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
portal-cornell.github.io/X-Diffusion

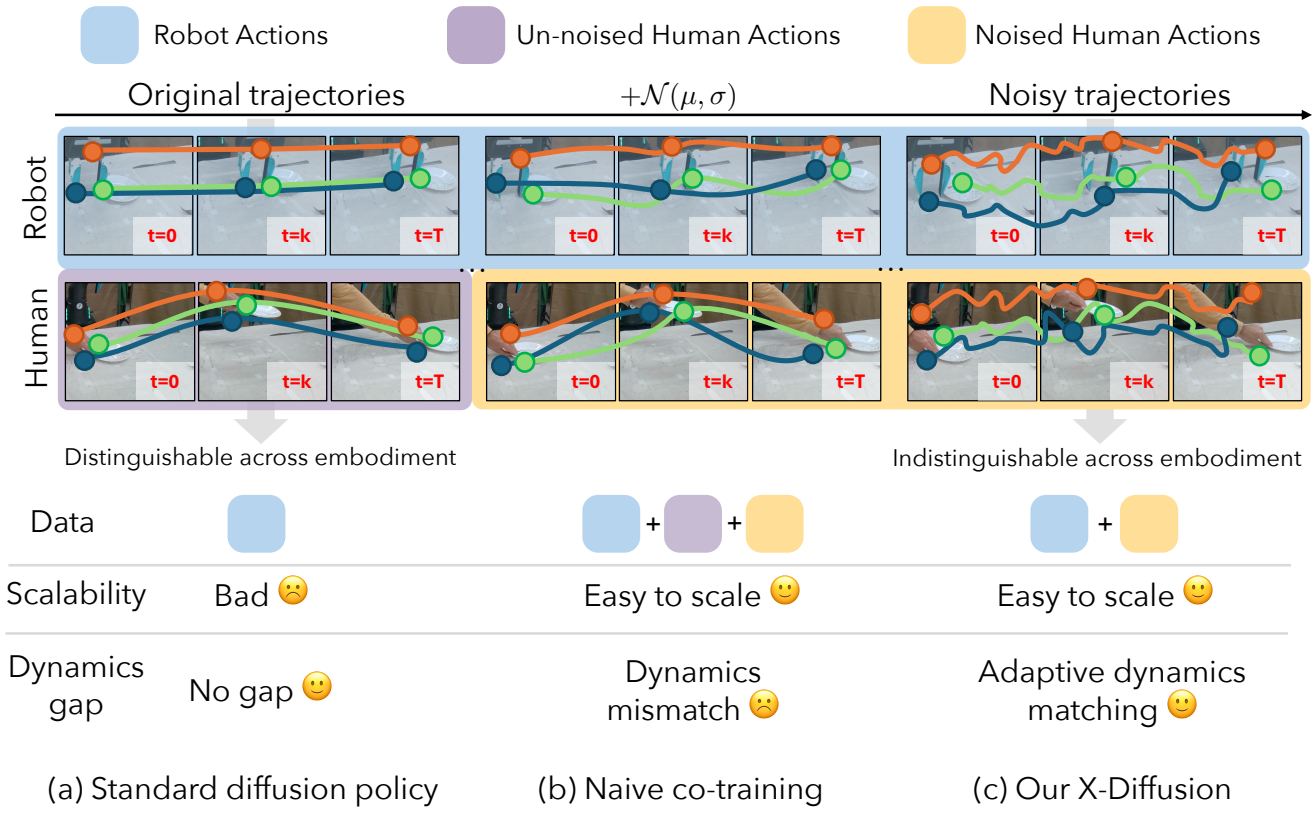

Fig. 1: **Overview of X-DIFFUSION:** We introduce X-DIFFUSION, a cross-embodiment learning framework that trains Diffusion Policies on human demonstrations even when their actions are not directly executable by the robot. Prior methods typically co-train on mixed human and robot datasets, which often 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 [7, 8]. 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 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. Should we train on all human data indiscriminately, or should misaligned demonstrations be identified and discarded to prevent degrading policy performance?

Similar challenges exist in generative modeling, where naively training on a mixture of low-quality and high-quality data often degrades model performance. Ambient Diffusion [9, 10] offers an exciting alternative by strategically integrating low-quality data only at higher-noise timesteps of diffusion. Building on this idea, 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. When sufficient noise is applied to both human and robot actions in the forward diffusion process, low-level embodiment differences fade away while the underlying task structure is preserved.

Selectively training Diffusion Policies on noised human actions thus 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 signal from all human data while avoiding infeasible motion.

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. APPROACH

We present X-DIFFUSION, a cross-embodiment learning framework based on Ambient Diffusion [9] that maximally utilizes cross-embodiment data for Diffusion Policy learning without degrading performance. X-DIFFUSION first trains a classifier to distinguish between noised human and robot actions, then integrates noised actions into policy training only when the classifier is confused about their embodiment.

Cross-Embodiment Equivalence under Noise. Diffusion Policies [1] learn by denoising action sequences corrupted with Gaussian noise. Given a clean action sequence $\mathbf{A_t^0}$ from either the human dataset $\mathcal{D}_H$ or robot dataset $\mathcal{D}_R$, the forward diffusion process produces progressively noisier versions via $q(\mathbf{A_t^{k+1}} \mid \mathbf{A_t^k}) = \mathcal{N}\big(\sqrt{1-\beta_k}\,\mathbf{A}_t^k,\ \beta_k I\big)$. Our key observation is that forward diffusion progressively removes embodiment-specific features: at high noise levels, human and robot trajectories become indistinguishable (Fig. 1). Letting p_H^k and p_R^k denote the distributions of human and robot actions at step k, we define the **minimum indistinguishability step** $k^\star$ as the earliest step where $D_{KL}(p_H^k \,\|\, p_R^k) \le \epsilon$ for small ϵ. Beyond this step, human demonstrations can safely supervise robot policy learning without transferring infeasible motions.

Noised Human-Robot Action Classifier. To estimate $k^\star$ per action, we train a classifier $c_\theta(k, \mathbf{A_t^k}, s_t)$ that predicts whether a noised action sequence originates from the robot ($y = 1$) or a human ($y = 0$), given the current state s_t. We sample actions from $\mathcal{D}_R$ and $\mathcal{D}_H$ with equal probability and optimize a binary cross-entropy loss:

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

For each human action, we define its minimum indistinguishability step as $k^\star(\mathbf{A}_t) = \min\{k : c_\theta(k, \mathbf{A}_t^k, s_t) \ge 0.5\}$, the earliest step at which the classifier is equally likely to label it as robot or human.

Classifier Integration into Diffusion Policy. During training, we only incur the action denoising loss on human data when $k \ge k^\star(\mathbf{A}_t)$, restricting mismatched human actions to provide only coarse guidance at high noise:

$$\begin{aligned}\mathcal{L}_{\text{X-DP}}(\theta) = \ & \mathbb{E}_{(k,\mathbf{A}_t,s_t)\sim\mathcal{D}_R}\ \ell\big(p_\theta, \mathbf{A}_t^k\big) \\ & + \mathbb{E}_{(k,\mathbf{A}_t,s_t)\sim\mathcal{D}_H}\ \mathbf{1}_{\{k\ge k^\star(\mathbf{A}_t)\}}\,\ell\big(p_\theta, \mathbf{A}_t^k\big)\,,\end{aligned} \tag{2}$$

where ℓ is the standard denoising loss. Fig. 2 visualizes $k^\star$ on `Pan On Plate`: kinematically feasible human actions have low $k^\star$ whereas infeasible actions have higher $k^\star$. This selective integration maximally utilizes human demonstrations without sacrificing kinematic feasibility.

Unifying state and action spaces. Following prior work [3, 4], we extract 3D hand keypoints via HaMeR [6]

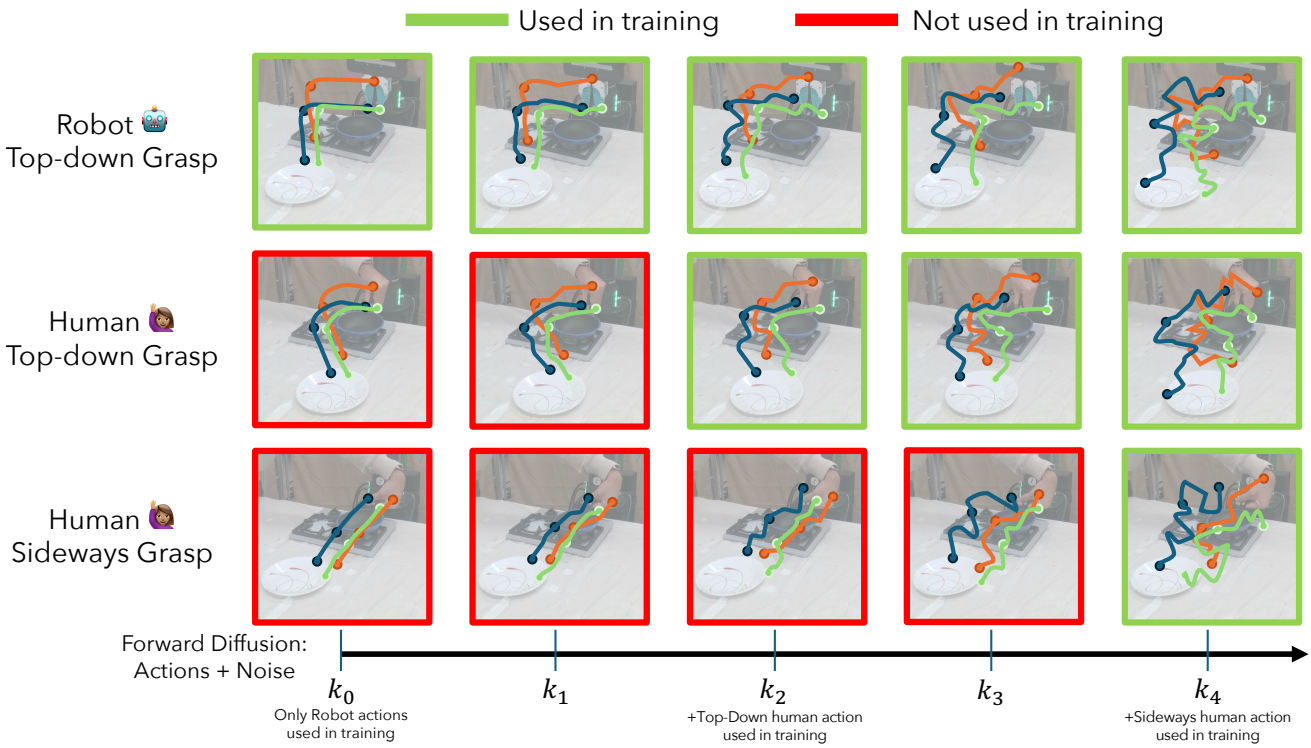

Fig. 2: **Visualizing Actions under Noise and Classifier Predictions at various Diffusion Steps**. Humans execute tasks in various ways. For example, when picking and placing a pan, a human can either execute a top-down grasp or a side grasp. Human actions that are feasible for robots (e.g. top-down 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.

from two calibrated RGB cameras, derive end-effector position, orientation, and a binary gripper status, and reduce the human–robot visual gap by segmenting task-relevant objects with Grounded-SAM 2 [11] and overlaying end-effector keypoints on the images.

III. Experiments

Experimental Setup. For each task, we collect 5 robot demonstrations and 100 human demonstrations. Human demonstrations are performed with a single hand; the robot is a 7-DOF Franka Emika Panda arm. We evaluate on: `Close Drawer` (closing a cabinet's top drawer), `Pan On Plate` (picking a frying pan 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). We report average success rate over 10 rollouts per task.

Baselines. We compare against: (1) **Diffusion Policy** [1]: trains only on 5 robot demonstrations. (2) **Point Policy** [4]: co-trains a Diffusion Policy on all human and robot data using object keypoints from DIFT [13] and Co-Tracker [14] plus hand keypoints. (3) **Motion Tracks** [3]: co-trains a Diffusion Policy on all human and robot data, using hand keypoints over raw images. (4) **DemoDiffusion** [12]: runs reverse diffusion with a human policy for the first 60% of steps and a robot policy for the remainder.

A. Comparison with Cross-Embodiment Baselines

X-Diffusion achieves higher success rates across tasks relative to Point Policy, Motion Tracks, and DemoDiffusion (Fig. 3). Naively co-training on uncurated human demonstrations yields little to no improvement (Motion Tracks, DemoDiffusion) and can even degrade performance (Point Policy) by learning suboptimal robot behaviors.

Qualitatively, these baselines share a failure mode: executing human actions infeasible for the robot (Fig. 4). 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. In contrast, X-Diffusion leverages its classifier to apply the denoising loss on human motions indistinguishable from robot motion.

B. Systematic Ablation of Co-Training Data Choices

To investigate the role of the human data distribution, we construct $\mathcal{D}_H^+$ by replaying each human demonstration on the robot via Inverse Kinematics (IK) and manually filtering out unsuccessful trajectories. Nearly all human demonstrations exhibit some mismatch, and approximately 50% result in kinematic or dynamic failures and are discarded. We train:

- Robot Only: trained 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 $k^\star$ during action denoising (Sec. II).

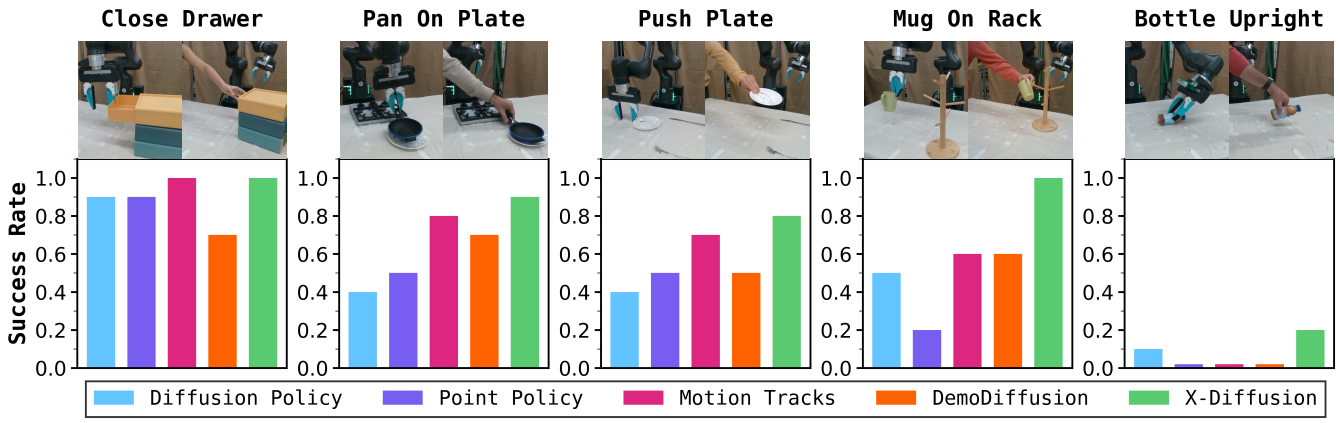

Fig. 3: **Performance vs. Baselines:** We report *task success rate* on 5 different manipulation tasks and compare X-DIFFUSION against a robot-only baseline (Diffusion Policy [1]) and various co-training baselines (Point-Policy [4], Motion Tracks [3]). DemoDiffusion [12] is another diffusion-based method, 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.

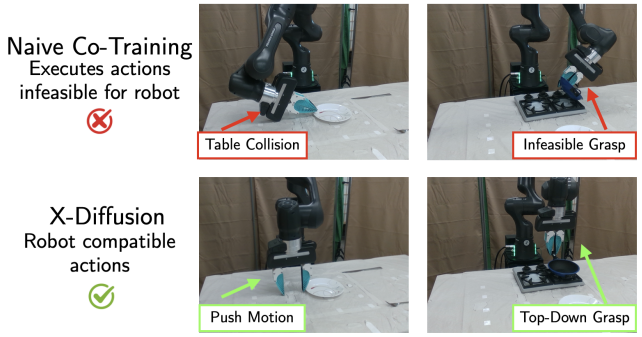

Fig. 4: **Naive Co-Training Learns Infeasible Robot Actions:** Including all human data in policy training can incentivize policies to learn strategies demonstrated by humans that are infeasible for robots. On multiple tasks, a human may manipulate objects in ways that are not realizable for a robot.

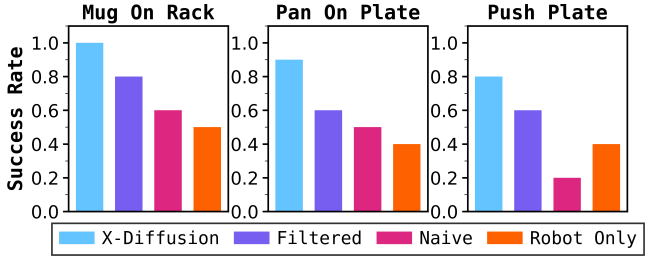

Fig. 5: **Performance vs. Human Data Usage:** We compare X-DIFFUSION with a policy co-trained on data verified as robot-feasible (FILTERED), a naively co-trained policy using all available human data (NAIVE), and policy trained only on robot data (ROBOT ONLY). X-DIFFUSION consistently outperforms all baselines.

Fig. 5 shows that FILTERED outperforms NAIVE, confirming that training on infeasible human demonstrations degrades performance. X-DIFFUSION takes an alternate approach: rather than discarding entire trajectories, it adaptively includes human data from $\mathcal{D}_H$ only beyond the noise level where human and robot distributions become indistinguishable, learning to denoise within the correct distribution for the robot. As shown in Fig. 2, $k^\star$ is lower for feasible human actions than for their infeasible counterparts. X-DIFFUSION

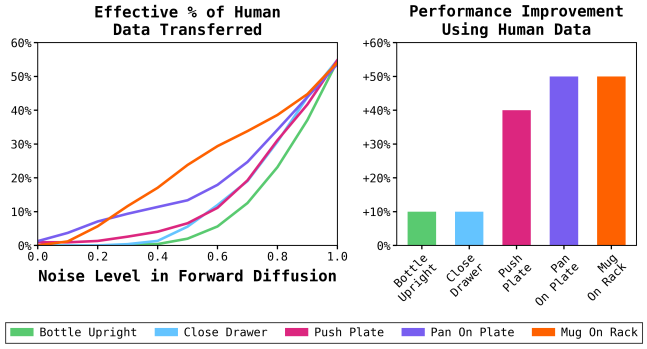

Fig. 6: **Quantifying Transfer Learning from Human Data in X-DIFFUSION: (Left)** For each manipulation task, we measure the fraction of human data incorporated into X-DIFFUSION during training. As the diffusion noise level increases, X-DIFFUSION uses a larger fraction of human data. This fraction varies 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`.

outperforms FILTERED across all tasks, demonstrating the ability to extract signal from infeasible human motion.

C. Quantifying Transfer Learning from Human Data

A central question in cross-embodiment learning is whether human demonstrations yield *positive transfer*—whether adding human data improves over robot-only training. X-DIFFUSION achieves positive transfer by selectively incorporating human data in a task-dependent manner. Fig. 6 (left) quantifies the fraction of human data incorporated into training across diffusion noise levels. Fig. 6 (right) reports the gain over a robot-only baseline: all tasks benefit, and tasks that integrate more human data show larger gains. In contrast, prior baselines often suffer from *negative* transfer (Fig. 3), showing that positive transfer requires selectively incorporating dynamically feasible human actions rather than simply adding more data.

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] G. Daras, K. Shah, Y. Dagan, A. Gollakota, A. Dimakis, and A. Klivans, "Ambient diffusion: Learning clean distributions from corrupted data," in *NeurIPS*, 2023.

[10] G. Daras, A. Rodriguez-Munoz, A. Klivans, A. Torralba, and C. C. Daskalakis, "Ambient diffusion omni: Training good models with bad data," in *NeurIPS*, 2025.

[11] T. Ren *et al.*, "Grounded SAM: Assembling open-world models for diverse visual tasks," 2024, *arXiv:2401.14159*.

[12] S. Park, H. Bharadhwaj, and S. Tulsiani, "DemoDiffusion: One-shot human imitation using pre-trained diffusion policy," in *ICRA*, 2026, to be published.

[13] L. Tang, M. Jia, Q. Wang, C. P. Phoo, and B. Hariharan, "Emergent correspondence from image diffusion," in *NeurIPS*, 2023.

[14] N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, and C. Rupprecht, "CoTracker: It is better to track together," in *ECCV*, 2024.

[15] H. Bharadhwaj, A. Gupta, S. Tulsiani, and V. Kumar, "Zero-shot robot manipulation from passive human videos," 2023, *arXiv:2302.02011*.

[16] C. Wang *et al.*, "MimicPlay: Long-horizon imitation learning by watching human play," in *CoRL*, 2023.

[17] M. Lepert, R. Doshi, and J. Bohg, "Shadow: Leveraging segmentation masks for zero-shot cross-embodiment policy transfer," in *CoRL*, 2024.

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

[19] 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*.

[20] P. Vitiello, K. Dreczkowski, and E. Johns, "One-shot imitation learning: A pose estimation perspective," in *CoRL*, 2023.

[21] J. Li *et al.*, "OKAMI: Teaching humanoid robots manipulation skills through single video imitation," in *CoRL*, 2024.

[22] 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.

[23] Z. Yuan *et al.*, "HERMES: Human-to-robot embodied learning from multi-source motion data for mobile dexterous manipulation," 2025, *arXiv:2508.20085*.

[24] P. Dan *et al.*, "X-Sim: Cross-embodiment learning via real-to-sim-to-real," in *CoRL*, 2025.

[25] 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.

[26] K. Schmeckpeper *et al.*, "Learning predictive models from observation and interaction," in *ECCV*, 2020.

[27] E. Jang *et al.*, "BC-z: Zero-shot task generalization with robotic imitation learning," in *CoRL*, 2021.

[28] V. Jain *et al.*, "Vid2Robot: End-to-end video conditioned policy learning with cross-attention transformers," in *RSS*, 2024.

[29] K. Kedia, P. Dan, A. Chao, M. A. Pace, and S. Choudhury, "One-shot imitation under mismatched execution," in *ICRA*, 2025.

[30] M. Xu, Z. Xu, C. Chi, M. Veloso, and S. Song, "XSkill: Cross embodiment skill discovery," in *CoRL*, 2023.

[31] R. Shah *et al.*, "MimicDroid: In-context learning for humanoid manipulation from human play videos," in *ICRA*, 2026, to be published.

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

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

[34] C. Agia *et al.*, "CUPID: Curating data your robot loves with influence functions," in *CoRL*, 2025.

[35] J. Hejna, C. A. Bhateja, Y. Jiang, K. Pertsch, and D. Sadigh, "ReMix: Optimizing data mixtures for large scale imitation learning," in *CoRL*, 2024.

[36] C. Zhou *et al.*, "LIMA: Less is more for alignment," in *NeurIPS*, 2023.

[37] J. Lehtinen *et al.*, "Noise2Noise: Learning image restoration without clean data," in *ICML*, 2018.

[38] A. Bora, E. Price, and A. G. Dimakis, "AmbientGAN: Generative models from lossy measurements," in *ICLR*, 2018.

[39] X. Dai *et al.*, "Emu: Enhancing image generation models using photogenic needles in a haystack," 2023, *arXiv:2309.15807*.

[40] G. Daras, Y. Dagan, A. Dimakis, and C. C. Daskalakis, "Consistent diffusion models: Mitigating sampling drift by learning to be consistent," in *NeurIPS*, 2023.

[41] G. Daras, A. G. Dimakis, and C. Daskalakis, "Consis-

tent diffusion meets Tweedie: Training exact ambient diffusion models with noisy data," in *ICML*, 2024.
[42] G. Daras, Y. Cherapanamjeri, and C. C. Daskalakis, "How much is a noisy image worth? Data scaling laws for Ambient Diffusion," in *ICLR*, 2025.
[43] N. Ravi *et al.*, "SAM 2: Segment anything in images and videos," 2024, *arXiv:2408.00714*.
[44] 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.

APPENDIX

A. RELATED WORKS

Learning from Human Hand Motion. Advances in hand-pose estimation have enabled retargeting actionless human videos into robot actions. One approach tracks 6DoF hand trajectories and maps them to the robot end-effector [15, 16]. Other works define corresponding keypoints between humans and robots to unify their data representations [3, 4], or overlay rendered robot arms on human videos [5, 17, 18]. Open-world vision models have further enabled object-aware retargeting [19–21]. These methods assume that retargeted hand motions will transfer cleanly to the robot, which 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 [22, 23], object-centric signals in real-to-sim-to-real pipelines [24, 25], or classifier judgments of task success [26]. 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 [27, 28], unify visual embeddings of humans and robots [29, 30], use a human video as a guide to retrieve task-relevant behaviors [31, 32], or prompt pretrained policies with retargeted trajectories [12]. 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. Recent work has focused on estimating demonstration quality via costly online interactions [33, 34] or proxy loss metrics [35] that often correlate poorly with real-world performance. In generative modeling, prior works have focused on extracting clean signals from noisy or uncurated datasets [36–39]. Our method builds on Ambient Diffusion [9, 10, 40–42], which incorporates low-quality samples into training only at sufficiently high noise, enabling large-scale learning from low-quality data without degrading outputs. We treat dynamically infeasible human demonstrations as low-quality data and adaptively extract guidance from uncurated human demonstrations.

B. BACKGROUND

Our goal is to learn a robot policy $\pi_\theta(\mathbf{A}_t \mid s_t)$ that predicts a sequence of S future actions $\mathbf{A}_t = a_{t:t+S}$ given the current robot state s_t. Training uses 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 contains trajectories $\xi = \{s_t, a_t\}_{t=1}^{T}$.

Co-Training of Robot Policies. A straightforward approach combines both datasets and trains on the aggregated mixture:

$$\mathcal{L}_{\text{co-train}}(\theta) = \mathbb{E}_{(s_t,\mathbf{A}_t)\sim\mathcal{D}_R\cup\mathcal{D}_H}\left[\ell\big(\pi_\theta(s_t), \mathbf{A}_t\big)\right], \quad (3)$$

assuming human and robot data have interchangeable dynamics. Because human actions are often physically infeasible for the robot, naive co-training can significantly degrade policy performance, motivating selective co-training.

Ambient Diffusion. Ambient Diffusion [9, 10, 44] trains diffusion models on low-quality data under sufficient noise. Its key insight is that high- and low-quality distributions p_{high} and p_{low} become ϵ-merged after k forward-diffusion steps if $D_{KL}(p^k_{\text{low}} \| p^k_{\text{high}}) \leq \epsilon$, enabling the use of low-quality data in high-noise regimes. We apply this idea to robot policy learning: we treat human and robot demonstrations as low- and high-quality samples respectively, and learn from noised human actions only when they match the robot's dynamics.

C. STATE AND ACTION REPRESENTATIONS

Our datasets contain trajectories $\xi = \{(s_t, a_t)\}_{t=1}^{T}$ with embodiment-agnostic state $s_t = (q_t, o_t)$. Proprioception $q_t = (p_t, r_t, g_t) \in \mathbb{R}^7$ includes 3D end-effector position p_t, rotation r_t, and gripper status g_t. The visual observation $o_t \in \mathbb{R}^{H\times W\times 3}$ contains 2D RGB segmentations of task-relevant objects with end-effector keypoint renderings overlaid. Actions are the next-step proprioception, $a_t = q_{t+1}$.

Robot Demonstrations. Robot proprioception is computed from joint angles and gripper status using forward kinematics. Visual observations are produced by applying Grounded-SAM 2 [11] with language prompts on a single-view RGB capture and overlaying end-effector keypoint renderings.

Human Demonstrations. We use HaMeR [6] to detect 21 2D keypoints per camera, select 5 keypoints along the index finger and thumb for retargeting to a parallel-jaw gripper, and triangulate across two calibrated cameras to obtain p_t in the robot frame. Rotation r_t is computed via the Kabsch algorithm. Gripper status g_t is thresholded on the distance between the index fingertip and thumb. Visual observations use the same pipeline as robot demonstrations.

D. TASK DESCRIPTIONS

- `Close Drawer`: close the top drawer of a cabinet, whose position and rotation are randomized along a line.

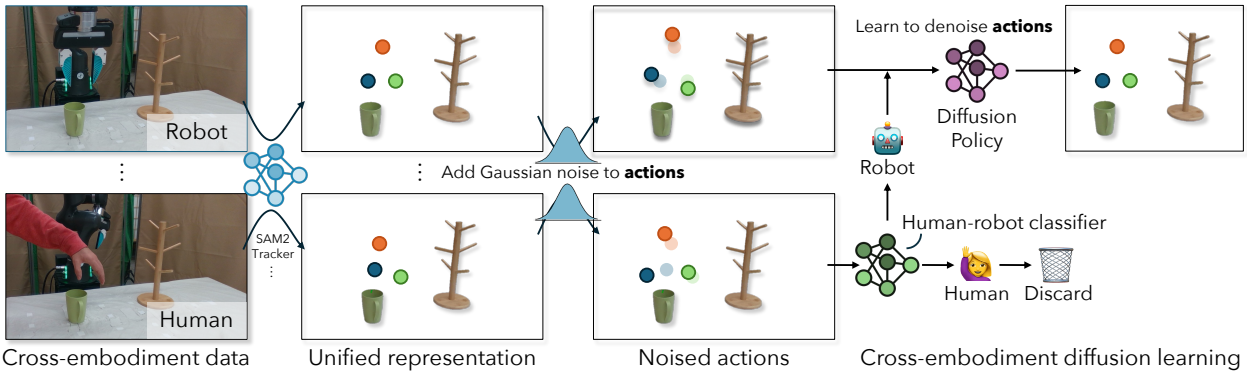

Fig. 7: **Pipeline:** X-DIFFUSION first unifies the state and action representation. State is represented by a colored segmentation mask of relevant objects using Grounded SAM 2 [11, 43]. Action is represented via end-effector/human hand pose utilizing HaMeR [6] for retargeting. To determine if the policy 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.

- `Pan On Plate`: pick up a frying pan from the stovetop and place it on a plate whose location is randomized along a line.
- `Push Plate`: slide a plate between a fork and a knife, whose positions are jointly randomized along the table.
- `Mug On Rack`: pick up a mug and place its handle on a rack peg; the rack position is randomized along a line.
- `Bottle Upright`: pick up a juice bottle lying on its side, reorient it, and release it standing upright.

E. BASELINE DETAILS

All policies use the Diffusion Policy [1] architecture on our unified state representation unless otherwise specified.

Diffusion Policy. Vanilla Diffusion Policy trained only on a small set of robot demonstrations.

Point Policy. Instead of segmented images, this baseline represents state via 3D keypoints of relevant objects. Keypoints are annotated on one training frame, with correspondences auto-detected at the start of other demonstrations and at inference time using DIFT [13]. Co-Tracker [14] tracks each point over time, and 3D object points come from two-camera triangulation. Trained with equal sampling of human and robot demonstrations.

Motion Tracks. Consumes raw RGB images (no segmentation) and end-effector proprioception. The original paper trains a keypoint retargeting network; we remove this by directly unifying the proprioception as end-effector position/rotation. Trained with equal sampling.

DemoDiffusion. Trains separate human policy π^H on $\mathcal{D}_H$ and robot policy π^R on $\mathcal{D}_R$. At inference, the reverse diffusion process uses π^H for the first 60% of denoising steps and π^R for the remaining 40%.

F. DISCUSSION AND LIMITATIONS

X-DIFFUSION is a cross-embodiment learning framework for co-training robot policies on human and robot data, viewing dynamically infeasible cross-embodiment demonstrations as an analog to low-quality data and leveraging advances in learning from noisy data [9, 10, 40–42] to integrate them into Diffusion Policy learning. X-DIFFUSION trains a classifier to identify the minimum noise level at which a human action becomes indistinguishable from a robot action, and incorporates human actions into training only when noised beyond this threshold. This provides coarse task guidance while avoiding the transfer of physically infeasible behaviors. In our experiments, X-DIFFUSION is trained on a limited number of robot and human demonstrations in a calibrated multi-camera environment; future work will scale to larger datasets and unstructured internet-scale human videos.

G. DIFFUSION POLICY HYPERPARAMETERS

TABLE I: 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