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/

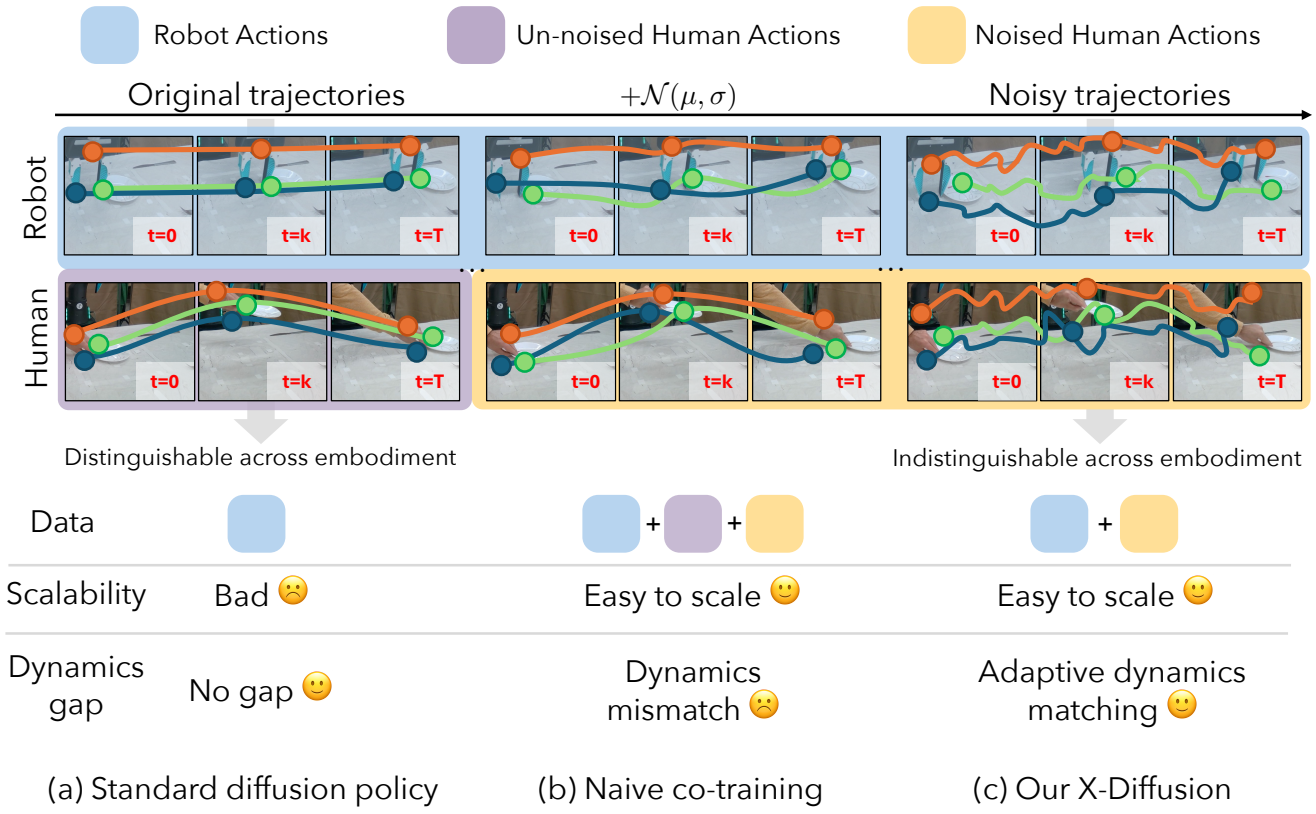

Fig. 1: **Overview of X-DIFFUSION:** We introduce X-DIFFUSION, a cross-embodiment learning framework that trains Diffusion Policies on human demonstrations even when their actions are not directly executable by the robot. Prior methods typically co-train on mixed human and robot datasets, which often causes the policy to learn actions that are dynamically infeasible on the robot. Instead, X-DIFFUSION integrates human actions into Diffusion Policy training only when they are sufficiently noised in the forward diffusion process, such that they are indistinguishable from robot actions. This enables the utilization of broad human data without sacrificing dynamic feasibility on the robot.

***Abstract*— Human videos are a scalable source of training data for robot learning. However, humans and robots significantly differ in embodiment, making many human actions infeasible for direct execution on a robot. Still, these demonstrations convey rich object-interaction cues and task intent. Our goal is to learn from this coarse guidance without transferring embodiment-specific, infeasible execution strategies. Recent advances in generative modeling tackle a related problem of learning from low-quality data. In particular, Ambient Diffusion is a recent method for diffusion modeling that incorporates low-quality data only at high-noise timesteps of the forward diffusion process. Our key insight is to view human actions as noisy counterparts of robot actions. As noise increases along the forward diffusion process, embodiment-specific differences fade away while task-relevant guidance is preserved. Based on these observations, we present X-DIFFUSION, a cross-embodiment learning framework based on Ambient Diffusion that selectively trains diffusion policies on noised human actions. This enables effective use of easy-to-collect human videos without sacrificing robot feasibility. Across five real-world manipulation tasks, we show that X-DIFFUSION improves average success rates by 16% over naive co-training and manual data filtering.**

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

* Equal contribution. † Equal advising.

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] because humans and robots significantly differ in embodiment.

To partially address this challenge, recent works propose to map human motions into the robot's action space [3–5]. By utilizing advances in 3D hand-pose estimation [6], hand motions extracted from human videos can be converted into robot end-effector actions via kinematic retargeting, making it possible to learn from large-scale human video datasets [7–10]. Yet such mappings only unify the representation of actions, not their physical realizability. Human executions often involve dynamics and contact strategies that are fundamentally mismatched with the robot's embodiment.

Consider the example in Fig. 1. Even for a simple manipulation task, humans and robots differ in execution style. When moving the plate, a human can dexterously slide their fingers underneath to pick it up, whereas a robot with a parallel-jaw gripper may more reliably push or slide the plate across the surface. This naturally raises a key question: how should we treat these human demonstrations? Even when the execution itself is not robot-feasible, human motions still provide rich cues about how objects could be manipulated and interacted with. Should we ignore the potential feasibility gap and train on all human data indiscriminately, or should those misaligned with the robot's capabilities be identified and discarded to prevent degrading policy performance?

Similar challenges exist in the field of generative modeling, where naively training on a mixture of low-quality and high-quality data often degrades model performance [11, 12]. While prior works filter low-quality samples [13, 14] or extract signals from noisy or corrupted data [15–18], Ambient Diffusion [19, 20] offers an exciting alternative by strategically integrating low-quality data into higher-noise timesteps of diffusion. In this paper, we build upon recent progress in learning from noisy data [19–23] to advance cross-embodiment learning. We show how these ideas can be integrated into prevailing robot-learning frameworks [1].

Our key idea is to *view human actions as a noisy counterpart to robot actions*. After mapping human and robot trajectories into a shared action space, embodiment-specific dynamics mismatches can be interpreted as manifestations of noise. During training, Diffusion Policies learn denoising networks by adding noise to action data. When a sufficient amount of noise is applied to both human and robot actions, low-level embodiment differences fade away while preserving the underlying task structure. Consequently, selectively training Diffusion Policies on noised human actions improves task performance without sacrificing robot feasibility.

Towards this goal, we train a classifier to distinguish between noised human and robot actions in the forward diffusion process. We then define the *minimum indistinguishability step* as the earliest diffusion step where the classifier can no longer discern an action's source embodiment. 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 Diffusion Policies to extract useful signals from all human data while avoiding degradation from execution mismatches.

We validate X-DIFFUSION on five real-world manipulation tasks exhibiting varying human-robot execution mismatch. While prior approaches that naively co-train on human data may generate infeasible robot actions, selectively training on human actions at high-noise levels improves upon naive co-training and even surpasses manual data filtering. X-DIFFUSION outperforms a range of cross-embodiment learning baselines by an average of 16% in task success.

II. RELATED WORK

Our work is related to the following topics:

Learning from Human Hand Motion. Advances in hand-pose estimation have enabled retargeting actionless human videos into robot actions. One approach is to track 6DoF hand trajectories and map them to the robot end-effector [24, 25]. Other works define corresponding keypoints between humans and robots to unify their data representations [3, 4], overlaying rendered robot arms on human videos [5, 26, 27]. Open-world vision models have further enabled object-aware retargeting [28–30]. These methods assume that retargeted hand motions will transfer cleanly to the robot, but this often fails in practice due to embodiment mismatch.

Extracting Rewards from Human Data. Reinforcement learning (RL) approaches leverage human data by defining rewards from tracking reference motion [31, 32], object-centric signals in real-to-sim-to-real pipelines [33, 34], and classifier judgments of task success [35]. However, these approaches are limited by the requirement of a realistic simulator or costly and unsafe real-world interactions. In contrast, we train Diffusion Policies directly on mixed human–robot data without requiring environment interactions.

One-Shot Imitation from Human Videos. Prior work has explored one-shot imitation, where robots attempt a task from a single human demonstration. Some methods learn correspondences from paired human–robot videos [38, 39], unify visual embeddings of humans and robots [40, 41], use a human video as a guide to retrieve task-relevant behaviors [42, 43], or prompt pretrained policies with retargeted trajectories [44], but these require costly paired data, large teleoperated datasets, or heavy reliance on base policies. Our method learns directly from multiple human demonstrations.

Learning from Sub-Optimal Data. Collecting large amounts of high-quality robot data is prohibitively expensive. As a result, recent work has focused on estimating demonstration quality via costly online interactions [45, 46] or proxy loss metrics [47] that often correlate poorly with real-world performance. In generative modeling, prior works have focused on extracting clean signals from noisy or uncurated datasets [11, 15, 16, 48]. Our method builds upon Ambient Diffusion [19–23], a method for training diffusion models on low-quality data to produce high-quality samples. Its core

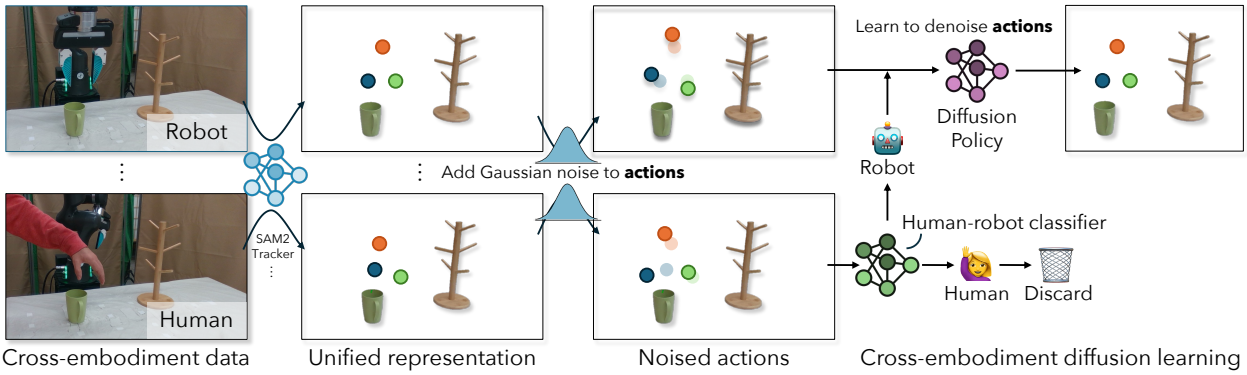

Fig. 2: **Pipeline:** X-DIFFUSION first unifies the state and action representation. State is represented by a colored segmentation mask of relevant objects using Grounded SAM 2 [36, 37]. Action is represented via end-effector/human hand pose utilizing HaMeR [6] for retargeting. To determine if the policy should learn to denoise noisy human 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.

principle is to incorporate low-quality samples into training only when they have been sufficiently noised in the diffusion process. This enables the diffusion model to learn from large amounts of low-quality data without degrading its outputs. Applying this to cross-embodiment robot learning, we treat dynamically infeasible demonstrations as low-quality data, exploiting Ambient Diffusion to adaptively extract useful guidance from uncurated human demonstrations.

III. Problem Formulation and Background

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 $\xi = \{s_t, a_t\}_{t=1}^{T}$.

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 is the behavior cloning loss. This assumes human and robot data have interchangeable dynamics, 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.

Ambient Diffusion. Ambient Diffusion [17, 19–22] is a recent method that trains diffusion models on low-quality data under sufficient noise. Their key insight is that high- and low-quality distributions p_{high} and p_{low} are close (ϵ-merged [17]) after k steps in the forward diffusion process if $D_{KL}\left(p^k_{\text{low}} \,\|\, p^k_{\text{high}}\right) \leq \epsilon$, enabling the use of low-quality data in high-noise regimes. We connect this idea to robot policy learning: when training Diffusion Policies [1], we view human and robot demonstrations as low- and high-quality samples, respectively, learning from noised human actions only when they match the robot's dynamics.

Unifying State and Action Spaces. Following prior work [3, 4], we unify the cross-embodiment data into a shared state $s_t = (q_t, o_t)$ and action $a_t = q_{t+1}$. The proprioception $q_t \in \mathbb{R}^7$ contains the end-effector 3D position, rotation, and gripper state. For human data, we assume access to the following: (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 each view and triangulate them to the 3D 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 [36, 37] and overlay a keypoint rendering of the end-effector pose on each frame, as depicted in Fig. 2. The policy input concatenates this masked image with the proprioceptive information.

IV. Approach

Naive co-training on human and robot demonstrations can degrade performance when execution styles are mismatched. In this section, we present X-DIFFUSION, a cross-embodiment learning framework based on Ambient Diffusion [19] to maximally utilize cross-embodiment data for Diffusion Policy learning without degrading performance. X-DIFFUSION first 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 dynamically infeasible robot actions.

A. Cross-Embodiment Equivalence under Noise

Due to embodiment differences, kinematic retargeting of human hand actions may result in physically infeasible robot motion. Still, human demonstrations provide rich cues for

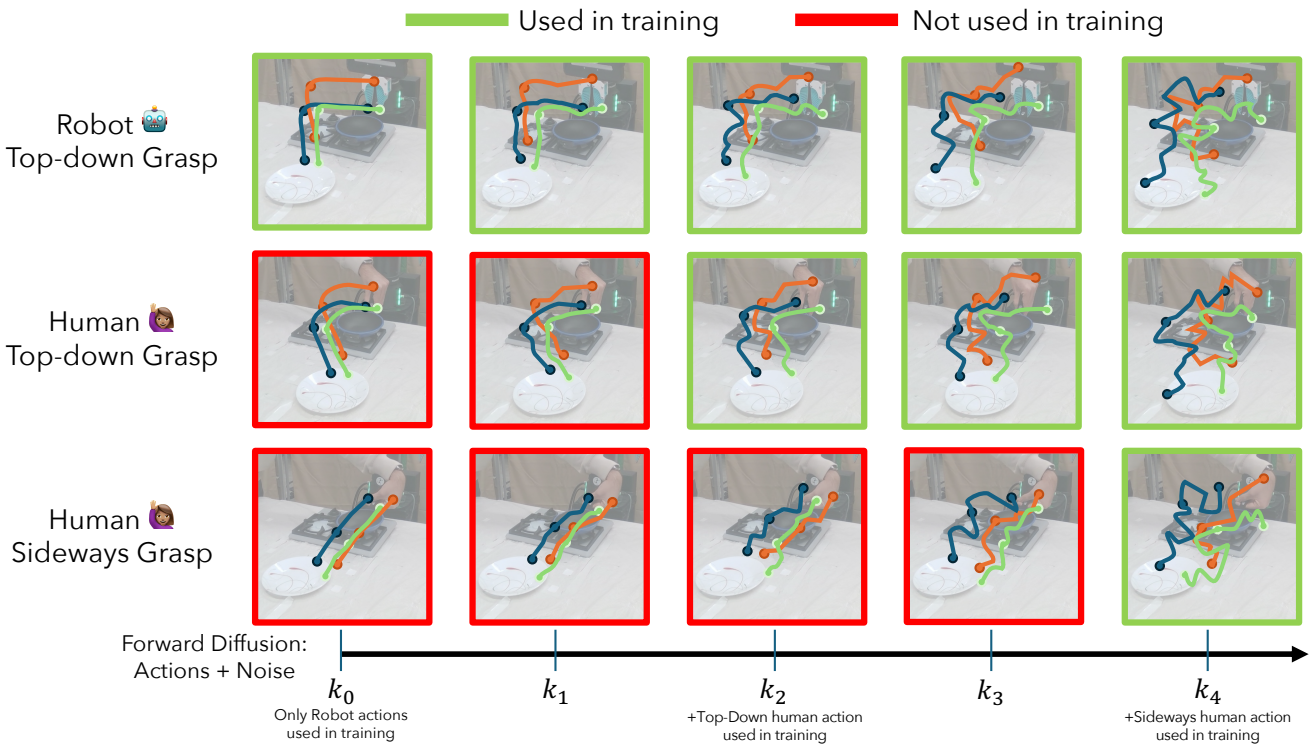

Fig. 3: **Visualizing Actions under Noise and Classifier Predictions at various Diffusion Steps**. Humans execute tasks in various ways. For example, when picking and placing a pan, a human can either execute a top-down grasp or a side grasp. Human actions that are feasible for robots (e.g. top-down 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.

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 clean 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 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 at diffusion step k. Similar to the ϵ-merging time in Ambient Proteins [17], 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:

$$k^\star = \min\Big\{k \mid D_{KL}\big(p_H^k \parallel p_R^k\big) \le \epsilon\Big\},$$

where ϵ is a small threshold. Intuitively, $k^\star$ identifies the point in the noising process at which human actions are sufficiently abstracted to resemble robot actions. Beyond this step ($k \ge k^\star$), human demonstrations can safely supervise robot policy learning without the transfer of infeasible motions.

B. Training a Noised Human-Robot Action Classifier

To determine the minimum indistinguishability timestep k^* for each action, we train a classifier that predicts the embodiment of a noised action. This idea is closely related to the classifier used in Ambient Diffusion Omni [20] to distinguish between low- and high-quality data. The classifier $c_\theta(\cdot|k, \mathbf{A_t^k}, s_t)$ takes in the diffusion step k, the noised action sequence $\mathbf{A}_t^{(k)}$, and the current state s_t, and outputs the probability of the action originating from the robot ($y = 1$) rather than a human ($y = 0$). Training samples are drawn from both the human dataset $\mathcal{D}_H$ and robot dataset $\mathcal{D}_R$. Since the human dataset is much larger than the robot dataset $|\mathcal{D}_H| \gg |\mathcal{D}_R|$, we sample actions from each with equal probability to avoid biasing toward the human label. The classifier is optimized with the binary cross-entropy loss:

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

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

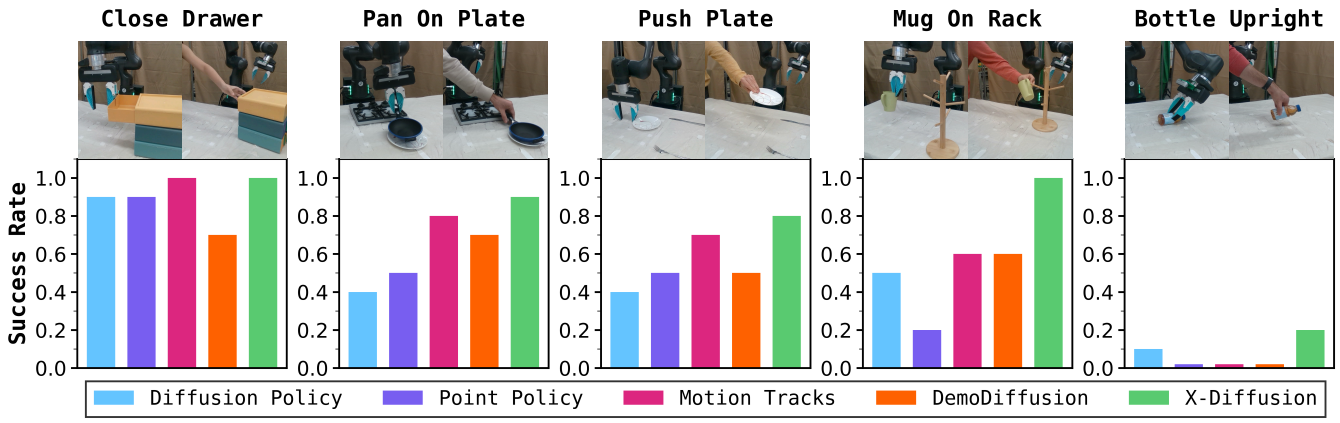

Fig. 4: **Performance vs. Baselines:** We report *task success rate* on 5 different manipulation tasks and compare X-DIFFUSION against a robot-only baseline (Diffusion Policy [1]) and various co-training baselines (Point-Policy [4], Motion Tracks [3]). DemoDiffusion [44] is another diffusion-based method, 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.

step $k^\star$ as the earliest diffusion step where the classifier assigns at least 50% probability to it being a robot action:

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

C. Classifier Integration into Diffusion Policy

Diffusion Policies model the reverse process of denoising. 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 until recovering the clean action sequence $\mathbf{A}_t^0$. 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 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 `Pan On Plate` task for different human actions. Actions that are kinematically feasible for the robot have low k^* whereas infeasible actions have higher k^*. Formally, our Diffusion Policy loss is:

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

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

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: `Close Drawer` (closing a cabinet's top drawer), `Pan On Plate` (picking a frying pan from a stovetop and placing it on a plate), `Push Plate` (sliding a plate between a fork and knife), `Mug On Rack` (inserting a mug's handle onto a rack peg), and `Bottle Upright` (reorienting a bottle to stand upright). These tasks span a wide range of manipulation skills and provide a comprehensive benchmark for assessing the value of human data in policy training. 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, lacking guidance from human data.
2) **Point Policy [4]**: This method co-trains a Diffusion Policy on all human and robot data. Its state is object keypoints from DIFT [49] and Co-Tracker [50] plus hand keypoints.
3) **Motion Tracks [3]**: This method co-trains a Diffusion Policy on all human and robot data. It unifies the action space as hand keypoints but uses raw image observations.
4) **DemoDiffusion [44]**: This method performs the reverse diffusion process using a human policy for the first 60% of steps and a robot policy for the remaining 40%.

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 baselines. We find that X-DIFFUSION achieves higher success rates across tasks relative to Point Policy, Motion Tracks, and DemoDiffusion (Fig. 4). Naively

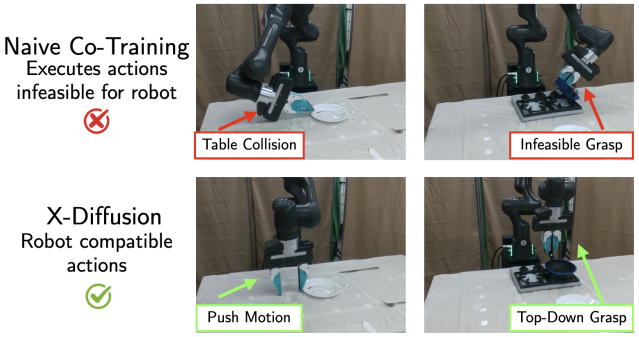

Fig. 5: **Naive Co-Training Learns Infeasible Robot Actions:** Including all human data in policy training can incentivize policies to learn strategies demonstrated by humans that are infeasible for robots. On multiple tasks, a human may manipulate objects in ways that are not realizable for a robot.

co-training on uncurated 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.

Qualitatively, these baselines share a failure mode: executing human actions that are infeasible for the robot (Fig. 5). In `Push Plate` and `Pan On Plate`, several human demonstrations grasp objects from the side (instead of top-down), a kinematically infeasible strategy for the robot.

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

B. Systematic Ablation of Co-Training Data Choices

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. We observe that while nearly all human demonstrations exhibit some degree of mismatch, approximately 50% of the original demonstrations resulted in kinematic or dynamic failures and were discarded. We train three policies with the same architecture but vary the data:

- **ROBOT ONLY:** 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.

Figure 6 shows that FILTERED dataset co-training outperforms NAIVE co-training, confirming the hypothesis that training on infeasible human demonstrations 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

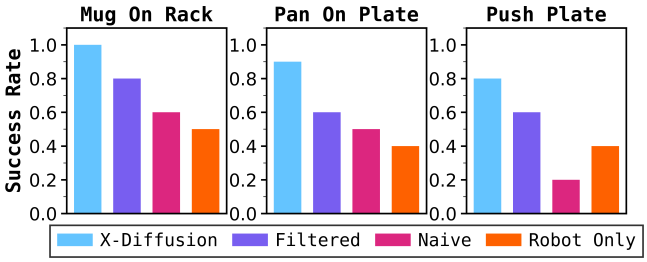

Fig. 6: **Performance vs. Human Data Usage:** We compare X-DIFFUSION with a policy co-trained on data verified as robot-feasible (FILTERED), a naively co-trained policy using all available human data (NAIVE), and policy trained only on robot data (ROBOT ONLY). X-DIFFUSION consistently outperforms all baselines.

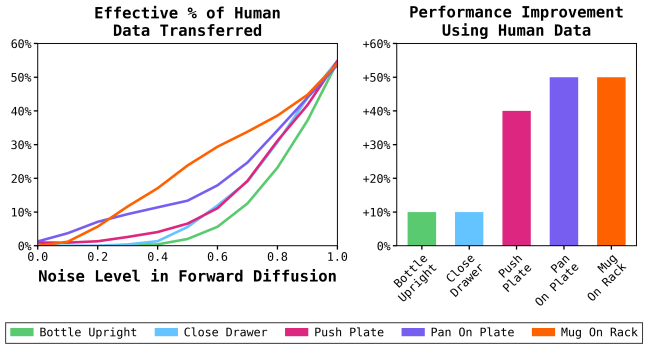

Fig. 7: **Quantifying Transfer Learning from Human Data in X-DIFFUSION: (Left)** For each manipulation task, we measure the fraction of human data incorporated into X-DIFFUSION during training. As the diffusion noise level increases, X-DIFFUSION uses a larger fraction of human data. This fraction varies across tasks; for example, `Mug On Rack` consistently uses a larger fraction of human data than `Bottle Upright`. **(Right)** We measure the performance gain of X-DIFFUSION when trained with human data relative to a baseline trained only on robot data. All tasks benefit from human data, and tasks that incorporate more of it into training, such as `Mug On Rack`, show larger improvements than tasks that use less, such as `Bottle Upright`.

$\mathcal{D}_H$ 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 observe that the minimum indistinguishability step is lower for feasible human actions than their infeasible counterparts. X-DIFFUSION outperforms the FILTERED policy across all tasks, demonstrating the ability to extract signal even from infeasible human demonstrations.

C. Quantifying Transfer Learning from Human Data

A central question in cross-embodiment learning is whether human demonstrations yield *positive transfer* for robot policy learning, i.e., whether adding human data improves performance relative to training on robot data alone. We find that X-DIFFUSION achieves positive transfer by selectively incorporating human data in a task-dependent manner. Figure 7 quantifies the amount of transfer across tasks. On the left, we quantify the fraction of human data incorporated into training across different noise levels in the diffusion process. We show that X-DIFFUSION benefits from transfer learning from human data to varying degrees across all five tasks. `Mug On Rack` and `Pan On Plate`

integrate a larger fraction of human data throughout the diffusion process. `Bottle Upright` integrates substantially less data, suggesting that its human demonstrations are less dynamically compatible with robot execution. On the right, we quantify *positive transfer* as the performance gain of X-DIFFUSION with human data relative to a robot-only baseline. Across all tasks, incorporating human data improves performance, and tasks that integrate more human data show larger gains. Together, these results show that the benefit of transfer learning from human data is task-dependent. Higher performance gains are observed when the human demonstrations are more aligned with the dynamics of robot execution.

Importantly, the transfer achieved by X-DIFFUSION is consistently *positive*. In contrast, Fig. 4 shows that prior cross-embodiment baselines often suffer from negative transfer and can perform worse than training on robot data alone. Fig. 6 provides a more systematic ablation by varying different choices of the data used to train X-DIFFUSION. This shows that the benefit of human supervision depends critically on selecting demonstrations that are truly transferable to the robot. Positive transfer does not arise simply from indiscriminately adding more data, but from selectively incorporating dynamically feasible human actions.

VI. DISCUSSION

In this paper, we propose X-DIFFUSION, a cross-embodiment learning framework for co-training robot policies on human and robot data. Our key idea is to view dynamically infeasible cross-embodiment demonstrations as an analog to low-quality data and leverage recent advances in learning from noisy data [19–23] to effectively integrate them into diffusion policy learning. X-DIFFUSION trains a classifier to identify the minimum noise level where a human action becomes indistinguishable from a robot action, incorporating human actions into training only when they are noised beyond this threshold. This provides coarse task guidance while avoiding the transfer of physically infeasible behaviors. This selective co-training enables effective use of human datasets for robot policy learning, allowing X-DIFFUSION to consistently outperform robot-only policies and prior co-training baselines across five manipulation tasks.

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 policies 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 *et al.*, "Diffusion policy: Visuomotor policy learning via action diffusion," *Int. J. Robot. Res.*, 2024.
[2] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *RSS*, 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," in *ICRA*, 2025.
[4] S. Haldar and L. Pinto, "Point Policy: Unifying observations and actions with key points for robot manipulation," in *CoRL*, 2025.
[5] M. Lepert, J. Fang, and J. Bohg, "Phantom: Training robots without robots using only human videos," in *CoRL*, 2025.
[6] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik, "Reconstructing hands in 3D with transformers," in *CVPR*, 2024.
[7] T. Tao, M. K. Srirama, J. J. Liu, K. Shaw, and D. Pathak, "DexWild: Dexterous human interactions for in-the-wild robot policies," in *RSS*, 2025.
[8] V. Liu *et al.*, "EgoZero: Robot learning from smart glasses," 2025, *arXiv:2505.20290*.
[9] M. Lepert, J. Fang, and J. Bohg, "Masquerade: Learning from in-the-wild human videos using data-editing," in *ICRA*, 2026, to be published.
[10] J. Shi *et al.*, "ZeroMimic: Distilling robotic manipulation skills from web videos," in *ICRA*, 2025.
[11] C. Zhou *et al.*, "LIMA: Less is more for alignment," in *NeurIPS*, 2023.
[12] M. Xia, S. Malladi, S. Gururangan, S. Arora, and D. Chen, "LESS: Selecting influential data for targeted instruction tuning," in *ICML*, 2024.
[13] G. Wang, S. Cheng, X. Zhan, X. Li, S. Song, and Y. Liu, "OpenChat: Advancing open-source language models with mixed-quality data," in *ICLR*, 2024.
[14] M. Li *et al.*, "Superfiltering: Weak-to-strong data filtering for fast instruction-tuning," in *ACL*, 2024.
[15] A. Bora, E. Price, and A. G. Dimakis, "AmbientGAN: Generative models from lossy measurements," in *ICLR*, 2018.
[16] J. Lehtinen *et al.*, "Noise2Noise: Learning image restoration without clean data," in *ICML*, 2018.
[17] G. Daras, J. Ouyang-Zhang, K. Ravishankar, C. C. Daskalakis, A. Klivans, and D. J. Diaz, "Ambient proteins - training diffusion models on noisy structures," in *NeurIPS*, 2025.
[18] S. Kondylatos, N. I. Bountos, I. Prapas, A. Zavras, G. Camps-Valls, and I. Papoutsis, "Probabilistic machine learning for noisy labels in Earth observation," *Sci. Rep.*, vol. 15, no. 1, 2025.
[19] G. Daras, K. Shah, Y. Dagan, A. Gollakota, A. Dimakis, and A. Klivans, "Ambient diffusion: Learning clean distributions from corrupted data," in *NeurIPS*, 2023.
[20] G. Daras, A. Rodriguez-Munoz, A. Klivans, A. Tor-

ralba, and C. C. Daskalakis, "Ambient diffusion omni: Training good models with bad data," in *NeurIPS*, 2025.
[21] G. Daras, Y. Dagan, A. Dimakis, and C. C. Daskalakis, "Consistent diffusion models: Mitigating sampling drift by learning to be consistent," in *NeurIPS*, 2023.
[22] G. Daras, A. G. Dimakis, and C. Daskalakis, "Consistent diffusion meets Tweedie: Training exact ambient diffusion models with noisy data," in *ICML*, 2024.
[23] G. Daras, Y. Cherapanamjeri, and C. C. Daskalakis, "How much is a noisy image worth? Data scaling laws for Ambient Diffusion," in *ICLR*, 2025.
[24] H. Bharadhwaj, A. Gupta, S. Tulsiani, and V. Kumar, "Zero-shot robot manipulation from passive human videos," 2023, *arXiv:2302.02011*.
[25] C. Wang *et al.*, "MimicPlay: Long-horizon imitation learning by watching human play," in *CoRL*, 2023.
[26] M. Lepert, R. Doshi, and J. Bohg, "Shadow: Leveraging segmentation masks for zero-shot cross-embodiment policy transfer," in *CoRL*, 2024.
[27] S. Bahl, A. Gupta, and D. Pathak, "Human-to-robot imitation in the wild," in *RSS*, 2022.
[28] Y. Zhu, A. Lim, P. Stone, and Y. Zhu, "Vision-based manipulation from single human video with open-world object graphs," 2024, *arXiv:2405.20321*.
[29] P. Vitiello, K. Dreczkowski, and E. Johns, "One-shot imitation learning: A pose estimation perspective," in *CoRL*, 2023.
[30] J. Li *et al.*, "OKAMI: Teaching humanoid robots manipulation skills through single video imitation," in *CoRL*, 2024.
[31] 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, 2018.
[32] Z. Yuan *et al.*, "HERMES: Human-to-robot embodied learning from multi-source motion data for mobile dexterous manipulation," 2025, *arXiv:2508.20085*.
[33] P. Dan *et al.*, "X-Sim: Cross-embodiment learning via real-to-sim-to-real," in *CoRL*, 2025.
[34] T. G. W. Lum, O. Y. Lee, C. K. Liu, and J. Bohg, "Crossing the human-robot embodiment gap with sim-to-real RL using one human demonstration," in *CoRL*, 2025.
[35] K. Schmeckpeper *et al.*, "Learning predictive models from observation and interaction," in *ECCV*, 2020.
[36] N. Ravi *et al.*, "SAM 2: Segment anything in images and videos," 2024, *arXiv:2408.00714*.
[37] T. Ren *et al.*, "Grounded SAM: Assembling open-world models for diverse visual tasks," 2024, *arXiv:2401.14159*.
[38] E. Jang *et al.*, "BC-z: Zero-shot task generalization with robotic imitation learning," in *CoRL*, 2021.
[39] V. Jain *et al.*, "Vid2Robot: End-to-end video conditioned policy learning with cross-attention transformers," in *RSS*, 2024.
[40] K. Kedia, P. Dan, A. Chao, M. A. Pace, and S. Choudhury, "One-shot imitation under mismatched execution," in *ICRA*, 2025.
[41] M. Xu, Z. Xu, C. Chi, M. Veloso, and S. Song, "XSkill: Cross embodiment skill discovery," in *CoRL*, 2023.
[42] R. Shah *et al.*, "MimicDroid: In-context learning for humanoid manipulation from human play videos," in *ICRA*, 2026, to be published.
[43] V. Vosylius and E. Johns, "Instant policy: In-context imitation learning via graph diffusion," in *ICLR*, 2025.
[44] S. Park, H. Bharadhwaj, and S. Tulsiani, "DemoDiffusion: One-shot human imitation using pre-trained diffusion policy," in *ICRA*, 2026, to be published.
[45] A. S. Chen, A. M. Lessing, Y. Liu, and C. Finn, "Curating demonstrations using online experience," in *RSS*, 2025.
[46] C. Agia *et al.*, "CUPID: Curating data your robot loves with influence functions," in *CoRL*, 2025.
[47] J. Hejna, C. A. Bhateja, Y. Jiang, K. Pertsch, and D. Sadigh, "ReMix: Optimizing data mixtures for large scale imitation learning," in *CoRL*, 2024.
[48] X. Dai *et al.*, "Emu: Enhancing image generation models using photogenic needles in a haystack," 2023, *arXiv:2309.15807*.
[49] L. Tang, M. Jia, Q. Wang, C. P. Phoo, and B. Hariharan, "Emergent correspondence from image diffusion," in *NeurIPS*, 2023.
[50] N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, and C. Rupprecht, "CoTracker: It is better to track together," in *ECCV*, 2024.