precisely labeled human demonstrations and a mixture of human demonstrations and a small number of robot demonstrations,

which can be considered our two contributions. Specifically, balancing cost of implementation and the requirement for precision, we leverage our control over the gripper’s appearance by assigning it a unique, easily segmentable color to facilitate robust application of Random Sample Consensus (RANSAC) and Iterative Closest Point (ICP) registration for precise labeling via pose estimation. We show that policies trained on human demonstrations reach on average 88.1% of the performance of the same architecture trained on the same number of robot demonstrations, while the addition of precisely labeled human demonstrations to a training set of robot demonstrations significantly boosts policy performance.

II. RELATED WORK

A. Learning from Human Demonstrations

As previously mentioned, the purpose of this category of work is to expand the robotics training set, and the two inherent obstacles every work in this category must in some way address are (1) the lack of action labels, and (2) the presence of an embodiment gap. It is interesting to observe how the two issues are addressed by different existing approaches.

One class of approaches, which we call latent action methods, uses a self-supervised objective to pretrain a model that learns a latent action representation from videos [Schmidt and Jiang, 2024, Ye et al., 2024]. This approach addresses problem (1) by supplying motion-encoded pseudo-labels to human videos, and relies on separate mechanisms to establish correspondences between latent actions and real actions [Ye et al., 2024], and similar motions on different embodiments [Bjorck et al., 2025], thus does not address problem (2) directly. This approach is both highly scalable and effective in learning

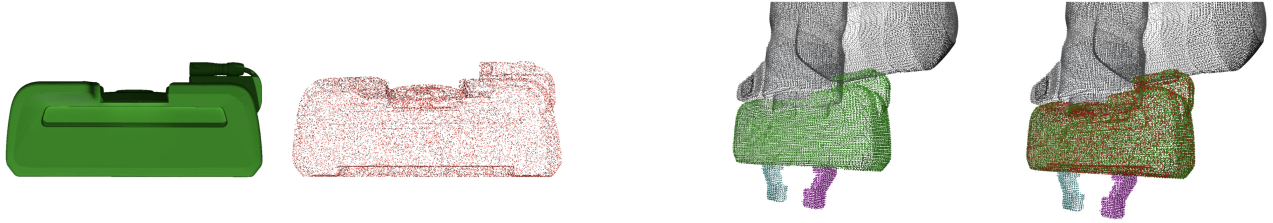


Fig. 1: Going from left to right: 1) The triangle mesh of the recolored panda gripper used in simulation; 2) the corresponding point cloud sampled uniformly from the mesh, which we use for pose estimation; 3) a point cloud of the robot (cropped from the scene for clarity) in a simulated demonstration; 4) Random Sample Consensus (RANSAC) and Iterative Closest Point (ICP) are applied to estimate a rigid body transformation that aligns the end-effector point cloud obtained from the mesh and the robot point cloud (i.e. the pose), resulting in precise and reliable end-effector pose estimation without the need to train a deep learning model.

embodiment-insensitive motion representations [Bjorck et al., 2025], but as the tradeoff does not provide precise guidance for learning fine-grained motion planning. Although also involving representation learning, self-supervised objectives, and internet-scale videos, vision encoders tailored for robotics are pretrained to capture visual semantic priors and sometimes state representations instead of just action representations [Radosavovic et al., 2022, Nair et al., 2022, Ma et al., 2023]. A related and also in principle highly scalable class of methods is reward-learning methods, which involves training reward functions that measure the progress or success of a trajectory, sometimes with respect to a language instruction [Shao et al., 2020, Chen et al., 2021, Ma et al., 2023]. All three categories share the properties of not explicitly providing action labels or aligning embodiments, but rather assist other modules of the system in these respects implicitly. Our work is similar to them in that we also do not explicitly align embodiments—although we do focus on providing fine-grained, precise, true action labels—and can be easily integrated with large scale imitation learning pipelines.

Less scalable methods that use structured, often locally-collected human manipulation videos, on the other hand, tend to attempt addressing both issues simultaneously.

Flow-based methods learn to predict the flows of points across the frame sequence, then use the predicted flow to guide a downstream policy [Xiong et al., 2021, Xu et al., 2024, Wen et al., 2023, Ren et al., 2025, Georgios Papagiannis and Johns, 2025]. Since flows both describe movements and abstract away the appearance of agents and objects, they address problem (1) and (2) simultaneously—but not completely, since as long as points on the agent is tracked, a morphology gap remains as humans and robots move in different ways. Compared to flow-based methods, our method contains fewer modules, no off-the-shelf components, can capture 3D motion more precisely, and can integrate with large-scale imitation learning frameworks easier.

Object-centric methods [Hsu et al., 2024, Heppert et al., 2024, Bahety et al., 2024, Zhu et al., 2024] circumvent problem (2) and provide at least post-grasp solution to problem

(1) by tracking objects instead of embodiments, but compared to our method are more constrained in terms of the range of tasks and types of objects they can effectively handle, and cannot easily integrate with large-scale imitation learning systems.

Another prominent class of methods, which we call behavior-prior learning methods, learns motion or planning priors, then uses these priors to improve policy performance and generalization [Bahl et al., 2022, Wang et al., 2023, Xu et al., 2023, Bharadhwaj et al., 2024, Jain et al., 2024]. These methods provide either implicit (latent) guidance [Wang et al., 2023, Xu et al., 2023, Bharadhwaj et al., 2024, Jain et al., 2024] or sparse labels via waypoints [Bahl et al., 2022], while our work focuses on providing precise labels at reasonable cost. Methods in this category address the embodiment gap problem in various ways: Wang et al. [2023] do so by decoupling high-level latent planning (learned from human play data) from low-level control (trained on robot demonstrations); Jain et al. [2024] by leveraging paired human-robot trajectory data; Xu et al. [2023] by learning a shared cross-embodiment skill representation through self-supervised clustering. These methods require either extra training [Wang et al., 2023, Xu et al., 2023] or extra idiosyncratic data collection [Jain et al., 2024]; in contrast, our work closes the embodiment gap simply with a common end-effector and does not require the end-to-end policy learning pipeline to be modified.

Lastly, video prediction methods train or fine-tune video generators that predict subsequent frames conditioned on the first frame, and use the predicted frame sequence at inference time as temporally dense motion guidance for the robot [Du et al., 2023, Ko et al., 2023, Liang et al., 2024]. If the human at training time and robot at test time use end-effectors with shared appearances, which is the key insight leveraged by Liang et al. [2024], the action trajectory can be precisely extracted via end-effector pose estimation, bypassing the embodiment difference, solving both problem (1) and (2) simultaneously. Video prediction methods have many advantages, but also three flaws. First is the computational cost of video generation models, which hinders closed-loop control.

Second is that the video generation models hallucinate. Third is the prohibitively large extra expense required to enable end-effector pose estimation with precision matching that of our method. Our work uses 3D scene reconstruction to maximize pose estimation precision, while 3D video generation or multi-camera 2D video generation are very computationally involved. Our work is inspired by the use of shared end-effectors in Liang et al. [2024], but involves no expensive video prediction.

B. Human Demonstration Collection Methods

Human demonstrations are labeled or unlabeled videos of human agents performing robotics tasks without robots. Compared to robot demonstrations, human demonstrations avoid two expensive factors: robot hardware and teleoperation. Therefore, although data still scales linearly with human labor, the cost of collecting human demonstrations is significantly lower than collecting robot demonstrations. Lepert et al. [2025] propose collecting videos of humans completing manipulation tasks with their hands, track the hands’ poses, then overlay virtual renderings of the target embodiment accordingly, thus addressing both the label and the embodiment problems. Our method in contrast uses a hand-held end-effector, which trades off the convenience of using less hardware for simple and more precise pose estimation. Different from Lepert et al. [2025], Song et al. [2020], Young et al. [2020], Shafiullah et al. [2023], Chi et al. [2024a] proposed the more portable solution of using hand-held hardware with ego-centric camera angles, which is advantageous for in-the-wild data collection. On the basis of prior work, Chi et al. [2024a] expanded the set of operable tasks by ameliorating occlusion with fisheye lenses and side mirrors, and achieved sub-centimeter action precision with a visual-inertial SLAM pipeline incorporating video and IMU data, which slightly increases cost and decreases portability [Chi et al., 2024a]. Chi et al. [2024a] constitute a thoroughly reasoned and well-tested real-world data collection solution, while our work, with its smaller scope, loosens the “no external camera” constraint and focus on devising and testing a minimalist precise labeling strategy in simulation.

III. METHOD

We follow a simple paradigm: given unlabeled human demonstrations, we label the human demonstrations via end-effector pose estimation, and use the labeled human demonstrations—either on their own or in concatenation with robot demonstrations—to train a visuomotor policy. To retain portability and cost-effectiveness, we devise a method that requires only accessible hardware, namely a 3D printed hand-held gripper and one or more RGB-D cameras, while maintaining precision.

A. Point Cloud Reconstruction

To most accurately capture 3D geometries in the scene, particularly that of the end effector, we deploy external depth cameras with known matrices and perform point-cloud reconstruction to obtain 3D scenes corresponding to each frame. The reconstruction method is standard.

We start with an unlabeled dataset $\mathcal{D} = \{\text{demo}_i\}_{i=1}^N$ of human demonstrations. With T_i denoting the number of timesteps, each demonstration $\text{demo}_i = (I_{i,1}, I_{i,2}, \dots, I_{i,T_i})$, where observation $I_{i,t} \in \mathbb{R}^{V \times 4 \times H \times W}$ represents V RGBD images from V camera angles with height H and width W , is an ordered sequence of observations. For each camera $c \in \{1, \dots, V\}$ in each observation $I_{i,t}$, denoted $I_{c,i,t}$, every pixel with coordinates (u, v) in $I_{c,i,t}$ with depth value $d_{c,i,t}(u, v)$ corresponds to a homogeneous 3D point $\mathbf{X}_{c,i,t} = [x, y, z, 1]^T$ in the global frame via:

$$\mathbf{X}_{c,i,t} = \mathbf{T}_c h \left(d_{c,i,t}(u, v) \mathbf{K}_c^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \right),$$

where \mathbf{K}_c and \mathbf{T}_c are respectively the intrinsic and extrinsic matrix of camera c , and h is the homogenization operator (i.e. $h(\mathbf{X}) = [\mathbf{X}^T, 1]^T$). We apply this to each view at each time step to obtain a series of partial point clouds, and merge partial point clouds to obtain fuller point clouds.

B. End-Effector Pose Estimation

Inspired by Liang et al. [2024], Chi et al. [2024a], we use a custom end-effector that work both hand-held and mounted on a robot. The design process naturally produces CAD models of the used end-effectors, which, along with the aforementioned point clouds, allows us to apply a simple and reliable model-based pose estimation algorithm that involves no training and minimal tuning.

Concretely, given the point clouds and a CAD model of the end-effector, we first segment out points belonging to the end-effector, then align the CAD model of the end-effector onto the points by estimating the best rigid-body transformation, i.e., the pose. To enable easy segmentation, we simply customize the end-effector to have a distinct color, e.g. green, and segmentation can be performed reliably by simply filtering out points with the color, and taking the largest cluster. To estimate the best rigid-body transformation, we first obtain a coarse alignment using the Random Sample Consensus (RANSAC) algorithm, and then refine this estimate with the Iterative Closest Point (ICP) algorithm. Since videos are temporally dense and adjacent frames capture fine motions, this “RANSAC plus ICP” combination needs only be applied for the first frame, and the estimated pose of the previous frame can be used as the initial coarse alignment for each subsequent frame to both increase efficiency and easily enforce rotational consistency for symmetric end-effectors. See Figure 1 for a visualization of this process.

This algorithm labels each frame with the corresponding absolute end-effector pose. To obtain an action-labeled dataset, we simply shift the series of poses forward by one time step so that the estimated pose at frame $t + 1$ becomes the goal pose at frame t .

IV. EXPERIMENTS

We evaluate our approach in robosuite simulated environments, with diffusion policy [Chi et al., 2024b] as the downstream behavioral cloning architecture of choice. Specifically,

TABLE I: **Downstream Policy Performance.** We report the success rates of the visual diffusion policy on different data mixtures, averaged across 50 different environment initial conditions. Each cell contains (best single test success rate)/(average success rate of best consecutive 10 checkpoints)/(starting epoch of the best consecutive 10 checkpoints). The best results in each column are in bold; the second best is underlined.

	Square D_0	Coffee D_1	Stack D_0	Threading D_0	Three Piece Assembly D_0
50TD + 0HD	0.70/0.63/0650	0.20/0.12/0050	0.74/0.48/0050	0.66/0.57/0200	0.38/0.30/0050
200TD + 0HD	0.98/0.94/2550	<u>0.76/0.63/0050</u>	1.00/0.98/0050	0.96/0.89/0650	<u>0.78/0.71/1400</u>
0TD + 200HD	0.86/0.82/1250	0.70/ <u>0.65</u> /0100	0.76/0.69/2550	0.02/0.00/0050	0.74/0.65/0150
50TD + 50HD	0.74/0.69/1250	0.46/0.26/0050	0.72/0.61/0050	0.84/0.70/0200	0.58/0.48/0050
50TD + 100HD	0.80/0.71/0100	0.54/0.41/0100	0.84/0.79/0050	0.86/0.79/0650	0.68/0.59/0600
50TD + 200HD	0.88/0.83/0050	0.68/0.63/0050	<u>0.96/0.91/0750</u>	0.94/0.85/0550	0.74/0.68/0550
50TD + 400HD	<u>0.94/0.88/0350</u>	0.88/0.81/0100	0.94/0.88/1350	0.96/0.93/1000	0.82/0.74/0150

we choose MimicGen [Mandlekar et al., 2023] tasks due to the larger datasets MimicGen provides (1000 demonstrations each task that can be downloaded directly and the flexibility to generate more demonstrations across multiple embodiments). MimicGen datasets are each generated from 10 source human teleoperated demonstrations by segmenting object-centric sub-tasks and leveraging trajectory equivariance in relations to new object poses, thus are different from fully human collected datasets. But since Mandlekar et al. [2023] demonstrated for the exact tasks experimented in this work that policy performance on MimicGen data “can be comparable to performance on an equal amount of human demonstrations,” and that increasing the number of MimicGen generated demonstrations in the training set has a diminishing return on performance, we conclude that the use of MimicGen datasets does not significantly impact trends observed in our results, at least not to our advantage, and that the usage is reasonable. MimicGen datasets are arranged into categories D_0, D_1, D_2 indicating different reset distribution variabilities in the object poses, with D_0 being the baseline reset distribution, and D_1 and D_2 having increasing variability. We choose to use the standard D_0 reset distribution for all tasks except Coffee, because for Coffee D_0 the diffusion policies reach a test success rate of 1 across all data mixtures quickly, and does not reveal useful information regarding our objective.

To simulate the embodiment gap, we use the KUKA iiwa manipulator to simulate human demonstrations (i.e., demonstrations carried out by a different embodiment) by relabeling the dataset with our pose estimation pipeline, and the Franka Panda for robot demonstrations and the embodiment at test time. As previously mentioned, we use the same colorized gripper, i.e., the Panda gripper on both embodiments. We train diffusion policies implemented by Chi et al. [2024b] out-of-the-box (U-Net with visual observations) for 3000 epochs, which we empirically observe to be sufficient to allow for convergence, following the input arrangement of the diffusion policy out-of-the-box, which include one external view (agentview, with dimension 84×84), one wrist camera view (84×84), and the end-effector pose, and report the results of training on different mixtures of precisely labeled demonstrations with an embodiment gap and demonstrations

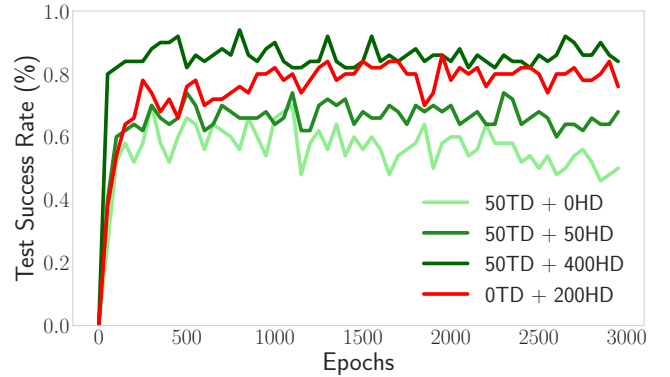


Fig. 2: **Downstream Policy Performance on Task Square D_0 During Training on Different Data Mixtures.** We visualize the task success rate of a visual diffusion policy averaged across 50 different environment initial conditions on different mixtures of teleoperated demonstrations (TD) and simulated human demonstrations (HD) during training.

with original labels without an embodiment gap. The results are summarized in Table I, which we discuss in the next paragraph, while Figure 2 visualizes the training curves for different data mixtures on the Square D_0 task. In Table I we choose to display the average of the best consecutive 10 checkpoints instead of the last 10 because the models reach convergence and overfit at different rates when trained on different data mixtures.

Quantitatively, results in Table I indicate that the success rate of the mixture 0TD + 200HD on average reach 88.1% of the success rate of the mixture 200TD + 0HD in terms of maximum test score, and 87.7% in terms of the average of 10 best checkpoints, excluding the extreme outlier task Threading. We hypothesize that the results observed for Threading are due to the fact that the point of insertion, which is vital for the success of Threading, lies outside the view of the wrist camera when the insertion is initiated. The wrist camera is embodiment invariant, but the policy has to rely on the agentview camera for Threading. On the other hand, with 50 robot/same-embodiment demonstrations added,

the policy’s performance on Threading does scale as we add more human/cross-embodiment demonstrations. This indicates 1) the presence of embodiment invariant views is important for zero-shot cross-embodiment generalization, and therefore the amelioration of occlusion in these views is vital for wide-range task applicability (as explored by Chi et al. [2024a]) and 2) when same-embodiment demonstrations, even a small number, are provided, the addition of precisely labeled cross-embodiment demonstrations increases policy performance despite occlusion in the embodiment-invariant view. All other tasks exhibit the same increased success rate as we increase the number of added cross-embodiment demonstrations on top of 50 same-embodiment demonstrations. On average across all tasks, policy performance of the best 50 TD data mixture on average reach 102.6% of the success rate of the mixture 200TD + 0HD in terms of maximum test success rate, and 104.1% in terms of the average of 10 best checkpoints. Moreover, from the starting epoch of the best consecutive 10 checkpoints we do not observe a definitive trend of slower convergence as the size of the datasets increases, at least for the dataset scale we have tested.

V. CONCLUSIONS, LIMITATIONS, FUTURE WORK

In this work we propose a simple and reliable method for extending policy learning data with precisely labeled human demonstrations, and demonstrate its effectiveness in simulated environments. Our results are limited on two fronts: first, the lack of real-world experiments, and second, the lack of an ablation study on larger scale datasets containing more demonstrations or categories of visuomotor data used for training frontier VLA models. The first limitation, caused by our lack of access to equipment, is admittedly nontrivial since our method relies to an extent on precise point cloud reconstruction, which may be subject to harsher hardware limitations in the real world. The second limitation is more appropriately addressed in a larger scale project seeking to propose data strategies for frontier VLA models. We leave both for future work to ameliorate. Otherwise, a topic of future work we suggest is the automated generation of precisely labeled human demonstrations with generative models.

REFERENCES

- Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale. In *arXiv preprint arXiv:2212.06817*, 2022.
- Dominik Schmidt and Minqi Jiang. Learning to act without actions. 01 2024.
- Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejeun Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos, 2024. URL <https://arxiv.org/abs/2410.11758>.
- Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llonet, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. Gr00t n1: An open foundation model for generalist humanoid robots, 2025.
- Shuran Song, Andy Zeng, Johnny Lee, and Thomas Funkhouser. Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations. *Robotics and Automation Letters*, 2020.
- Sarah Young, Dhiraj Gandhi, Shubham Tulsiani, Abhinav Gupta, Pieter Abbeel, and Lerrel Pinto. Visual imitation made easy, 2020.
- 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.
- 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. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024a.
- Marion Lepert, Jiaying Fang, and Jeannette Bohg. Phantom: Training robots without robots using only human videos, 2025. URL <https://arxiv.org/abs/2503.00779>.
- Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. *CoRL*, 2022.
- Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation, 2022.
- 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. *ICLR*, 2023.
- 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.

- Annie Chen, Suraj Nair, and Chelsea Finn. Learning generalizable robotic reward functions from “in-the-wild” human videos. In *Robotics: Science and Systems XVII*, RSS2021. Robotics: Science and Systems Foundation, July 2021. doi: 10.15607/rss.2021.xvii.012. URL <http://dx.doi.org/10.15607/RSS.2021.XVII.012>.
- 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, 2021. doi: 10.1109/IROS51168.2021.9636080.
- Mengda Xu, Zhenjia Xu, Yinghao Xu, Cheng Chi, Gordon Wetzstein, Manuela Veloso, and Shuran Song. Flow as the cross-domain manipulation interface, 2024.
- Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning, 2023.
- Juntao Ren, Priya Sundareshan, Dorsa Sadigh, Sanjiban Choudhury, and Jeannette Bohg. Motion tracks: A unified representation for human-robot transfer in few-shot imitation learning, 2025.
- Pietro Vitiello Georgios Papagiannis, Norman Di Palo and Edward Johns. R+x: Retrieval and execution from everyday human videos. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.
- 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, 2024.
- Nick Heppert, Max Argus, Tim Welschehold, Thomas Brox, and Abhinav Valada. Ditto: Demonstration imitation by trajectory transformation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, page 7565–7572. IEEE, October 2024. doi: 10.1109/iros58592.2024.10801982. URL <http://dx.doi.org/10.1109/IROS58592.2024.10801982>.
- Arpit Bahety, Priyanka Mandikal, Ben Abbatematteo, and Roberto Martín-Martín. Screwmimic: Bimanual imitation from human videos with screw space projection, 2024.
- Yifeng Zhu, Arisrei Lim, Peter Stone, and Yuke Zhu. Vision-based manipulation from single human video with open-world object graphs, 2024.
- Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. In *Robotics: Science and Systems XVIII*, RSS2022. Robotics: Science and Systems Foundation, June 2022. doi: 10.15607/rss.2022.xviii.026. URL <http://dx.doi.org/10.15607/rss.2022.xviii.026>.
- 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.
- Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela Veloso, and Shuran Song. Xskill: Cross embodiment skill discovery, 2023.
- 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)*, page 6904–6911. IEEE, May 2024. doi: 10.1109/icra57147.2024.10610288. URL <http://dx.doi.org/10.1109/ICRA57147.2024.10610288>.
- Vidhi Jain, Maria Attarian, Nikhil J Joshi, Ayzaan Wahid, Danny Driess, Quan Vuong, Pannag R Sanketi, Pierre Sermanet, Stefan Welker, Christine Chan, Igor Gilitschenski, Yonatan Bisk, and Debidatta Dwibedi. Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers, 2024.
- Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *arXiv e-prints*, pages arXiv–2302, 2023.
- Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B Tenenbaum. Learning to Act from Actionless Videos through Dense Correspondences. *arXiv:2310.08576*, 2023.
- Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate: Real-world visuomotor policy learning via video generation, 2024.
- 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*, 2024b.
- Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. In *7th Annual Conference on Robot Learning*, 2023.