X-DIFFUSION: Training Diffusion Policies on Cross-Embodiment Human Demonstrations

Maximus A. Pace* Prithwish Dan* Chuanruo Ning Atiksh Bhardwaj Audrey Du
Edward W. Duan Wei-Chiu Ma† Kushal Kedia†
Cornell University

https://portal-cornell.github.io/X-Diffusion/

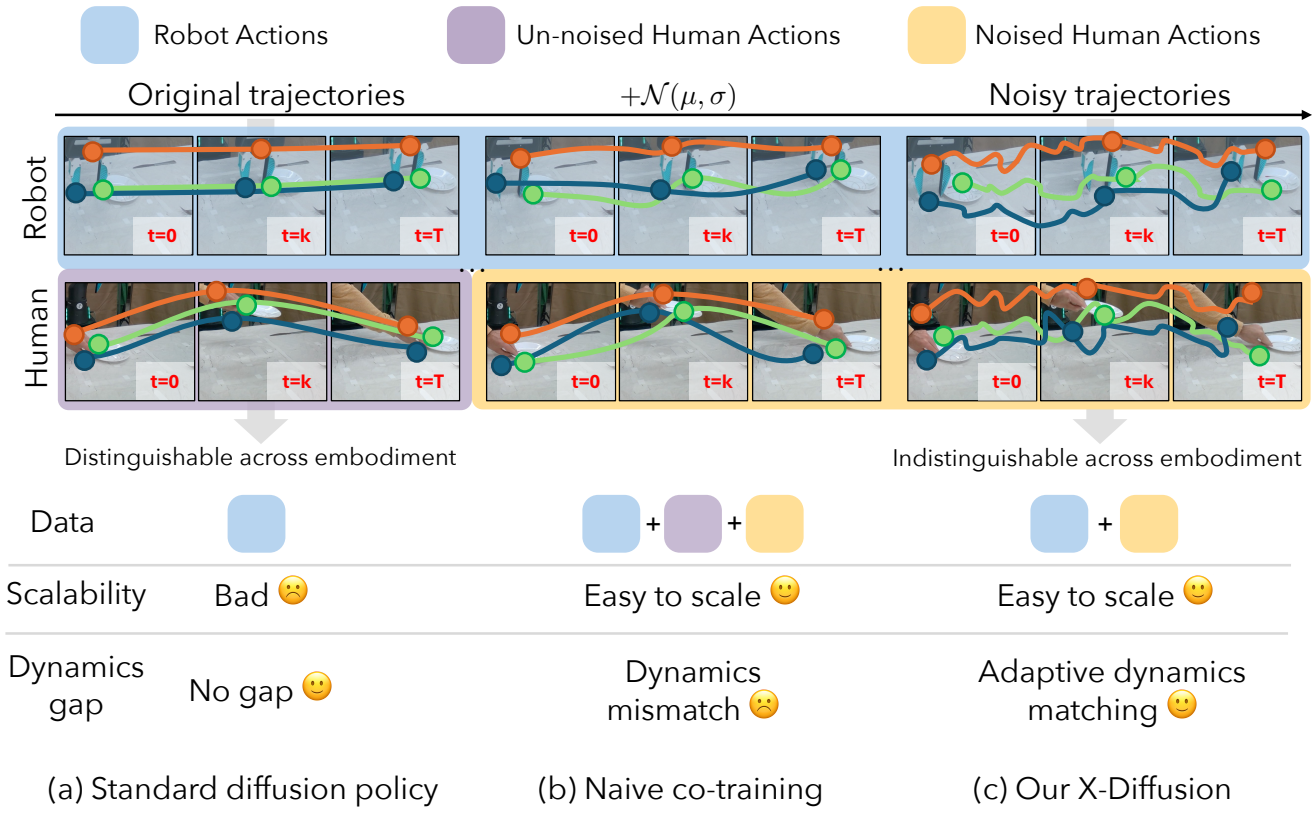

Fig. 1: **Overview of X-DIFFUSION:** We introduce X-DIFFUSION, a framework to train diffusion policies on cross-embodiment human data containing a variety of execution styles. Naively co-training diffusion policies on human and robot datasets with mismatched dynamics can lead the denoising process to output dynamically infeasible actions for the robot, degrading performance below standard robot-only diffusion policy training. Instead, X-DIFFUSION trains a classifier to distinguish between noised human and robot actions, and integrates noised human actions into policy training only when the classifier is unsure of which embodiment produced the actions, thus, effectively learning from large and diverse human demonstrations.

***Abstract*— Human videos can be recorded quickly and at scale, making them an appealing source of training data for robot learning. However, humans and robots differ fundamentally in embodiment, resulting in mismatched action execution. Direct kinematic retargeting of human hand motion can therefore produce actions that are physically infeasible for robots. Despite these low-level differences, human demonstrations provide valuable motion cues about how to manipulate and interact with objects. Our key idea is to exploit the forward diffusion process: as noise is added to actions, low-level execution differences fade while high-level task guidance is preserved. We present X-DIFFUSION, a principled framework for training diffusion policies that maximally leverages human data without learning dynamically infeasible motions. X-DIFFUSION first trains a classifier to predict whether a noisy action is executed by a human or robot. Then, a human action is incorporated into policy training only after adding sufficient noise such that the classifier cannot discern its embodiment. Actions consistent with robot execution supervise fine-grained denoising at low noise levels, while mismatched human actions provide only coarse guidance at higher noise levels. Our experiments show that naive co-training under execution mismatches degrades policy performance, while X-DIFFUSION consistently improves it. Across five manipulation tasks, X-DIFFUSION achieves a 16% higher average success rate than the best baseline.**

* Equal contribution. † Equal advising.

I. Introduction

Imitation learning (IL) is an effective and flexible method for teaching robot skills, but collecting large amounts of robot data is costly and slow. Human video demonstrations offer a scalable alternative, since they are easier and faster to collect. However, such data cannot be directly used to train state-of-the-art IL methods [1, 2], since humans and robots significantly differ in embodiment.

To leverage human data, recent works aim to unify human and robot action spaces [3–5]. Utilizing advances in 3D hand-pose estimation [6], hand motions recorded from human videos can be converted into robot actions via kinematic retargeting. Consequently, these methods have been applied to crowd-sourced human video collection [7, 8] and datasets of egocentric human videos [9, 10]. Nevertheless, these approaches typically rely on fine-tuning with robot teleoperation data to produce actions compatible with the robot's kinematics and dynamics. If final robot performance is the goal, is it optimal to train on all human demonstrations indiscriminately — or can some demonstrations, mismatched in execution with the robot, hurt policy learning?

We study this problem setting where we have access to a large dataset of human demonstrations, and a small dataset of high-quality robot teleoperation data. Even for the same manipulation task, humans and robots often differ in execution style. For example, in Fig. 1, when tasked with moving a plate on the table, a human can dexterously slide their fingers underneath to pick and place it. However, a robot with a parallel-jaw gripper may more reliably push or slide the plate across the table. Even with such differences in action execution, human videos still provide rich motion cues about how objects should be manipulated and interacted with. In diffusion policy learning, as noise is added to both human and robot actions, these low-level discrepancies start to fade away, preserving high-level guidance on how to complete the task. ***Our key insight is that training diffusion policies on noised human actions can improve task performance without sacrificing robot feasibility.***

We propose X-Diffusion, a framework for co-training diffusion policies with large-scale human demonstrations and a smaller set of robot teleoperation data (Fig. 1). Before policy training, we train a classifier to distinguish between noised human and robot actions in the forward diffusion process. To maximally leverage human data without introducing dynamically infeasible behaviors, we define the *minimum indistinguishability step*: the earliest diffusion step at which the classifier can no longer discern whether an action comes from a human or a robot. Actions that are compatible with robot kinematics and dynamics are integrated at lower noise levels, while actions that diverge from the robot's execution style are only included at higher noise levels. As a result, feasible human and robot demonstrations provide precise, low-level supervision throughout the diffusion process, whereas mismatched human actions contribute only coarse, high-level guidance. This enables our method to extract useful signal from all human data while avoiding degradation from execution mismatches. Our contributions:

1) We propose X-Diffusion, a framework for training diffusion policies on cross-embodiment human data while preserving dynamically feasible robot motions.
2) We demonstrate that prior methods that train on all human demonstrations often generate infeasible robot actions. Through ablations and analyses, we demonstrate that our selective training strategy outperforms both naive co-training and manual human annotation.
3) Across 5 manipulation tasks, we show that X-Diffusion outperforms a range of cross-embodiment learning baselines by 16% on average.

II. Related Work

Our work is related to the following topics:

Learning from Human Hand Motion. Human videos typically lack action annotations that can be directly executed by robots. Recent progress in hand-pose estimation has enabled retargeting human hand motion into robot actions. A common approach is to track 6DoF hand trajectories and map them to the robot end-effector [11–13], which is particularly effective for dexterous robotic hands [14–17]. Other works define corresponding keypoints between human and robot hands [3, 4] to unify their state and action spaces. Retargeting has also been leveraged to synthesize robot data by overlaying rendered robot arms on human videos [5, 18, 19]. When combined with open-world vision models, these methods can further enable object-aware retargeting of human motion [20–22]. While promising, these approaches often assume that every human hand motion can be feasibly executed by a robot. This assumption breaks down in practice, as many human actions involve kinematics or dynamics outside a robot's capabilities, limiting the reliability of direct retargeting.

Extracting Rewards from Human Data Beyond imitation, reinforcement learning (RL) can leverage human data by defining rewards from tracking reference motion [23, 24], video similarity [25–27], language alignment [28, 29], or object-centric signals in real-to-sim-to-real pipelines [30–32]. Preference learning methods further derive rewards from object interactions or classifier judgments of task success [33–36]. However, a common limitation of all these approaches is the requirement of a realistic simulator or costly and unsafe real-world environment interactions for RL. In contrast, we train diffusion policies directly on mixed human–robot data without requiring interactions with the environment to learn actions that match rewards extracted from human videos.

One-Shot Imitation from Human Videos. Without a direct mapping from human to robot actions, prior work has explored one-shot imitation, where robots attempt a task after a single human demonstration. Some methods learn correspondences from paired human–robot videos [38, 39], but such datasets are costly to scale. Others unify visual embeddings of humans and robots [40, 41], yet require large teleoperated robot datasets. Recent works have framed one-shot imitation as an in-context learning problem, where the human video acts a guide to retrieve the task-relevant

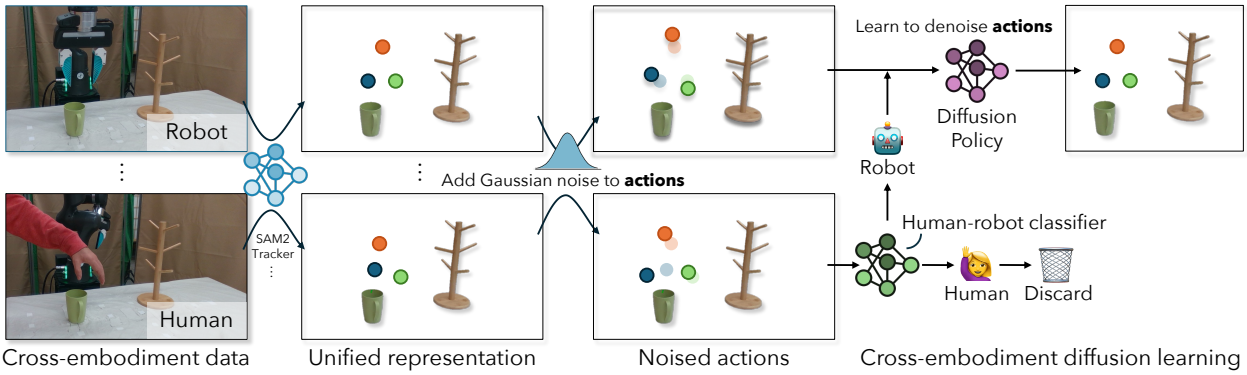

Fig. 2: **Pipeline:** X-DIFFUSION first unifies the state and action representation. State is represented by a colored segmentation mask of relevant objects using Grounded-SAM2 [37]. Action is represented via end-effector/human hand pose utilizing HaMeR [6] for retargeting. During the policy's forward diffusion process, Gaussian noise is sampled and added to the clean actions. To determine if the policy should learn to denoise noisy human actions into robot actions, X-DIFFUSION utilizes a classifier trained to distinguish the source embodiment of noised actions. Actions are only included for training the denoising process if the classifier is fooled into thinking it's from a robot. Thus, we learn from broad human data without learning infeasible actions.

behaviors [42, 43]. DemoDiffusion [44] is a diffusion-based approach that prompts a pretrained diffusion policy with a retargeted human trajectory as the initial noise. Still, this remains limited to a single video and relies heavily on the base robot policy. In contrast, our method learns directly from multiple human demonstrations, storing the knowledge into a single policy.

Learning from Sub-Optimal Data. As robot datasets scale, filtering low-quality demonstrations becomes critical. Prior work down-weights or removes poor trajectories using online rollouts [45, 46], which is costly, or by proxy loss metrics [47], which poorly predict real-world performance. In generative modeling, similar challenges are addressed by training classifiers to detect data quality under noise and selectively updating diffusion models [48–50]. Inspired by this, we adapt the idea to robotics: using diffusion policies [1] as the generative model and treating cross-embodiment human as a low-quality data source.

III. Problem Formulation

Our goal is to learn a robot policy $\pi_\theta(\mathbf{A_t}|s_t)$, which predicts a sequence of future actions $\mathbf{A_t} = a_{t:t+S}$ over the next S timesteps given the current robot state s_t. Training relies on two sources of supervision: a small, high-quality dataset of robot demonstrations $\mathcal{D}_R$ and a larger dataset of human demonstrations $\mathcal{D}_H$. Each dataset contains trajectories of state–action pairs $\xi = \{s_t, a_t\}_{t=1}^{T}$. Following prior work [3, 4], we unify the state and action spaces of humans and robots: states comprise proprioceptive inputs and third-person camera views, while actions are represented by the motion of the human hand or the robot end-effector.

Co-Training of Robot Policies. Cross-embodiment datasets are typically leveraged for policy learning by *co-training* with the robot dataset. A straightforward approach is to simply combine the robot dataset $\mathcal{D}_R$ and the human dataset $\mathcal{D}_H$ and train on the aggregated mixture:

$$\mathcal{L}_{\text{co-train}}(\theta) = \mathbb{E}_{(s_t, \mathbf{A_t}) \sim \mathcal{D}_R \cup \mathcal{D}_H} \left[\ell\big(\pi_\theta(s_t), \mathbf{A_t}\big) \right], \quad (1)$$

where ℓ denotes the behavior cloning loss function. This co-training paradigm treats human and robot data as interchangeable, assuming human and robot action dynamics are matched, i.e., $p_H(\mathbf{A_t} = a_{t:t+S}|s_t) \approx p_R(\mathbf{A_t} = a_{t:t+S}|s_t)$. However, differences in embodiment and execution style mean that human actions are often physically infeasible for the robot. As a result, naive co-training can significantly degrade policy performance, motivating the need for more selective co-training strategies.

IV. Approach

Naive co-training on human and robot demonstrations can degrade performance when execution styles are mismatched. We present X-DIFFUSION, a framework to maximally utilize cross-embodiment data for diffusion policy learning without degrading performance. At its core, X-DIFFUSION trains a classifier to distinguish between noised human and robot actions. Noised human actions are integrated into policy training only when the classifier is confused about its embodiment. This approach allows us to utilize large datasets of cross-embodiment demonstrations without learning to execute physically infeasible robot actions.

A. Cross-Embodiment Equivalence under Noise

Due to differences in embodiment, kinematic retargeting of human hand actions may result in physically infeasible robot motion. Still, human hand motion provides rich cues for what steps to follow, which objects to interact with, and how to interact with them. The usefulness of these cues depends on their alignment with the robot's action dynamics.

Diffusion policies [1] learn by denoising action sequences corrupted with Gaussian noise. Given the ground-truth robot or human action sequence $\mathbf{A_t^0}$, the *forward diffusion process* q produces progressively noisier versions $\mathbf{A_t^1}, \ldots, \mathbf{A_t^K}$ via:

$$q(\mathbf{A_t^{k+1}} \mid \mathbf{A_t^k}) = \mathcal{N}\Big(\sqrt{1-\beta_k}\,\mathbf{A}_t^k,\ \beta_k I\Big)$$

, where β_k controls the amount of additive Gaussian noise at diffusion step k. Our key observation is that the *forward diffusion* process progressively removes embodiment-specific

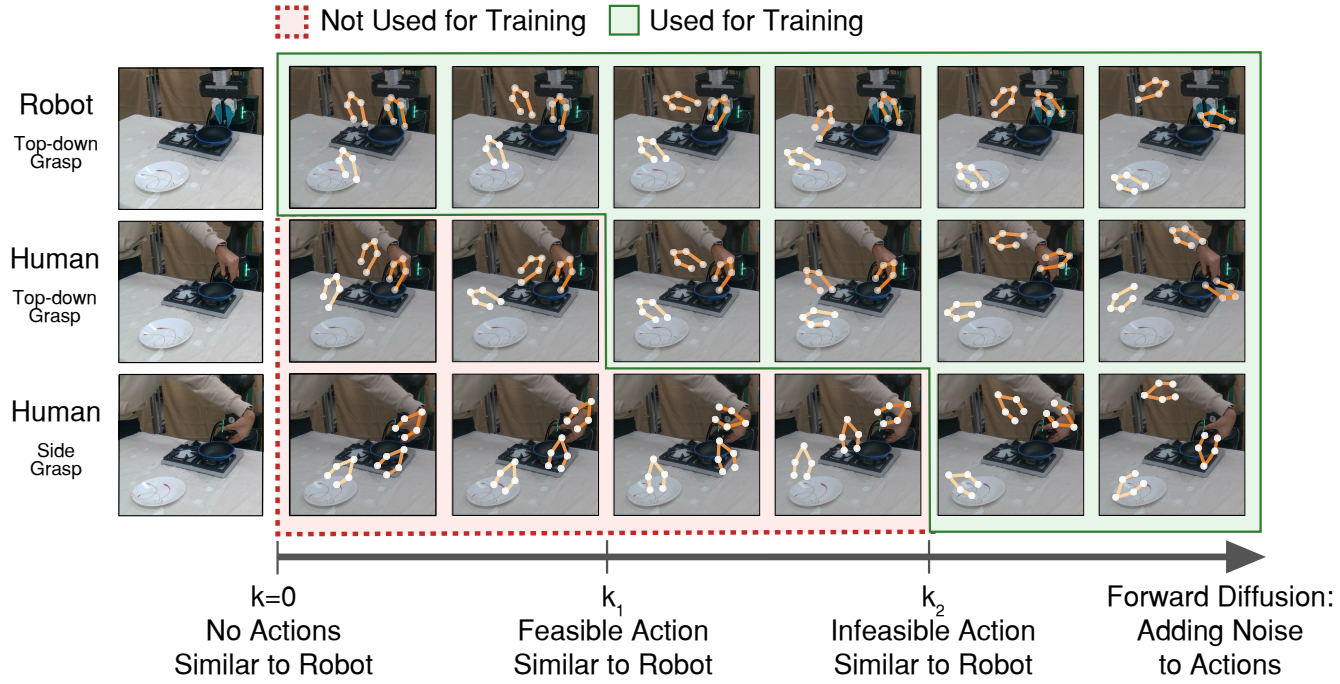

Fig. 3: **Visualizing Actions under Noise and Classifier Predictions at various Diffusion Steps**. Humans execute tasks in various ways. For example, when picking and placing a pan, a human can either execute a top-down grasp or a side grasp. Human actions that are feasible for robots (e.g. top-down grasp) overlap with robot action distribution under low noise timesteps. This data fools the classifier into believing it could have been executed by a robot, so we include it in the diffusion denoising process during policy training. In contrast, human actions that are kinematically and dynamically infeasible for robots (e.g. side grasp) are accurately identified as human actions by the classifier until significantly more noise is added in the forward diffusion process, restricting their impact on policy learning to only supervise coarse guidance at high noise.

features from actions. As shown in Fig. 1, *at high noise levels, human and robot trajectories become indistinguishable*.

Formally, let p_H^k and p_R^k denote the distributions of human and robot actions after k steps of the forward diffusion process. We define the **minimum indistinguishability step** $\mathbf{k}^\star$ as the earliest diffusion step where the two distributions overlap such that they cannot be reliably distinguished:

$$k^\star = \min\left\{k \mid D_{KL}\left(p_H^k \parallel p_R^k\right) \leq \epsilon\right\},$$

where ϵ is a small threshold. Intuitively, $k^\star$ identifies the point in the noising process at which human actions are sufficiently abstracted that they resemble robot actions. Beyond this step ($k \geq k^\star$), human demonstrations can provide effective supervision for robot policy learning without risking the transfer of infeasible motions.

B. Training a Noised Human-Robot Action Classifier

To determine the minimum indistinguishability timestep $k^\star$ for each action, we train a classifier that predicts the embodiment of a noised action. The classifier $c_\theta(\cdot|k, \mathbf{A_t^k}, s_t)$ takes as input the diffusion step k, the noised action sequence $\mathbf{A}_t^{(k)}$, and the current state s_t, and outputs the probability that the action originated from the robot ($y = 1$) or a human ($y = 0$). Training samples are drawn from both the human dataset $\mathcal{D}_H$ and the robot dataset $\mathcal{D}_R$. Since the size of the human dataset is much larger than the robot dataset $|\mathcal{D}_H| \gg |\mathcal{D}_R|$, we sample actions from both datasets with equal probability to prevent the classifier from being biased toward predicting the human label. The classifier is optimized using the binary cross-entropy loss:

$$\begin{aligned}\mathcal{L}_{\text{class}}(\theta) = \mathbb{E}_{(k,\mathbf{A_t^k},s_t)\sim\mathcal{D}_R}\left[-\log c_\theta(k, \mathbf{A_t^k}, s_t)\right] \\ + \mathbb{E}_{(k,\mathbf{A_t^k},s_t)\sim\mathcal{D}_H}\left[-\log\big(1 - c_\theta(k, \mathbf{A_t^k}, s_t)\big)\right].\end{aligned} \tag{2}$$

The classifier enables us to annotate human demonstrations with the timestep at which their noised actions become indistinguishable from robot actions. For each human action sequence $\mathbf{A}_t$, we define the **minimum indistinguishability step** $k^\star$ as the earliest diffusion step where the classifier assigns at least 50% probability to it being a robot action:

$$k^\star(\mathbf{A}_t) = \min\left\{k : c_\theta(k, \mathbf{A}_t^k, s_t) \geq 0.5\right\}. \tag{3}$$

C. Classifier Integration into Diffusion Policy

Diffusion policies model the reverse process of denoising with a neural network. Starting from Gaussian noise $\mathbf{A}_t^K$, the reverse model $p_\theta(\mathbf{A}_t^{k-1} \mid k, \mathbf{A}_t^k, s_t)$ iteratively denoises step by step until recovering the clean action sequence $\mathbf{A}_t^0$. Each reverse step attempts to predict a slightly less noisy action, conditioned on the current state s_t. Naive co-training (Eq. 1) supervises the reverse process using human actions across all diffusion steps. If human data is used indiscriminately at all noise levels, the policy is forced to denoise toward actions that may be kinematically infeasible for the robot.

Integration beyond the indistinguishability step. Our classifier resolves this problem by identifying, for each

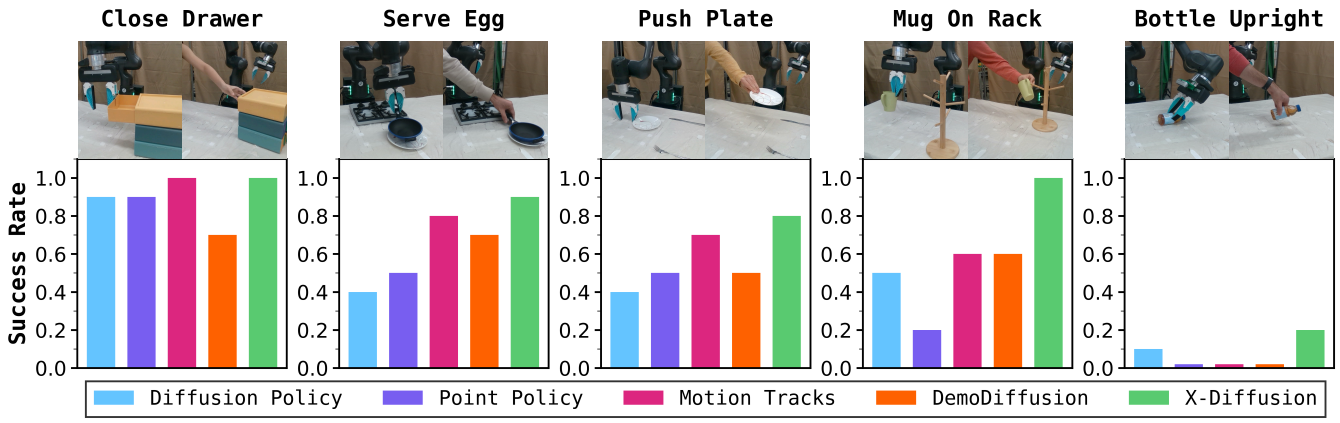

Fig. 4: **Performance vs. Baselines:** We report *task success rate* on 5 different manipulation tasks and compare X-DIFFUSION against a robot-only baseline (Diffusion Policy) and various co-training baselines (Point-Policy, MotionTracks). DemoDiffusion is another diffusion-based method, but it doesn't train the robot policy on human demonstrations. We find that X-DIFFUSION is the highest performing model on all tasks, effectively incorporating human action data into its training recipe even when execution styles are mismatched. One human and robot demonstration is visualized for each task.

human action, the minimum indistinguishability step $k^\star$ where the action distribution sufficiently overlaps with the robot action distribution under noise. During diffusion policy training, we only integrate human actions into the loss when $k \geq k^\star$ (using Eq. 2). Fig. 3 shows the minimum indistinguishability step on the `Serve-Egg` task for different human actions in the dataset. Actions that are kinematically feasible for the robot have low k^* whereas infeasible actions have higher k^*. Formally, our diffusion policy loss is:

$$\begin{aligned}\mathcal{L}_{\text{X-DP}}(\theta) = \mathbb{E}_{(k,\mathbf{A}_t,s_t)\sim\mathcal{D}_R}\ \ell\big(p_\theta, \mathbf{A}_t^k\big) \\ + \mathbb{E}_{(k,\mathbf{A}_t,s_t)\sim\mathcal{D}_H}\ \mathbf{1}_{\{k\geq k^\star(\mathbf{A}_t)\}}\,\ell\big(p_\theta, \mathbf{A}_t^k\big)\,,\end{aligned} \tag{4}$$

where ℓ denotes the denoising loss. This selective integration ensures that we maximally utilize human demonstrations without sacrificing kinematic feasibility of action execution.

D. Unifying State and Action Spaces

We convert human videos into robot-aligned state–action pairs with a minimal pipeline. Assumptions: (i) single-hand demonstrations that begin with an open grasp, and (ii) two calibrated RGB cameras. Using HaMeR [6], we detect 2D hand keypoints in both views and triangulate them to 3D in the robot frame. The grasp point is the mean of the thumb and index fingertips; orientation is obtained by fitting a local hand frame and retargeting to the robot end-effector following prior work [3, 4]. Gripper state is inferred using the distance between the thumb and index keypoints. To reduce the visual domain gap, we segment task-relevant objects with Grounded-SAM 2 [51] and overlay a keypoint rendering of the end-effector pose on each frame, as depicted in Fig. 2. The policy input is the masked image with overlaid keypoints, concatenated with proprioceptive information. More details are provided in the Appendix.

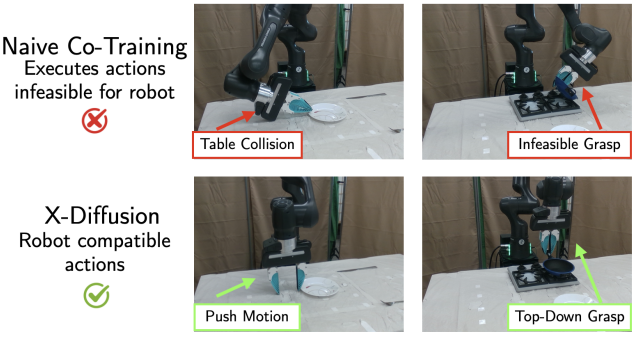

Fig. 5: **Naive co-training learns infeasible robot actions:** Including all human data in policy training can incentivize policies to learn strategies demonstrated by humans but infeasible for robots. On multiple tasks, a human may manipulate objects in ways that are not realizable for a robot.

V. EXPERIMENTS

We evaluate the ability of X-DIFFUSION to learn 5 different manipulation skills from cross-embodiment human data. Our experiments are designed to address four key questions:

1) Does X-DIFFUSION outperform prior cross-embodiment learning approaches?
2) Does naive co-training generate kinematically or dynamically infeasible motion on the robot?
3) How does the learned classifier compare to manual data filtering via human annotation?
4) How does the usefulness of human data vary across tasks?

Experimental Setup. For each manipulation task, we collect 5 robot demonstrations and 100 human demonstrations. Human demonstrations are performed with a single hand, while the robot is a 7-DOF Franka Emika Panda arm. We evaluate across five diverse tasks: pick-and-place (`Serve Egg`), non-prehensile manipulation (`Close Drawer`, `Push Plate`), precise insertion (`Mug On Rack`), and reorientation (`Bottle Upright`). These tasks span a wide range of manipulation skills and provide a

	X-DIFFUSION	FILTERED	NAIVE	ROBOT
Mug On Rack	**10/10**	8/10	6/10	5/10
Serve Egg	**9/10**	6/10	5/10	4/10
Push Plate	**8/10**	6/10	2/10	4/10

TABLE I: We compare the performance of X-DIFFUSION with a policy trained only on human demonstrations verified as robot-feasible (FILTERED), a naively trained policy using all available human data (NAIVE), and a policy trained only on robot data (ROBOT). We find X-DIFFUSION outperforms all baselines for each task.

comprehensive benchmark for assessing the value of human data in policy training for different manipulation skills. We evaluate each method over 10 real-world rollouts per task and report average success rates.

Baselines. We compare against the following baselines:

1) **Diffusion Policy [1]**: This method trains only on 5 robot demonstrations, potentially lacking coverage of the initial state distribution.
2) **Point Policy [4]**: This method co-trains on all human and robot data, unifying the cross-embodiment observation and actions via hand and object keypoints.
3) **Motion Tracks [3]**: This method co-trains on all human and robot data. It unifies the action space as hand keypoints but uses raw RGB image observations.
4) **DemoDiffusion [44]**: This method conditions a robot-only diffusion policy with a human trajectory. More details are provided in the Appendix.

A. Comparison with Cross-Embodiment Learning Baselines.

We evaluate X-DIFFUSION's ability to learn from human demonstrations, and compare performance against existing cross-embodiment policy learning methods. We find that X-DIFFUSION achieves higher success rates across tasks relative to Point Policy, Motion Tracks, and DemoDiffusion (Fig. 4). Qualitatively, we observe that these approaches share a common failure mode: executing actions that appear in human demonstrations which are infeasible for the robot, as shown in Fig. 5. In `Push Plate` and `Serve Egg`, several human demonstrations grasp objects from the side (instead of top-down), a strategy which is kinematically infeasible for the robot to perform. Naively co-training with an uncurated set of human demonstrations yields little to no improvements (Motion Tracks, DemoDiffusion) over robot-only training, and can even degrade performance (Point Policy) by learning suboptimal robot behaviors.

In contrast to all these methods, X-DIFFUSION leverages its classifier to directly filter out action sequences that have low probabilities of being classified as a robot trajectory, only applying the action denoising loss on (noisy) human motions which are indistinguishable from robot motion. This training recipe consistently improves performance over robot-only training and naive co-training by carefully including human data from a wider state distribution.

B. Comparing Classifier with Manual Human Annotations

To further investigate the human data distribution and its impact on policy learning, we design an experiment with a FILTERED policy. We replay human demonstrations on the robot via Inverse Kinematics (IK) and manually filter out unsuccessful trajectories to construct $\mathcal{D}_H^+$, a dataset of feasible human demonstrations. Concretely, we train three policies with the same architecture but vary the data:

- **ROBOT:** Trained only on $\mathcal{D}_R$.
- **NAIVE:** Trained on $\mathcal{D}_R \cup \mathcal{D}_H$.
- **FILTERED:** Trained on $\mathcal{D}_R \cup \mathcal{D}_H^+$.
- **X-DIFFUSION:** Trained on $\mathcal{D}_R \cup \mathcal{D}_H$, discarding human data below the **minimum indistinguishability step** (Sec. IV) during action denoising.

Table I shows that FILTERED dataset co-training outperforms NAIVE co-training, confirming the hypothesis that the inclusion of infeasible human demonstrations at train time degrades policy performance. X-DIFFUSION takes an alternate approach—instead of discarding entire trajectories and applying the action denoising loss at all noise levels for successful human trajectories in $\mathcal{D}_H^+$, it adaptively includes human data from $\mathcal{D}_H$ only beyond noise levels where the human and robot data distributions are indistinguishable, thus learning to denoise within the correct distribution for the robot. We visualize this phenomenon in Fig. 3: as Gaussian noise is added to human actions, our classifier is unable to identify which embodiment executed the actions. We also observe that the minimum indistinguishability step is lower for feasible human actions than their infeasible counterpart. X-DIFFUSION outperforms the FILTERED policy across all tasks, demonstrating the ability to extract signal even from infeasible human demonstrations.

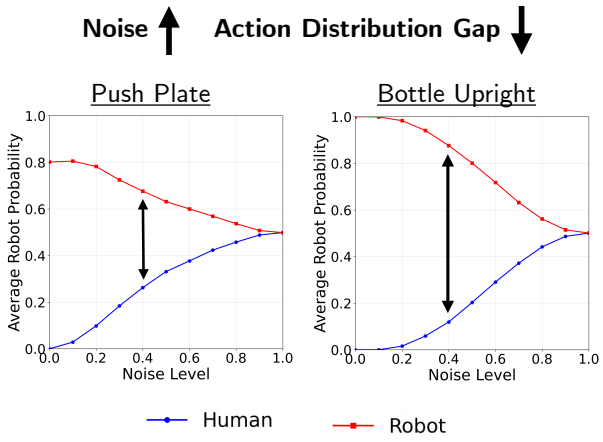

Fig. 6: **Classifier Robot Probability across forward diffusion process:** As the noise levels increase, the human action distribution becomes more similar to the robot action distribution. The similarity of human actions with robot actions varies across tasks: as shown on the graphs, the distance between the human and robot action distributions at every noise level is smaller for `Push Plate` data compared to `Bottle Upright`. Consequently, we find that our policy improves performance more by training on the former task.

C. Analyzing Human Data Quality Across Tasks

Human demonstrations are widely unstructured compared to robot teleoperation, varying in strategy and speed. This leads to non-uniform performance gains across tasks for all

cross-embodiment methods (Fig. 4). We systematically analyze the quality of human data, particularly focused on `Push Plate` and `Bottle Upright`. In the `Push Plate` task, all cross-embodiment methods outperform robot-only DP, with the biggest beneficiary being X-DIFFUSION (90% vs. 40%). In contrast, X-DIFFUSION only slightly improved over DP (20% vs. 10%) on `Bottle Upright`, while all other methods dropped to 0% success.

We probe X-DIFFUSION's learned classifier at various noise levels for both tasks, and plot the average predicted robot probability over all the human and robot data in Fig. 6. Notably, we observe that at low noise regimes, the classifier is extremely confident and accurate in its class predictions given `Bottle Upright` actions as input. For `Push Plate`, as noise is added, we quickly see the gap between human and robot probability gap shrink as their (noisy) action distributions become more similar, indicating a strong correlation with the policy success rates. These observations also support our findings from the data, where we noticed that `Bottle Upright` human demonstrations were (a) much faster than robot execution and (b) prone to retargeting errors, making them detrimental when naively co-training without careful data curation. X-DIFFUSION autonomously discards these suboptimal demonstrations with the learned classifier, avoiding policy degradation.

VI. DISCUSSION

In this paper, we propose X-DIFFUSION, a scalable framework for co-training robot policies on cross-embodiment data by selectively incorporating human actions according to their feasibility for the robot. X-DIFFUSION employs a classifier to determine the noising timestep in the forward diffusion process at which a human action becomes indistinguishable from a robot action. By including human data only once it is sufficiently noised, our approach integrates demonstrations that are compatible with robot execution to provide strong denoising signal, while filtering out those that could otherwise degrade performance and generate dynamically infeasible motion. This enables effective use of large-scale human datasets that are not curated for robot learning. Our empirical evaluation across five manipulation tasks shows that X-DIFFUSION consistently outperforms both robot-only policies and prior co-training baselines, even when the human data is of low quality.

Limitations. In our work, we train X-DIFFUSION on a limited number of robot and human demonstrations in a calibrated multi-camera environment. Future works will attempt to train classifiers on large-scale datasets and learn from unstructured internet-scale human videos.

VII. ACKNOWLEDGMENTS

The research is partially supported by a gift from Ai2, a NVIDIA Academic Grant, and DARPA TIAMAT program No. HR00112490422. This research is also supported in part by Google Faculty Research Award, OpenAI Super-Alignment Grant, ONR Young Investigator Award, NSF RI #2312956, and NSF FRR #2327973. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of DARPA.

REFERENCES

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

[2] T. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *ArXiv*, vol. abs/2304.13705, 2023.

[3] J. Ren, P. Sundaresan, D. Sadigh, S. Choudhury, and J. Bohg, "Motion tracks: A unified representation for human-robot transfer in few-shot imitation learning," *ArXiv*, vol. abs/2501.06994, 2025.

[4] S. Haldar and L. Pinto, "Point policy: Unifying observations and actions with key points for robot manipulation," *ArXiv*, vol. abs/2502.20391, 2025.

[5] M. Lepert, J. Fang, and J. Bohg, "Phantom: Training robots without robots using only human videos," *ArXiv*, vol. abs/2503.00779, 2025.

[6] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. F. Fouhey, and J. Malik, "Reconstructing hands in 3d with transformers," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9826–9836, 2023.

[7] T. Tao, M. K. Srirama, J. J. Liu, K. Shaw, and D. Pathak, "Dexwild: Dexterous human interactions for in-the-wild robot policies," *ArXiv*, vol. abs/2505.07813, 2025.

[8] V. Liu, A. Adeniji, H. Zhan, R. M. Bhirangi, P. Abbeel, and L. Pinto, "Egozero: Robot learning from smart glasses," *ArXiv*, vol. abs/2505.20290, 2025.

[9] M. Lepert, J. Fang, and J. Bohg, "Masquerade: Learning from in-the-wild human videos using data-editing," *ArXiv*, vol. abs/2508.09976, 2025.

[10] J. Shi, Z. Zhao, T. Wang, I. Pedroza, A. Luo, J. Wang, J. Ma, and D. Jayaraman, "Zeromimic: Distilling robotic manipulation skills from web videos," *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16 939–16 947, 2025.

[11] H. Bharadhwaj, A. Gupta, S. Tulsiani, and V. Kumar, "Zero-shot robot manipulation from passive human videos," vol. abs/2302.02011, 2023.

[12] H. Bharadhwaj, R. Mottaghi, A. Gupta, and S. Tulsiani, "Track2act: Predicting point tracks from internet videos enables diverse zero-shot robot manipulation," *arXiv preprint arXiv:2405.01527*, 2024.

[13] C. Wang, L. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu, and A. Anandkumar, "Mimicplay: Long-horizon imitation learning by watching human play," 2023.

[14] A. Sivakumar, K. Shaw, and D. Pathak, "Robotic telekinesis: Learning a robotic hand imitator by watching humans on youtube," 2022.

[15] K. Shaw, S. Bahl, and D. Pathak, "Videodex: Learning

dexterity from internet videos," in *Conference on Robot Learning*, 2022.

[16] S. P. Arunachalam, S. Silwal, B. Evans, and L. Pinto, "Dexterous imitation made easy: A learning-based framework for efficient dexterous manipulation," 2022.

[17] J. Ye, J. Wang, B. Huang, Y. Qin, and X. Wang, "Learning continuous grasping function with a dexterous hand from human demonstrations," vol. 8, 2022, pp. 2882–2889.

[18] M. Lepert, R. Doshi, and J. Bohg, "Shadow: Leveraging segmentation masks for cross-embodiment policy transfer," *ArXiv*, vol. abs/2503.00774, 2025.

[19] S. Bahl, A. Gupta, and D. Pathak, "Human-to-robot imitation in the wild," 2022.

[20] Y. Zhu, A. Lim, P. Stone, and Y. Zhu, "Vision-based manipulation from single human video with open-world object graphs," *arXiv preprint arXiv:2405.20321*, 2024.

[21] P. Vitiello, K. Dreczkowski, and E. Johns, "One-shot imitation learning: A pose estimation perspective," in *Conference on Robot Learning*, 2023.

[22] 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," *ArXiv*, vol. abs/2410.11792, 2024.

[23] X. B. Peng, P. Abbeel, S. Levine, and M. van de Panne, "Deepmimic: Example-guided deep reinforcement learning of physics-based character skills," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 143:1–143:14, Jul. 2018.

[24] Z. Yuan, T. Wei, L. Gu, P. Hua, T. Liang, Y. Chen, and H. Xu, "Hermes: Human-to-robot embodied learning from multi-source motion data for mobile dexterous manipulation," 2025.

[25] K. Zakka, A. Zeng, P. R. Florence, J. Tompson, J. Bohg, and D. Dwibedi, "Xirl: Cross-embodiment inverse reinforcement learning," in *Conference on Robot Learning*, 2021.

[26] D. Yang, D. Tjia, J. Berg, D. Damen, P. Agrawal, and A. Gupta, "Rank2reward: Learning shaped reward functions from passive video," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 2806–2813.

[27] W. Huey, H. Wang, A. Wu, Y. Artzi, and S. Choudhury, "Imitation learning from a single temporally misaligned video," *ArXiv*, vol. abs/2502.05397, 2025.

[28] L. Shao, T. Migimatsu, Q. Zhang, K. Yang, and J. Bohg, "Concept2robot: Learning manipulation concepts from instructions and human demonstrations," vol. 40, 2020, pp. 1419 – 1434.

[29] A. S. Chen, S. Nair, and C. Finn, "Learning generalizable robotic reward functions from "in-the-wild" human videos," vol. abs/2103.16817, 2021.

[30] P. Dan, K. Kedia, A. Chao, E. W. Duan, M. A. Pace, W.-C. Ma, and S. Choudhury, "X-sim: Cross-embodiment learning via real-to-sim-to-real," 2025.

[31] T. Ga, W. Lum, O. Y. Lee, C. K. Liu, J. Bohg, and P.-M. H. Pose, "Crossing the human-robot embodiment gap with sim-to-real rl using one human demonstration," 2025.

[32] M. Xu, Z. Xu, Y. Xu, C. Chi, G. Wetzstein, M. Veloso, and S. Song, "Flow as the cross-domain manipulation interface," in *Conference on Robot Learning*, 2024.

[33] A. Nair, D. Chen, P. Agrawal, P. Isola, P. Abbeel, J. Malik, and S. Levine, "Combining self-supervised learning and imitation for vision-based rope manipulation," 2017, pp. 2146–2153.

[34] M. Sieb, X. Zhou, A. Huang, O. Kroemer, and K. Fragkiadaki, "Graph-structured visual imitation," in *Conference on Robot Learning*, 2019.

[35] K. Schmeckpeper, A. Xie, O. Rybkin, S. Tian, K. Daniilidis, S. Levine, and C. Finn, "Learning predictive models from observation and interaction," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX*. Springer, 2020, pp. 708–725.

[36] S. Kumar, J. Zamora, N. Hansen, R. Jangir, and X. Wang, "Graph inverse reinforcement learning from diverse videos," 2022.

[37] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," 2024.

[38] 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," vol. abs/2202.02005, 2022.

[39] V. Jain, M. Attarian, N. Joshi, A. Wahid, D. Driess, Q. Vuong, P. R. Sanketi, P. Sermanet, S. Welker, C. Chan, I. Gilitschenski, Y. Bisk, and D. Dwibedi, "Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers," vol. abs/2403.12943, 2024.

[40] K. Kedia, P. Dan, A. Chao, M. A. Pace, and S. Choudhury, "One-shot imitation under mismatched execution," *arXiv preprint arXiv:2409.06615*, 2024.

[41] M. Xu, Z. Xu, C. Chi, M. Veloso, and S. Song, "XSkill: Cross embodiment skill discovery," in *7th Annual Conference on Robot Learning*, 2023.

[42] R. Shah, S. Liu, Q. Wang, Z. Jiang, S. Kumar, M. Seo, R. Martín-Martín, and Y. Zhu, "Mimicdroid: In-context learning for humanoid manipulation from human play videos," *arXiv preprint arXiv:2509.09769*, 2025.

[43] V. Vosylius and E. Johns, "Instant policy: In-context imitation learning via graph diffusion," 2025.

[44] S. Park, H. Bharadhwaj, and S. Tulsiani, "Demodiffusion: One-shot human imitation using pre-trained diffusion policy," 2025.

[45] A. S. Chen, A. M. Lessing, Y. Liu, and C. Finn, "Curating demonstrations using online experience," 2025.

[46] C. Agia, R. Sinha, J. Yang, R. Antonova, M. Pavone, H. Nishimura, M. Itkina, and J. Bohg, "Cupid: Curating data your robot loves with influence functions," 2025.

[47] J. Hejna, C. Bhateja, Y. Jiang, K. Pertsch, and

D. Sadigh, "Re-mix: Optimizing data mixtures for large scale imitation learning," 2024.
[48] X. Dai, J. Hou, C.-Y. Ma, S. Tsai, J. Wang, R. Wang, P. Zhang, S. Vandenhende, X. Wang, A. Dubey, M. Yu, A. Kadian, F. Radenovic, D. Mahajan, K. Li, Y. Zhao, V. Petrovic, M. K. Singh, S. Motwani, Y. Wen, Y. Song, R. Sumbaly, V. Ramanathan, Z. He, P. Vajda, and D. Parikh, "Emu: Enhancing image generation models using photogenic needles in a haystack," 2023.
[49] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "Laion-5b: An open large-scale dataset for training next generation image-text models," 2022.
[50] G. Daras, A. Rodriguez-Munoz, A. Klivans, A. Torralba, and C. Daskalakis, "Ambient diffusion omni: Training good models with bad data," 2025.
[51] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan *et al.*, "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv*, 2024.
[52] L. Tang, M. Jia, Q. Wang, C. P. Phoo, and B. Hariharan, "Emergent correspondence from image diffusion," *Advances in Neural Information Processing Systems*, vol. 36, pp. 1363–1389, 2023.
[53] N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, and C. Rupprecht, "Cotracker: It is better to track together," in *European conference on computer vision*. Springer, 2024, pp. 18–35.

APPENDIX

A. Contributions

- **Maximus A. Pace:** Investigated different algorithms for using human data in policy learning, set up the data collection pipeline using teleoperation and human videos, major role in real world experiments and paper writing, worked on figure design, and co-led the project.
- **Prithwish Dan:** Set up the human-robot classifier, helped in real-world experiments, worked on figure design, paper writing, and co-led the project.
- **Chuanruo Ning:** Worked on the perception pipeline, setting up SAM2 real-world tracking and HaMeR.
- **Atiksh Bhardwaj:** Contributed to cross-embodiment learning baseline experiments and data collection.
- **Audrey Du:** Contributed to cross-embodiment learning baseline experiments and data collection.
- **Edward W. Duan:** Contributed to cross-embodiment learning baseline experiments and data collection.
- **Wei-Chiu Ma:** Suggested the idea of training diffusion policies with human actions after adding noise, major role in designing figures, and co-advised the project.
- **Kushal Kedia:** Conceived the goals of the project, suggested baselines and ablations, involved in coding, major role in paper writing, and co-advised the project.

B. Task Descriptions

We provide descriptions of the tasks described in Sec. V and Fig. 4.

- `Close Drawer`: close the top drawer of a cabinet. The cabinet's position and rotation is randomized across a line.
- `Serve Egg`: pick up a frying pan from stovetop and place it on a plate. The plate's location is randomized across a line.
- `Push Plate`: move a plate to between the fork and knife. The position of the fork and knife are randomized together along the table.
- `Mug On Rack`: pick up a mug and place its handle on a peg of a rack. The position of the rack is randomized across a line.
- `Bottle Upright`: pick up a juice bottle lying on its side from the table, reorient it, and release it standing upright.

C. State and Action Representations

Training a policy on cross-embodiment data requires unification of the state and action representations. Our datasets of robot demonstrations $\mathcal{D}_R$ and human demonstrations $\mathcal{D}_H$ contain trajectories of state–action pairs $\xi = \{s_t, a_t\}_{t=1}^{T}$. The embodiment-agnostic state $s_t = \{q_t, o_t\}$ consists of proprioception q_t and a visual observation of the scene o_t at each timestep t. The proprioception $q_t = \{p_t, r_t, g_t\} \in \mathbb{R}^7$ includes the 3D position $p_t \in \mathbb{R}^3$, rotation $r_t \in \mathbb{R}^3$, and gripper status $g_t \in \mathbb{R}$ of the end effector. The visual observation $o_t \in \mathbb{R}^{H \times W \times 3}$ includes 2D RGB segmentations of task-relevant objects with end-effector keypoint renderings overlaid. We simply consider the action at timestep t to be the proprioception at $t+1$, i.e. $a_t = q_{t+1}$. We record the states for each embodiment as follows:

1) Robot Demonstrations: The robot's proprioception q_t is computed using forward kinematics given its joint angles and gripper status (e.g. open or closed) at timestep t. Visual observations o_t are obtained by applying Grounded-SAM 2 [51] with language prompts on a single-view RGB capture of the scene and overlaying end-effector keypoint renderings.

2) Human Demonstrations: We use HaMeR [6] to detect a set of 21 keypoints in 2D pixel space for each camera. We select 5 of these keypoints along the index finger and thumb to be retargeted into a parallel jaw. Using two cameras with known parameters, we triangulate these keypoints into the same 3D coordinate frame as the robot to obtain p_t and apply the Kabsch algorithm to compute the rotation r_t. Finally, we calculate the gripper status g_t as a function of the distance between the tip of the index finger and the thumb, considering it to be closed under a fixed threshold and open otherwise. Visual observations o_t are processed using the same pipeline as robot demonstrations.

D. Baseline Implementation Details

All policies are trained using the Diffusion Policy [1] architecture, which consumes our unified state representation (unless otherwise specified) to predict action sequences.

1) Diffusion Policy: This baseline uses the vanilla Diffusion Policy architecture trained only on a small set of robot demonstrations.

2) Point Policy: Instead of using segmented images in its visual observation o_t, this baseline represents state via 3D keypoints of relevant objects at each timestep t. The keypoints are annotated in the first frame of one training demonstration, and correspondences are automatically detected at the start of all other demonstrations and at inference time using DIFT [52]. Co-Tracker [53] then tracks each point over time, and 3D object points are computed via triangulation from two cameras. This baseline is trained by equally sampling human and robot demonstrations.

3) Motion Tracks: This baseline consumes the raw RGB image (without segmentations) and end-effector proprioception as input. The original paper for MOTION TRACKS uses a keypoint retargeting network to minimize any gap between hand and end-effector keypoints, which we alleviate in our implementation by unifying the proprioception directly into end-effector position and rotation. This baseline is trained by equally sampling human and robot demonstrations.

4) DemoDiffusion: This baseline leverages two Diffusion Policies: human policy π^H is trained on the full human dataset $\mathcal{D}_H$, and robot policy π^R is trained on the full robot dataset $\mathcal{D}_R$. The reverse diffusion process is completed by using the human policy π^H for the initial denoising steps, followed by the robot policy π^R for the remainder of the denoising steps. We follow the original paper by using the human policy π^H for the first 60% of denoising steps and the robot policy π^R for the remaining 40%.

E. Diffusion Policy Hyperparameters

We use the same hyperparameters for training all of our policies and baselines. All of our policies use Diffusion Policy UNet architecture. The hyperparameters and values are provided in Table II.

TABLE II: Hyperparameters for Training Diffusion Policy

Diffusion Settings	
Diffusion timesteps (training)	100
Diffusion timesteps (inference)	20
Model Architecture	
Backbone CNN	ResNet50
Policy backbone	UNet
Image size	96×96
Temporal Horizon	
Observation horizon	1
Prediction horizon	8
Action horizon	8
Training	
Batch size	128
Learning rate	1×10^{-4}
Weight decay	0
Gradient clipping	5.0
Epochs	30
Gradient Steps Per Epoch	10,000
EMA decay rate	0.01
Evaluation	
Validation split ratio	0.15