

Diffusion Co-Policy for Synergistic Human-Robot Collaborative Tasks

Eley Ng¹, Ziang Liu², and Monroe Kennedy III^{1,2}

Abstract—Modeling multimodal human behavior has been a key barrier to increasing the level of interaction between human and robot, particularly for collaborative tasks. Our key insight is that an effective, learned robot policy used for human-robot collaborative tasks must be able to express a high degree of multimodality, predict actions in a temporally consistent manner, and recognize a wide range of frequencies of human actions in order to seamlessly integrate with a human in the control loop. We present Diffusion Co-policy, a method for planning sequences of actions that synergize well with humans during test time. The co-policy predicts joint human-robot action sequences via a Transformer-based diffusion model, which is trained on a dataset of collaborative human-human demonstrations, and directly executes the robot actions in a receding horizon control framework. We demonstrate in both simulation and real environments that the method outperforms other state-of-art learning methods on the task of human-robot table-carrying with a human in the loop. Moreover, we qualitatively highlight compelling robot behaviors that demonstrate evidence of true human-robot collaboration, including mutual adaptation, shared task understanding, leadership switching, and low levels of wasteful interaction forces arising from dissent.

Index Terms—Human-Robot Collaboration, Deep Learning Methods, Imitation Learning

I. INTRODUCTION

MULTIMODAL behavior poses a key barrier to achieving effective human-robot coordination in collaborative tasks. In collaborative tasks, decentralized agents execute joint actions, which are defined in cognitive neuroscience as “any form of social interaction whereby two or more individuals coordinate their actions in space and time to bring about a change in the environment” [1]. For such scenarios, the ability to anticipate and predict partners’ actions is crucial, as it can significantly enhance the capacity to plan synergistic actions that contribute to the team’s success with an understanding of the task (i.e. how the team’s actions affects the dynamics). Collaborative table carrying, an exemplar of such tasks, demands on-the-fly mutual adaptation and

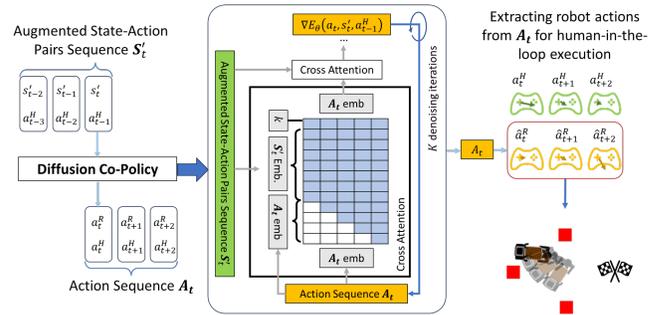


Fig. 1: Framework for human-robot table-carrying with diffusion co-policy. At time step t , the co-policy takes as input the latest T_o steps of augmented state data and past human actions, denoted as the augmented state-action pair sequence S'_t , and outputs T_a steps of joint actions A_t . To generate the output, the Transformer-based decoder architecture denoises a sampled action sequence after K iterations through multiple cross-attention layers. The robot extracts the robot action sequence from A_t and executes it simultaneously with control inputs from a human.

precise timing of movements. This paper introduces a new approach to coordinating and executing behaviors seamlessly, demonstrated through human-robot collaborative carrying.

Given the recent successes of denoising diffusion probabilistic models (DDPM) in learning single agent behaviors [2], [3], [4], we propose leveraging DDPMs in human-robot collaboration, and demonstrate that the gains in model expressivity enabled by DDPM are highly suited for tasks that involve humans. Using human-human collaborative demonstrations, we train a robot co-policy that conditions on past observations and human actions to generate sequences of future joint human and robot actions, and directly execute the robot actions using receding horizon control. We show the effectiveness of the diffusion-based co-policy in both simulation and real robot experiments by highlighting compelling collaborative behaviors exhibited by the robot and human.

Contributions: Our primary contribution is the application of diffusion models for learning robot collaborative policies that can synergize with a real, human partner. Our approach utilizes a Transformer-based diffusion model to predict future joint action sequences conditioned on past observations and human actions, facilitating smooth, coordinated actions that can be executed without further processing or use of hand-engineered collaborative reward functions. We demonstrate on the table-carrying task that this method outperforms state-

Manuscript received: May 20, 2023; Revised: August 16, 2023; Accepted: October 16, 2023.

This paper was recommended for publication by Editor Angelika Peer upon evaluation of the Associate Editor and Reviewers’ comments. The first author is supported by the NSF Graduate Research Fellowship. This work was supported by Stanford Institute for Human-Centered Artificial Intelligence (HAI) and conducted under IRB-65022. Link to project site: <https://sites.google.com/view/diffusion-co-policy-hrc>.

¹Department of Mechanical Engineering, Stanford University, USA. {eleyng, monroek}@stanford.edu.

²Department of Computer Science, Stanford University, USA. {ziangliu}@stanford.edu.

Digital Object Identifier (DOI): see top of this page.

of-the-art imitation learning approaches in both simulated and real-world experiments with humans, achieving higher task success rates and lower, wasteful interaction forces. By leveraging the generative capabilities of diffusion models, we have presented a significant step towards enabling effective human-robot collaboration on continuous state, joint-action tasks that require rapid mutual co-adaptation, behavior multimodality, coordination, and shared task understanding via learned, implicit constraints from human interaction.

II. RELATED WORK

Imitation learning: A multitude of work has been done in improving the quality of policies trained with offline datasets, particularly to account for multimodality. In particular, behavioral cloning has made significant progress due to advances in policy representation, by moving from *explicit* representations (e.g. LSTM-GMM [5], Transformers [6], normalizing flows [7]) to *implicit* representations (energy-based models [8], score-based models [9], diffusion models [4]), which allow for the expression of multimodal outcomes. Chi, et. al. showed that the diffusion policy representation can surpass other policy representations in expressing multimodality over a variety of tasks [4]. While these methods demonstrate the generative power of diffusion models, their effectiveness in scenarios that involve human behavior recognition remains to be seen.

Human intent modelling for HRI: Prior works in human-robot interaction (HRI) have generally fallen along a spectrum defined by the degree to which the human or the world is modelled. Theory of Mind (ToM) methods, which ascribe mental states to the human with whom the robot interacts, generally involve learning human reward functions [10], [11], learning user type [12], human motion prediction [13], or latent strategies [14]. These approaches are predicated on the hypothesis that human behavior follows a goal-directed policy, but are not necessarily structured to allow for multimodal behaviors given an attribute. Contrary to ToM methods, black box methods leverage data to directly train robot policies, though many approaches are a mix of both. Recent advances in HRI suggest that a promising approach for modelling multimodal behavior is to leverage the expressiveness of generative models for planning (e.g. variational autoencoders [15], [16], [17], variational recurrent neural networks [18]), or learning a joint action policy (e.g. Co-GAIL [19]).

In this work, we leverage diffusion models to learn a co-policy, wherein we predict future action sequences of both agents given sequences of past partner actions and observations. Nikolaidis, et. al. [20] similarly investigates the human-robot table carrying task, using a discrete state-action formulation to model human adaptation during interaction. While our continuous state-action formulation does not explicitly model adaptation, it enables the robot to execute dynamic behaviors alongside a human in real-time.

III. DIFFUSION CO-POLICY FORMULATION

This section describes the formulation of the collaborative robot policy as a DDPM. First, we describe the collaborative

task and motivate the use of diffusion models, leading to the formulation of a co-policy which incorporates human action and scene conditioning for significantly improved human-robot coordination on continuous state-action tasks.

A. Problem Setting

Consider a human-robot system wherein the dynamics of the world state $s \in \mathcal{S} \subset \mathbb{R}^n$. Let $a^i \in \mathcal{A} \subset \mathbb{R}^m, i \in [H, R]$ define the action space of the human H and robot R . At any point in time t , the human can take action a_t^H jointly with the robot action a_t^R , and the joint action is denoted by the concatenated vector $\mathbf{a}_t = (a_t^H, a_t^R)$. The state progresses with the following dynamics:

$$s_{t+1} = f(s_t, \mathbf{a}_t) \quad (1)$$

Provided with a control sequence of joint actions, a rollout starting from initial state s_0 results in trajectory $\tau = (s_0, \mathbf{a}_0, \dots, s_T, \mathbf{a}_T)$.

Typically, the goal is to find a sequence of actions $\mathbf{a}_{0:T}^*$ that maximizes the sum of rewards $\sum_{t=0}^T r(s_t, \mathbf{a}_t)$ via trajectory optimization or reinforcement learning methods.

An instance of such a system is the collaborative carrying task [18], a human and a robot both carry a table at opposite ends, moving it from a start pose to a goal location while avoiding obstacles. In this task, we assume full observability of the state. The state is a 7-dim. vector of the 2D table pose (p_x, p_y, θ) and its velocity in the world frame: $s = [p_x, p_y, \cos\theta, \sin\theta, \dot{p}_x, \dot{p}_y, \dot{\theta}]$. Furthermore, the state can be *augmented* by concatenating the following: the initial pose of the table, the 2D goal position of the table, and a 6-dim observation of the map, which consists of a vector concatenation of the world-frame positions of a maximum of three obstacles in the map. We refer to this 18-dim augmented state as s' . Perception was not the focus of this work, so we used a low-dimensional vector representation; however, future work can amend this representation to an arbitrary number of local obstacles, or leverage visual information. Each agent's action is a 2D force applied at opposite ends of the table; thus, the joint action is 4-dim.

B. Approach

While we could formulate the diffusion model for planning with RL using classifier-guided sampling [2], [21], doing so would require hand-designing a collaborative reward function. However, manually designing reward functions is tricky and prone to over-specification, particularly for multi-agent collaborative scenarios, where multimodal behaviors arise from preference and various factors like diversity and inconsistency. Despite promising directions in increasing the nuance of learned reward functions [22], querying methods would not be viable in interactive, long-horizon tasks like collaborative carrying, as specifying the query itself would be non-trivial.

Our **key insight** is that an effective, learned robot policy used for human-robot collaborative tasks must be able to express a high degree of multimodality, predict actions in a

temporally consistent manner, and recognize a wide range of frequencies of human actions in order to seamlessly integrate with a human in the control loop. Given recent empirical breakthroughs in the generative quality of diffusion models [23], [24], we propose leveraging diffusion models to enable coordination on long-horizon, continuous state-action, human-robot collaborative tasks in novel settings during test time by relying on demonstrations to capture interaction dynamics.

C. Denoising Diffusion Probabilistic Models

DDPMs [25], [26], [23] are generative models that approach sample generation with an iterative denoising process modeled by Langevin dynamics. Data generation via diffusion works by denoising (reversing) a forward diffusion process that iteratively adds noise to data until it resembles as standard Gaussian. Specifically, samples from a standard Gaussian prior, $p(x_K) = \mathcal{N}(0, I)$ pass through K iterations of noise reduction based on a fixed iteration-dependent variance schedule (parameterized by σ_k , α_k and γ_k), producing K intermediate latent variables, x_{k-1}, \dots, x_0 , where x_0 is the noiseless output. The Gaussian noise predicted is parameterized by a network, $\epsilon_\theta(x_k, k)$. Thus, to sample $x_{k-1} \sim p(x_{k-1}|x_k)$, we compute:

$$x_{k-1} = \alpha_k \left(x_k - \gamma_k \epsilon_\theta(x_k, k) \right) + \sigma_k z \quad (2)$$

where $z \sim \mathcal{N}(0, I)$.

Clarification of notation: In this work, we use two different time steps in subscript: k to denote the diffusion timestep, and t to denote the prediction timestep, i.e. $s_{t,k}$ is the t^{th} state in the k^{th} diffusion step. Subscripts of noiseless quantities are omitted, e.g. s_t . Subscripts of constants parameterized only by k do not have a time-indexed subscript, e.g. ϵ_k .

D. Diffusion Co-Policy for Coordinating with Humans

We consider the task of modeling a robot co-policy as learning a probabilistic model for the robot that infers future *sequences* of *joint human-robot* actions, conditioned on past states, map information, and past human actions. More specifically, we seek to model the conditional distribution $p(\mathbf{a}_t | s'_t, a_{t-1}^H)$. Thus, we modify the DDPM in two ways: 1. Conditioning on past human actions, which allows the robot to derive an understanding of human strategy from past human action trajectories to aid future team predictions; and 2. High-level goal conditioning, wherein the robot can condition its predictions on where the carried table should land. Eq. 2 can be modified to model the conditional distribution:

$$\mathbf{a}_{t,k-1} = \alpha_k \left(\mathbf{a}_{t,k} - \gamma_k \epsilon_\theta(\mathbf{a}_{t,k}, s'_t, a_{t-1}^H, k) \right) + \sigma_k z \quad (3)$$

Eq. 3 can also be interpreted as a noisy gradient descent step, with the gradient of the energy-based model (EBM), $\nabla E_\theta(a_t, s'_t, a_{t-1}^H)$, and learning rate γ :

$$\mathbf{a}_t \leftarrow \mathbf{a}_t - \gamma \nabla E_\theta(\mathbf{a}_t, s'_t, a_{t-1}^H) \quad (4)$$

In other words, $\epsilon_\theta(\mathbf{a}_{t,k}, s'_t, a_{t-1}^H, k)$ predicts $\nabla E_\theta(\mathbf{a}_t, s'_t, a_{t-1}^H)$, which approximates the action-score gradient, i.e. $\nabla \log p_\theta(\mathbf{a}_t | s'_t, a_{t-1}^H)$. Song, et. al. [9] provides further background on score-based models and this relationship, but we summarize the implications. By learning the parameters of the action-score gradient (thus, the distribution $p_\theta(\mathbf{a}_t | s'_t, a_{t-1}^H)$), we can circumvent approximating the intractable normalization constant, a problem which pervades likelihood-based models by affecting training stability and limiting model expressivity [9]. By performing Stochastic Langevin Dynamics sampling on this gradient field, the network can express arbitrary, multimodal distributions, which is beneficial for learning behaviors in human-robot interactive tasks.

E. Network Architecture

We adopt the Transformer-Based Diffusion architecture from Chi, et. al. [4], which uses the minGPT [27] transformer decoder model for action prediction, and modify the inputs as follows. The model takes as input a sequence of T_o steps of augmented state-action pairs; more specifically, these pairs consist of augmented states s' and past human actions a^H . The model outputs a sequence of T_a joint action steps denoised by the diffusion model. Noise-injected joint actions, $\mathbf{a}_{t,k}$, are tokenized and passed to the transformer decoder, which uses a sinusoidal embedding to encode the k^{th} diffusion step inputs as well as k , which is prepended as first token. Positional embedding is applied to conditional inputs, s'_t and a_{t-1}^H , which are converted to a sequence before passed to the transformer decoder. The decoder then predicts the noise corresponding to each input in the time dimension for the k^{th} iteration. A causal attention mask constrains the attention of each action to itself and prior actions. The predicted joint action sequence is constructed only after the predicted noise is propagated through the noise scheduler following the reverse diffusion process.

As expected [4], the 1D temporal CNN diffusion model does not work as well as the Transformer-based model for this task due to oversmoothing the action space. For the task we focus on, the actions are inertial forces. Depending on the user and playing style, people tend to apply short impulses due to damping and oscillations when pivoting around obstacles or correcting speed, making the Transformer-based architecture useful for the table-carrying task.

F. Training

To train the diffusion model, unnoised joint action data, $\mathbf{a}_{t,0}$, and a value of k are randomly sampled, the latter of which is then used to sample noise ϵ_k with variance σ_k . $\epsilon_\theta(\mathbf{a}_{t,k}, s'_t, a_{t-1}^H, k)$ then predicts the noise from the data sample (with the noise added). The loss for the noise model is:

$$\mathcal{L} = \|\epsilon_k - \epsilon_\theta(s'_t, a_{t-1}^H, \mathbf{a}_{t,0} + \epsilon_k, k)\|_2^2 \quad (5)$$

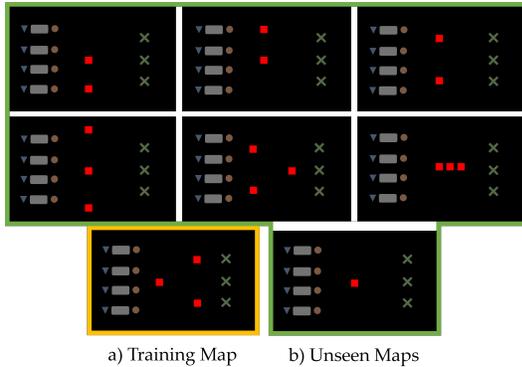


Fig. 2: Co-planning evaluation training (yellow box) and test (green box) maps. Start poses and goal positions (green “X”s) are shown, along with obstacle layouts (red boxes).

IV. EVALUATION

A key advantage of the Diffusion Co-policy is its generative quality; however, the model’s effectiveness on human-robot collaborative tasks remains unexplored. The collaborative carrying task itself poses several interesting questions for evaluation: 1) Can it learn to effectively condition on obstacles? 2) Can it learn to effectively condition on its partner’s behaviors and shared task representation? 3) Can it mutually adapt with real humans in test time? To attempt to address these questions, we compare our diffusion co-policy against several state-of-art imitation learning methods on a suite of evaluations described in the following sections.

We select three learning-based methods for comparison:

- **BC-LSTM-GMM [5]**: Several works in HRI have leveraged a variant of this method, notably for the GMM output layer. We adapt the implementation from [5] and did not condition inputs on past human actions; doing so led to worse performance on the task.
- **VRNN Planner [18]**: This sampling-based planner autoregressively predicts team waypoints learned from demonstrations in receding horizon and does not condition on human actions.
- **Co-GAIL [19]**: Co-GAIL learns collaborative behaviors from demonstrations and maps human behaviors to a latent space, which is then used to train a co-policy with Generative Adversarial Imitation Learning (GAIL).

Some baselines do not incorporate map information. To improve task performance, we trained them using the augmented state representation, s'_t . We also compare two variants of the Diffusion Co-Policy: one with past human action conditioning (**CoDP-H**), and the other without (**CoDP**).

A. Experimental Setup

We trained the diffusion co-policy, CoDP, and variant (CoDP-H) to output joint actions at 10Hz, which were executed on a simulator running at 30Hz, applying a zero-order hold of 3 time steps for each planned. All other baselines were trained to produce outputs at 30Hz. We conduct the

following experiments and user studies for the collaborative carrying task to address the questions posed above.

1) **Co-planning (in simulation environment)**: To test each method’s ability to complete the task without a human in the loop and learn a representation of the map without the potential added benefit of human co-piloting corrections, we varied out-of-distribution obstacle locations while keeping the same distribution of initial and goal states from the training data. We executed $T_a = 8$ actions sampled with 100 denoising steps before replanning, which takes roughly 0.3 sec on a NVIDIA 3090Ti GPU.

2) **Human-in-the-loop evaluation (in simulation environment)**: In this user study, the robot policies complete the task with a real human-in-the-loop, in various out-of-training-distribution settings. The human teleoperates the orange circle agent in the simulation using joystick control. The sampling scheme for the human-in-the-loop simulation evaluation is different than the co-planning setting due to having a human in the loop. To account for visual latency and reaction time, we execute $T_a = 1$ sampled actions with 34 inference steps to allow for planning time of roughly 0.1 sec, and zero-order hold each planned action for 3 time steps before executing the next planned action, resulting in low visual latency in the simulator.

Here, we study the effect of the policy or planning method on success rate (i.e. the percentage of trials in which the user succeeds in completing the task with the robot). Our *hypothesis* is that the diffusion policies will enable the robot to complete the task with a human at a higher success rate than the other methods, particularly since the diffusion policies are generating multimodal, multi-step predictions.

3) **Human-in-the-loop evaluation (in real environment)**: In this user study, a human uses a joystick to teleoperate an Interbotix Locobot that is pin-connected via a rigid rod to another Locobot, operated by a policy or planner. We use the same policies trained in simulation and deploy them in the real environment in zero-shot sim-to-real transfer. We also use the same sampling scheme from the human-in-the-loop evaluation in simulation. To address the sim-to-real gap of the state space, we mapped a space in the experiment room and scaled it to the simulation environment coordinates, using data from motion capture. For this experiment, we consider two initial configurations: in the first, which we denote as “Human Front”, the human is placed closest to an obstacle such that they are inclined to make the decision to go above or below the obstacle before the robot; and in the second, which we denote as “Robot Front”, the robot is placed closest to an obstacle, implicitly forcing the robot to make the same decision before the human.

In this study, we investigate the robustness of the robot’s policy or planner to its initial configuration by determining whether there is a significant interaction effect of those two factors. Our *hypothesis* is that robots running the diffusion methods should not see an interaction effect with initial configuration since diffusion methods can express a high degree of multimodality.

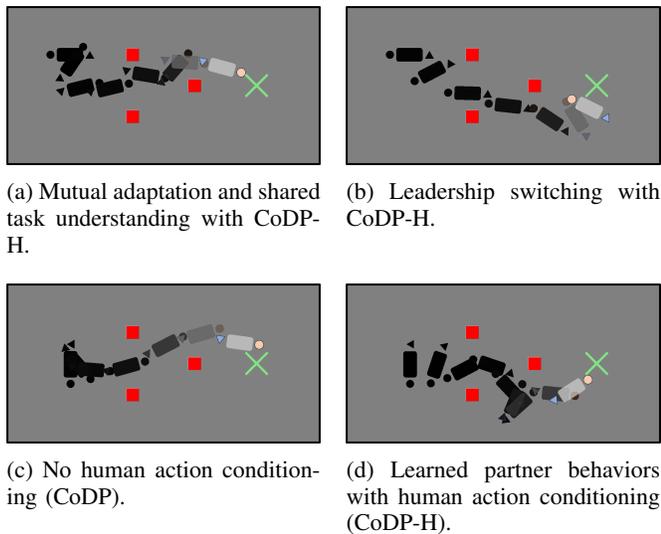


Fig. 3: **Qualitative Observations:** (a) *Mutual adaptation and shared task understanding* between human (orange circle) and CoDP-H agent (blue triangle). (b) *Leadership switching* with CoDP-H. (c) *Without conditioning on past human behavior*, robot behavior is notably affected: the human becomes de-facto leader as the robot displays passive behavior. (d) *With conditioning* on its partner’s past actions, the robot actively takes the lead, and displays interesting leadership switching behaviors via pivoting.

B. Dataset and Map Details

1) *Training data:* The training dataset consists of 376 human-human demonstrations (179,993 environment interactions) on the collaborative carrying task collected by 5 distinct pairs of people on a total of 36 possible configurations. Due to the added complexity of obstacle representation learning, cost of demonstration collection, and requirement of hand-engineered obstacle placement to allow for multimodal behaviors, we used up to a maximum of three obstacles in each training demonstration, with each initial, goal, and obstacle location illustrated in Fig. 2a. Note that while many offline dataset learning methods [28] augment data with trained RL policies, planners paired with PID controllers, etc., we recognize that such augmented data could skew our dataset distribution, particularly if these methods do not contain demonstrations of multimodal, sub-optimal, and inconsistent, yet “human-like” behaviors. For example, RRT planners do not exhibit the same behaviors (e.g. rotations, distance from obstacles) as human demonstrators on the collaborative carrying task [18]. The demonstration data in this work contains multimodal behaviors, as seen in Fig. 4.

2) *Test maps:* For the co-planning evaluation, we evaluated on all possible combinations of unseen map settings outlined in Fig. 2b. For the human-in-the-loop simulation evaluation, we sampled from a subset of unseen maps, goals, in addition to different initial orientations. For reference, π is the initial table orientation depicted in Fig. 2, and we included four total initial orientations in our sampling, i.e. $[0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}]$. We then evaluated on the same sampled subset per method.

On the real robot evaluation, we used a one obstacle unseen map with two initial orientations: robot facing the obstacle first, and human facing the obstacle first.

Method	Unseen Maps Success (%)	Test Holdout Maps Success (%)
CoDP-H	40.48 / 32.54	78.57 / 77.38
CoDP	32.14 / 28.97	72.62 / 67.86
BC-LSTM-GMM[5]	19.05 / 17.06	60.71 / 55.16
Co-GAIL[19]	25.00 / 22.22	47.62 / 39.28
VRNN[18]	22.62 / 19.44	22.62 / 20.24

TABLE I: **Co-planning Results.** Reported success rates for each method as (max performing seed / average performance over 3 random training seeds) over a total of 84 randomly selected test holdout maps (Fig. 2a) and 84 novel configurations on unseen maps (Fig. 2b). Our results show that the diffusion co-policy conditioned on past human partner actions (CoDP-H) outperformed all state-of-art imitation learning methods and baselines on the co-planning evaluation.

C. Simulation Results: Key Findings

1) *Co-planning:* We subjected the co-policies to the collaborative carrying task on maps seen in training, as well as maps unseen in training. **CoDP-H** consistently outperforms other baselines on all maps tested (Table I). Considering the few maps and obstacle configurations used during training, methods leveraging diffusion models exhibited unexpectedly high performance on maps with novel obstacle locations, suggesting that the diffusion co-policy was able to learn obstacle representations efficiently.

2) *Human-in-the-loop trials:* The diffusion methods outperform other baselines on the human-in-the-loop evaluation on task success rate (Table II). We validate this trend by conducting a one-way repeated measures ANOVA to examine the effect that the methods had on task success rate. Results showed statistically significant differences in success rate ($F(4, 16) = 11.65, p < .001$). We ran a post-hoc analysis with Tukey HSD corrections for multiple comparisons, which showed the human-in-the-loop performance of the diffusion methods to be significantly different from that of Co-GAIL, BC-LSTM-GMM, and VRNN, and all contrasts with $p < .0001$. The contrasts support our hypothesis that using a diffusion network for the robot policy yields higher success rates on the collaborative human-robot task, as it surpasses other methods at predicting multimodal behavior. Fig. 4 shows heat maps of the state visitation frequencies for the human-in-the-loop simulation evaluations for each policy or planner, as well as rollouts from the human-human demonstration dataset. Note in the test maps that obstacles were placed in locations where the table frequently visited in the training data. This suggests that the diffusion methods are able to learn a shared task representation and perform well on unseen maps with a real human partner.

Table II also demonstrates the planning time disadvantage of diffusion-based methods; yet, despite planning less frequently and requiring interpolation methods, diffusion-based methods achieve higher success rate on the task.

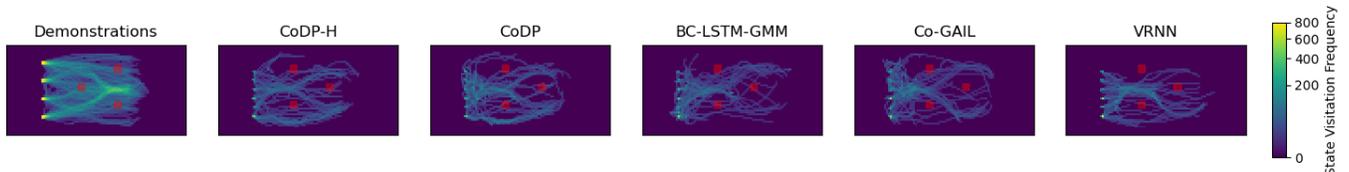


Fig. 4: Heat map visualization of *state visitation frequencies*, from left to right, of human-human demonstration data for training the models, and the human-in-the-loop simulation evaluations on novel, unseen maps for each policy or planner. In each scenario, one to three obstacles are placed in the the potential obstacle locations shown as red squares for each map.

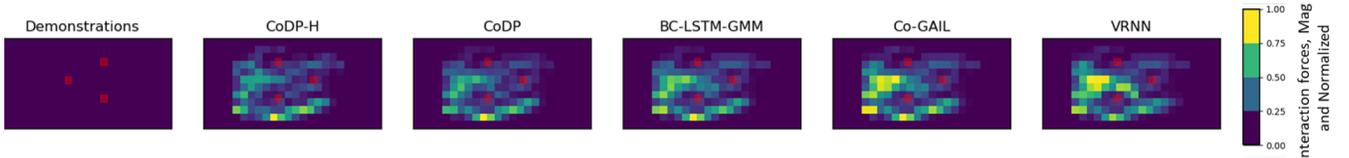


Fig. 5: Comparison of interaction forces from the human-human demonstrations and simulation evaluations with a human-in-the-loop on novel, unseen maps. For each trajectory, we binned the interaction force *magnitudes* (see color-map bar for bins) at each location in the discretized map, then *normalized* the magnitudes across *all* trajectories. Each map shows 1 – 3 obstacles at potential obstacle locations (red squares). The diffusion methods show overall lower magnitudes of interaction forces across all human-in-the-loop evaluations in simulation.

Method	Success (%)	Time (s)	Plan Time (ms)
CoDP-H	68 ± 2.1	32.3 ± 0.2	93.5 ± 4.0
CoDP	61 ± 3.7	31.0 ± 0.1	87.0 ± 3.8
BC-LSTM-GMM[5]	36 ± 1.4	22.3 ± 0.5	0.745 ± 0.004
Co-GAIL[19]	37 ± 1.9	23.9 ± 0.9	0.267 ± 0.003
VRNN[18]	35 ± 3.6	15.8 ± 1.4	18.52 ± 0.02

TABLE II: **Human-in-the-Loop Simulation Results.** The max performing model seed was used for each robot planner or policy, which played with human subjects ($N = 5$) for a total of 60 randomly selected configurations on unseen maps (Fig. 2b). Standard error (SE) is reported for all measurements, including success rate (%), time to completion (s) for successful trajectories, and average time for the model to plan (ms). Our results show that the diffusion co-policy conditioned on past human partner actions (CoDP-H) outperformed all baselines for the human-in-the-loop evaluation in simulation.

D. Interesting Behaviors in Human-in-the-loop Evaluation

Diffusion co-policy demonstrated interesting collaborative behaviors on novel configurations in simulation evaluation. We highlight them qualitatively as follows:

1) *Mutual adaptation and shared task understanding*: Fig. 3a demonstrates an instance of mutual adaptation as well as shared task understanding. Initially, both agents simultaneously choose different strategies: the robot (blue triangle) applies a downward force while the human (orange circle) applies an upward force, resulting in an in-place moment on the table. The human leads, while the robot maintains awareness of obstacles and rotates to avoid them, demonstrating shared task understanding between human and robot. These behaviors lend insight into the better performance of CoDP-H in human-in-the-loop evaluations.

2) *Leadership switching*: Fig. 3b demonstrates an instance of leadership switching, which occurs several times over the

course of the trajectory. The human starts leading by moving below the obstacle, but the robot takes over by maintaining its lead in front. Both agents approach the goal past the final obstacle. This demonstrates the ability to switch roles while maintaining task understanding.

3) *Learning partner behaviors*: Conditioning on past partner actions allows the robot to develop a better understanding of the task and its partner’s intentions. Without this past action conditioning, the robot acts passively, leaving the human to lead (Fig. 3c). Fig. 3d shows that the CoDP-H robot is capable of pivoting, a proactive behavior, since it has learned to associate past partner actions with observations. This behavior was not seen in the demonstration data.

4) *Low interaction forces*: Stretching or compressing may occur during transport of an object, indicating non-zero interaction forces [29]. Interaction forces do not contribute to motion, and can lend insight into periods of collaboration, disagreement, and other decision points in the trajectory. Fig. 5 shows a 2D heat map of normalized interaction force magnitudes over all trajectories for each method. Interaction forces less than 0.25 are generally negligible, and all human-human demonstrations displayed a negligible frequency of non-zero interaction forces. Forces between 0.25 - 0.75 may indicate a decision point or dissent; those above 0.75 are strong indicators of dissent or human corrective action. Fig. 6 shows interaction forces across an example map configuration for each method with a human in the loop. Across all methods, the diffusion policies display lower interaction forces over most areas in the maps.

E. Real robot evaluation

Table III shows that the diffusion methods outperform the other baselines, except for BC-LSTM-GMM in the “Robot Front” case. While BC-LSTM-GMM appears to outperform the diffusion methods in the “Robot Front” case, it performs

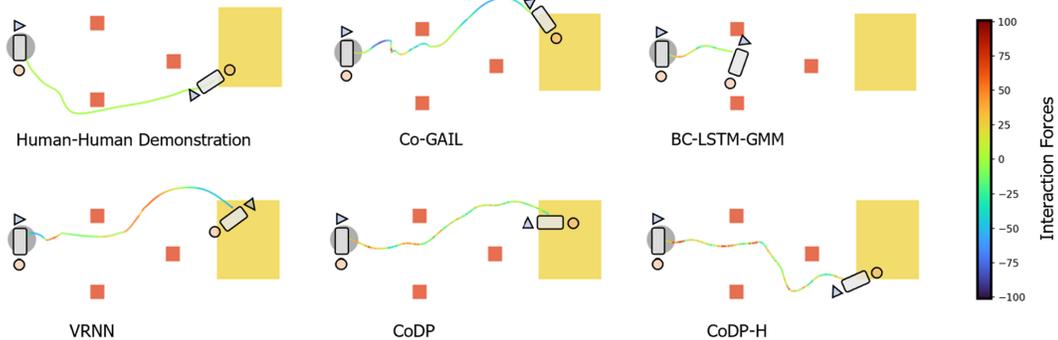


Fig. 6: Visualization of interaction forces over rollouts on a sample map configuration played between each robot policy or planner (blue triangle) with a human (orange circle) for evaluation in simulation, with the goal (yellow region) and obstacles (red squares) shown. Closer to zero is better, as demonstrated in the human-human exemplar. The diffusion policies show overall lower (in magnitude) interaction forces over the course of the trajectory.

Method	Human Front Success (%)	Robot Front Success (%)
CoDP-H	75.0 ± 4.6	91.7 ± 3.4
CoDP	66.7 ± 4.3	91.7 ± 3.4
BC-LSTM-GMM[5]	33.3 ± 4.3	100.0 ± 0.0
Co-GAIL[19]	66.7 ± 4.3	50.0 ± 7.4
VRNN[18]	75.0 ± 4.6	75.0 ± 7.0

TABLE III: **Human-in-the-loop Real Robot Experimental Results.** We tested on a real-robot scenario with a single centered obstacle (see bottom unseen map in Fig. 2), in two initial configurations: one with the human facing the obstacle first (“human front”), and the other with the robot facing the obstacle first (“robot front”). Average success rate (%) is reported with std. dev. over subjects ($N = 6$), with a total of 12 trajectories per initial configuration. BC-LSTM-GMM outperforms CoDP-H in the robot front setting, but does significantly worse in the human front setting.

poorly in the “Human Front” case. BC-LSTM-GMM prefers a route below the obstacle; if the human partner happens to adapt to the robot or pick the same route, this tends to lead to success. However, unlike CoDP-H, it is unable to adapt to move above the obstacle when necessary to achieve task success, as seen in Fig. 7. This suggests a significant interaction between the policy or planner method and the initial configuration of the robot with respect to the obstacle, which is confirmed by results from a two-way repeated measures ANOVA for interaction effects, $F(4,20) = 3.170$, $p = 0.036$. Main effects on task success rate were also significant for initial configuration, $F(1, 5) = 7.857$, $p = 0.038$, and for method, $F(4, 20) = 3.152$, $p = 0.037$. We further investigated the interaction effect of the initial configuration for each method. Adjusted P-values using Holm multiple testing corrections show that the effect of initial configuration on success rate was significant for BC-LSTM-GMM ($p = 0.001$), but not for the other methods. This supports our hypothesis, and suggests that BC-LSTM-GMM is affected by the initial configuration and therefore less robust to multimodal outcomes that arise from human-robot interactions. While the other methods do

not show significant interaction effects, they perform poorly in the task compared to the diffusion methods.

V. CONCLUSION

In this work, we explore using action predictions from diffusion models to plan collaborative actions that synergize well with real humans in the loop. We show that a co-policy developed with a Transformer-based diffusion model and conditioned on past human actions can not only plan multimodal action sequences with real humans-in-the-loop to achieve high success rates, but also qualitatively display compelling collaborative behaviors in novel, out-of-training-distribution settings, including mutual adaptation, shared task understanding, and leadership switching.

Our study has several limitations. The time required to generate a sample with the diffusion policy is longer than other methods. We also faced several limitations in the real robot experiments, including physical capabilities of the robots, physical space constraints, and human subject variance. To extend this method for a co-manipulation task similar to [30], a dataset with tactile feedback from both agents co-manipulating a table would be highly beneficial for the learned task representation. Despite limitations in this work, the diffusion co-policy has demonstrated the significance of an expressive policy for human-robot collaboration, i.e. one that can capture a high degree of multimodality, predict actions in a temporally consistent manner, and recognize a wide range of frequencies of actions in order to seamlessly integrate with a human.

REFERENCES

- [1] Natalie Sebanz, Harold Bekkering, and Günther Knoblich. Joint action: bodies and minds moving together. *Trends in cognitive sciences*, 10(2):70–76, 2006.
- [2] Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022.
- [3] Julen Urain, Niklas Funk, Jan Peters, and Georgia Chalvatzaki. Se (3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5923–5930. IEEE, 2023.

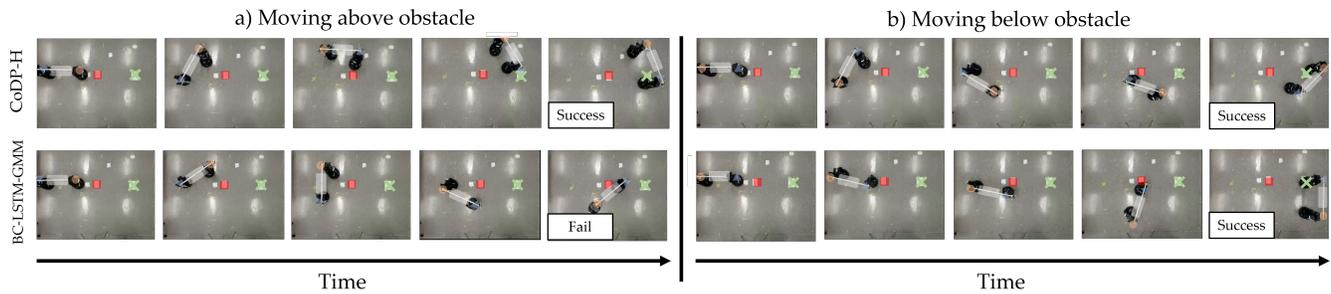


Fig. 7: All robots are represented by the blue triangle. The human is the orange circle, the goal is a green "X", and the obstacle is a red square. BC-LSTM-GMM prefers a route *below* the obstacle, leading to greater success rate if the human also chooses the route below the obstacle; otherwise, it can lead to failure. With the BC-LSTM-GMM robot facing the obstacle first, the human serves a more passive role, resulting in a higher success rate compared to when the human faces the obstacle first, since the robot will choose the route below the obstacle.

- [4] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023.
- [5] Ajay Mandlekar, Danfei Xu, J. Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Mart'ın-Mart'ın. What matters in learning from offline human demonstrations for robot manipulation. In *5th Annual Conference on Robot Learning*, 2021.
- [6] Nur Muhammad Shafiullah, Zichen Cui, Ariuntuya Arty Altanzaya, and Lerrel Pinto. Behavior transformers: Cloning k modes with one stone. *Advances in neural information processing systems*, 35:22955–22968, 2022.
- [7] Julen Urain, Michele Ginesi, Davide Tateo, and Jan Peters. Imitation-flow: Learning deep stable stochastic dynamic systems by normalizing flows. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5231–5237. IEEE, 2020.
- [8] Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In Aleksandra Faust, David Hsu, and Gerhard Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 158–168. PMLR, 08–11 Nov 2022.
- [9] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [10] Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.
- [11] Dorsa Sadigh, Shankar Sastry, Sanjit A Seshia, and Anca D Dragan. Planning for autonomous cars that leverage effects on human actions. In *Robotics: Science and systems*, volume 2, pages 1–9. Ann Arbor, MI, USA, 2016.
- [12] Stefanos Nikolaidis, Ramya Ramakrishnan, Keren Gu, and Julie Shah. Efficient model learning from joint-action demonstrations for human-robot collaborative tasks. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, pages 189–196, 2015.
- [13] Georgia Chalvatzaki, Xanthi S Papageorgiou, Petros Maragos, and Costas S Tzafestas. Learn to adapt to human walking: A model-based reinforcement learning approach for a robotic assistant rollator. *IEEE Robotics and Automation Letters*, 4(4):3774–3781, 2019.
- [14] Annie Xie, Dylan Losey, Ryan Tolsma, Chelsea Finn, and Dorsa Sadigh. Learning latent representations to influence multi-agent interaction. In *Conference on robot learning*, pages 575–588. PMLR, 2021.
- [15] Edward Schmerling, Karen Leung, Wolf Vollprecht, and Marco Pavone. Multimodal probabilistic model-based planning for human-robot interaction. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3399–3406. IEEE, 2018.
- [16] Judith Bütepage, Ali Ghadirzadeh, Özge Öztimur Karadağ, Märten Björkman, and Danica Kragic. Imitating by generating: Deep generative models for imitation of interactive tasks. *Frontiers in Robotics and AI*, 7:47, 2020.
- [17] Vignesh Prasad, Dorothea Koert, Ruth Stock-Homburg, Jan Peters, and Georgia Chalvatzaki. Mild: Multimodal interactive latent dynamics for learning human-robot interaction. In *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, pages 472–479. IEEE, 2022.
- [18] Eley Ng, Ziang Liu, and Monroe Kennedy III. It takes two: Learning to plan for human-robot cooperative carrying. *arXiv preprint arXiv:2209.12890*, 2022.
- [19] Chen Wang, Claudia Pérez-D'Arpino, Danfei Xu, Li Fei-Fei, Karen Liu, and Silvio Savarese. Co-gail: Learning diverse strategies for human-robot collaboration. In *Conference on Robot Learning*, pages 1279–1290. PMLR, 2022.
- [20] Stefanos Nikolaidis, David Hsu, and Siddhartha Srinivasa. Human-robot mutual adaptation in collaborative tasks: Models and experiments. *The International Journal of Robotics Research*, 36(5-7):618–634, 2017.
- [21] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.
- [22] Vivek Myers, Erdem Biyik, Nima Anari, and Dorsa Sadigh. Learning multimodal rewards from rankings. In *Conference on Robot Learning*, pages 342–352. PMLR, 2022.
- [23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- [24] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- [25] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.
- [26] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.
- [27] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [28] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning, 2020.
- [29] Vijay R Kumar and Kenneth J Waldron. Force distribution in closed kinematic chains. *IEEE Journal on Robotics and Automation*, 4(6):657–664, 1988.
- [30] Doganay Sirintuna, Alberto Giammarino, and Arash Ajoudani. Human-robot collaborative carrying of objects with unknown deformation characteristics. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10681–10687. IEEE, 2022.