

# Enhancing Exploration with Diffusion Policies in Hybrid Off-Policy RL: Application to Non-Prehensile Manipulation

Huy Le, Miroslav Gabriel, Tai Hoang, Gerhard Neumann, Ngo Anh Vien

**Abstract**—Learning diverse policies for non-prehensile manipulation is essential for improving skill transfer and generalization to out-of-distribution scenarios. In this work, we enhance exploration through a two-fold approach within a hybrid framework that tackles both discrete and continuous action spaces. First, we model the continuous motion parameter policy as a diffusion model, and second, we incorporate this into a maximum entropy reinforcement learning framework that unifies both the discrete and continuous components. The discrete action space, such as contact point selection, is optimized through Q-value function maximization, while the continuous part is guided by a diffusion-based policy. This hybrid approach leads to a principled objective, where the maximum entropy term is derived as a lower bound using structured variational inference. We propose the Hybrid Diffusion Policy algorithm (HyDo) and evaluate its performance on both simulation and zero-shot sim2real tasks. Our results show that HyDo encourages more diverse behavior policies, leading to significantly improved success rates across tasks - for example, increasing from 53% to 72% on a real-world 6D pose alignment task. Project page: <https://leh2rng.github.io/hydo>

## I. INTRODUCTION

The ability to manipulate objects in ways beyond simple grasping is a vital aspect of human dexterity, underscoring the significance of learning advanced non-prehensile manipulation skills. These complex skills are essential for a wide range of tasks, from daily activities to advanced industrial applications. Teaching robots to achieve a level of dexterity similar to the one of humans remains a significant challenge for the field of robotics [1], [2]. Previous research has made significant advances in this area, but often suffers from limitations in object generalization and motion complexity [3], [4], [5]. To address these challenges, motion primitives (MPs) are frequently employed to simplify the representation of long-horizon actions and thus the overall problem complexity. In addition, object-centric action representations are utilized to decrease the sample complexity and to enable a more efficient learning process. Reinforcement Learning (RL) can be used to learn such representations, especially within hybrid action spaces that combine discrete contact points with continuous parameters for MPs [6], [7].

Developing policies that can learn diverse behaviors in a RL context is motivated by two key factors: First, these poli-

cies are potentially able to improve generalization to out of distribution states and observations. Policies that only overfit to a narrow range of experiences, on the other hand, usually do not perform sufficiently well in unseen environments. Continually learning from a more diverse set of experiences forces the policy to capture the underlying principles of the task rather than merely optimizing for specific scenarios encountered during training [8], [9]. Secondly, developing such policies enhances skill transfer learning: Continuous exposure to diverse experiences in online RL facilitates the transfer of skills across different but related tasks. An agent that learns from a broad spectrum of interactions is more likely to develop a robust set of skills that can be applied to a variety of tasks. This increases its versatility and overall learning capability [8], [9], [10].

In this work, we introduce a novel approach to enhancing exploration in non-prehensile manipulation tasks under a hybrid off-policy reinforcement learning framework [7]. Our method handles both discrete and continuous action spaces by incorporating maximum entropy principles to encourage diverse behaviors. Specifically, we represent the continuous motion parameter policy using a diffusion model [11], [12], [13], [14], while the discrete action space, such as contact points, is optimized through Q-value function maximization. This formulation leads to the development of a Hybrid Diffusion Policy algorithm, called **HyDo**, which integrates two main components: diffusion-based policies and maximum entropy optimization over both discrete and continuous actions. The entropy maximization term, embedded in the soft actor-critic (SAC) [15] algorithm, is derived as a lower bound using structured variational inference. The overall framework is illustrated in Fig. 1. We evaluate the impact of combining maximum entropy regularization with diffusion in both simulation and zero-shot sim2real tasks. The results show that this combination helps to learn more diverse behavior policies. In the zero-shot sim2real transfer, this improved exploration leads to a significant increase in success rates, from 53% to 72% on a 6D pose alignment task with a physical Franka Panda robot.

In summary, our contributions are: i) a hybrid RL framework that enhances exploration by incorporating diffusion-based policies; ii) the integration of maximum entropy regularization to encourage diverse behaviors across both action spaces; and iii) a theoretical justification, showing that the new objective is a lower bound derived via structured variational inference. We validate our methods on both simulated and zero-shot sim2real non-prehensile manipulation tasks.

Huy Le is with the Bosch Center for Artificial Intelligence, Renningen, Germany and also with the Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany. [baohuy.le@de.bosch.com](mailto:baohuy.le@de.bosch.com)

Miroslav Gabriel, Ngo Anh Vien are with the Bosch Center for Artificial Intelligence, Renningen 71272, Germany

Tai Hoang, Gerhard Neumann are with the Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany.

Our project page is available at <https://leh2rng.github.io/hydo>

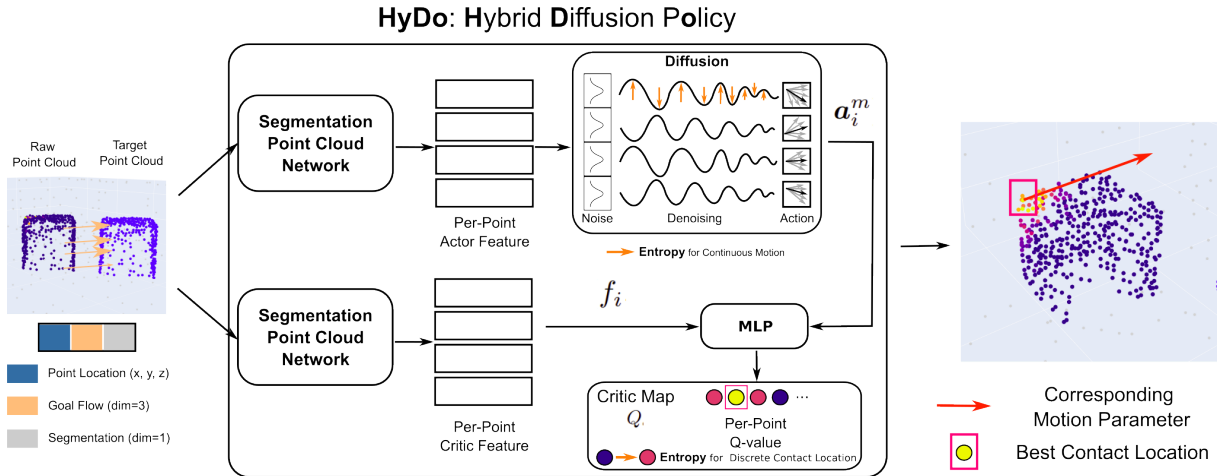


Fig. 1: Overview of HyDO: The network takes point clouds, goal flow, and segmentation (indicating object and background points) as input. These are passed through the actor and critic networks. The actor is enhanced exploration on the continuous motion parameter with the entropy regularizer applied during the diffusion process and outputs the motion parameter. The state-action pair is then evaluated by the critic, which also integrates entropy regularization for exploration on the discrete contact location. The action with the highest Q-value is selected and executed by the robot.

## II. RELATED WORK

### A. Diffusion-based offline and online RL

There have been different efforts, such as diffusion-based generative models DDPMs [12], to use diffusion models as representations for RL policies because they are (among others) capable of learning multi-modal and diverse behaviors. Most works focus on offline RL in which all data is available at training time. Diffusion-Q learning [16] proposes an explicit regularization of the cloned behavior policy. A follow-up work [13] proposes to use consistency models as policies since they improve inference speeds such that they can be applied to online RL. Other variants of diffusion policies introduce actor-critic methods via implicit Q-learning [17], Q-guided policy optimization with a new formulation for intermediate guidance in diffusion sampling process [18], diffusion-constrained Q-learning on latent spaces [19] or implicit Q-learning as an actor-critic method [17], [20].

Analogically, diffusion policies are applied in imitation learning [14], goal-conditioned imitation learning [21], human behavior imitation learning [22], offline direct policy search [23], [24] based on advantage-weighted regression (AWR) or reward-weighted regression (RWR) [25]. There are also few works which apply online RL to fine-tune pretrained diffusion models, such as RWR-based methods [26], [27] or advanced bandit setting-based methods [28], [29]. While these RWR-and AWR-based methods are considered to implicitly enforce entropy-max regularization, they are only 1-step MDPs and ignore the exploration problem for training from scratch. Closest to our work is Q-score matching for actor-critic (QSM) [30] which applies soft policy iteration, but without direct use of the soft Q-value function. Our work however optimizes the entropy-max objective using soft Q-value functions for the policy improvement step as well as the soft Bellman updates of the policy evaluations step.

### B. Manipulation skill learning with motion primitives

Non-prehensile manipulation involves manipulating objects without grasping them. Recent learning-based methods in this field are limited either by skill complexity or skill diversity [5], [2], [31]. Our work tries to address these challenges by focusing on learning diverse behaviors for 6D object manipulation with MPs and by optimizing a class of multi-modal policies. Traditional RL algorithms focus either on discrete or on continuous action spaces, despite the fact that some applications require hybrid action spaces where the agent selects a discrete action along with some continuous parameter. Recent approaches [6], [7], [32] address these shortcomings by using a spatial action representation with discrete actions defined over a visual input map but fail to incorporate profound exploration strategies because they either ignore exploration over spatial maps [6] or use a simple  $\epsilon$ -greedy strategy to explore over a continuous action parameter space [7]. Another recent work proposes a diffusion-based MP policy [33] that is similar to us that the policy generates only MP parameters which are then used to generate a full motion profile via ProDMP [34].

## III. BACKGROUND

### A. MDP and Off-Policy Actor-Critic Methods

The underlying problem of RL can be formulated as a Markov decision process (MDP) [35] which is defined by a tuple  $\{S, A, R, P, \gamma, P_0\}$  with state space  $S$ , action space  $A$ , rewards  $R$ , transition probabilities  $P$ , discount factor  $\gamma$ , and initial state distribution  $P_0$ . The objective is to maximize the discounted cumulative reward  $R = \sum_{i=0}^{\infty} \gamma^i r(s_{t+i}, a_{t+i})$  with states  $s_t$  and actions  $a_t$ . Off-policy actor-critic methods decouple the policy used to generate data from the policy being optimized. The *actor* is represented by a policy  $\pi_{\theta}(a|s)$  with parameters  $\theta$ , whereas the *critic* estimates the

value function  $Q_\phi(s, a) = \mathbb{E}[\sum_t \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a]$  with parameters  $\phi$ . TD3 [36] and SAC [15] are two prominent off-policy approaches. The objectives of SAC’s actor and critic are defined as follows:

$$\begin{aligned} L(\theta) &= \mathbb{E}_{s, a \sim D} [-Q_\phi(s, a) + \alpha \log \pi_\theta(a|s)], \\ L(\phi) &= \frac{1}{2} \mathbb{E}_{s_t, a_t, s_{t+1} \sim D} [\|Q_\phi(s_t, a_t) - y_t\|^2], \end{aligned} \quad (1)$$

where  $y_t$  is the target

$$y_t = r_t + \gamma \mathbb{E}_{a_{t+1} \sim \pi_\theta(s_{t+1})} [Q_{\phi'}(s_{t+1}, a_{t+1}) - \alpha \log \pi(a_{t+1}|s_{t+1})].$$

In contrast to TD3 that optimizes deterministic policies and exploration is handled by an  $\epsilon$ -greedy strategy, SAC uses a stochastic  $\pi_\theta$  and adds an entropy maximization term to the objective to encourage exploration. This promotes a more diverse set of behaviors and provides an improvement in exploration and training stability.

### B. Diffusion and Consistency Models

Diffusion-based generative models DDPMs [12], [11] assume  $p_\theta(x_0) := \int p_\theta(x_{0:T}) dx_{1:T}$ , where  $x_1, \dots, x_T$  are latent variables with the same dimensionality as the data  $x_0 \sim p(x_0)$ . In a forward diffusion chain  $q$ , given by  $q(x_{1:T}|x_0) := \prod_{t=1}^T q(x_t|x_{t-1})$ ,  $q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I)$ , noise with a predefined variance schedule  $\beta_i$  is gradually added to the data over a fixed amount of time steps  $T$ . Subsequently, a reverse diffusion chain  $p$ , defined as  $p_\theta(x_{0:T}) := \mathcal{N}(x_T; 0, I) \prod_{t=1}^T p_\theta(x_{t-1}|x_t)$ , is optimized by maximizing the evidence lower bound. Inference then requires sampling the reverse diffusion chain from  $t = T$  to  $t = 0$ .

Consistency models [37] extend diffusion models by adopting the form of a probability flow ordinary differential equation (ODE). The reverse process along the ODE path  $\{\hat{x}_\tau\}_{\tau \in [\epsilon, T]}$  generates data starting from  $\hat{x}_T \sim \mathcal{N}(0, T^2 I)$ , where  $\epsilon$  is a small value close to zero. The consistency model retains the effectiveness of a diffusion model but accelerates sampling by reducing the number of time steps.

### C. Diffusion and Consistency Models as RL Policy

Diffusion models have been used as a new class of policies in offline RL [16], [38] for actor-critic architectures. These works share a similar parametric policy representation as the reserve process of the conditional diffusion model which is defined as

$$\pi_\theta(\mathbf{a}|\mathbf{s}) := \pi_\theta(\mathbf{a}^{0:K}|\mathbf{s}) = N(\mathbf{a}^K; \mathbf{0}, I) \prod_{k=1}^K p_\theta(\mathbf{a}^{k-1}|\mathbf{a}^k, \mathbf{s}) \quad (2)$$

where  $\mathbf{a}^0$  is the action executed by the agent and sampled at step 0. The probability distribution  $p_\theta$  can be based on DDPM [16], [38] in which  $p_\theta(\mathbf{a}^{k-1}|\mathbf{a}^k, \mathbf{s})$  is a Gaussian with a trainable mean  $\mu(\mathbf{a}^k; k, \mathbf{s})$  and a fixed time-dependent covariance  $\Sigma(\mathbf{a}^k; k, \mathbf{s}) = \beta^k I$ .

With the above policy representation, Wang et. al. [16] propose Diffusion Q-learning to optimize  $\theta$  using a similar update scheme as TD3+BC [39]. Given an offline dataset

$D = \{\mathbf{s}_t, \mathbf{a}_t^0, r_t, \mathbf{s}_{t+1}\}_{t=0:T}$ , the objective of the policy evaluation step is

$$L(\phi) = \mathbb{E}_{\mathbf{s}_t, \mathbf{a}_t^0, r_t \sim D} \left[ \sum_{i=1}^2 \|y_t - Q_{\phi_i}(\mathbf{s}_t, \mathbf{a}_t^0)\|^2 \right],$$

where  $y_t = r(\mathbf{s}_t, \mathbf{a}_t^0) + \gamma \min_{i=1,2} Q_{\phi_i'}(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}^0)$ ,  $\mathbf{a}_{t+1}^0 \sim \pi_\theta(\mathbf{a}|\mathbf{s})$ , and  $\{Q_{\phi_i}\}_{i=1,2}$  are twin Q networks with parameters  $\phi = \{\phi_1, \phi_2\}$  and with target networks  $Q_{\phi_i'}$ . The objective of the policy improvement step is  $L(\theta) = L_{\text{behavior\_cloning}}(\theta) - \alpha \mathbb{E}_{\mathbf{s} \sim D, \mathbf{a}^0 \sim \pi_\theta} [Q_{\phi_1}(\mathbf{s}, \mathbf{a}^0)]$ , where  $\alpha$  is the trade-off hyperparameter between two losses. The behavior cloning loss is a standard supervised learning loss of DDPM, i.e. fitting a diffusion prediction model  $\pi_\theta(\cdot|\mathbf{s})$  to predict  $\mathbf{a}^0$ .

The main drawback of optimizing the objectives  $L(\phi)$  and  $L(\theta)$  is the high computational demand of having to perform a multitude of sampling steps in order to obtain  $\mathbf{a}^0 \sim \pi_\theta(\cdot|\mathbf{s})$ . Ding et. al. [13] propose using consistency models  $f_\theta(\mathbf{s}, \mathbf{a}^\tau, \tau)$  with  $\pi_\theta(\mathbf{s}) = \text{Consistency\_Inference}(\mathbf{s}, f_\theta)$  [37] to reduce the number of steps and show that they can effectively be applied for online RL.

### D. Hybrid Actor-Critic for Non-prehensile Manipulation

Our work is based on Feldman et al. [6] and HACMan [7]. Both propose to use a hybrid action space which consists of a continuous action space for motion prediction and a discrete action space for inferring contact locations, and employ similar actor-critic based network architectures. They learn an actor which predicts a per-point motion parameter map  $\mathbf{a}^m = \{\mathbf{a}_i^m = \pi_\theta(X) \mid i = 1 \dots N\}$  for a given input point cloud  $X$ , as well as a critic which determines a per-point Q-value map  $Q = \{Q_i = Q_\phi(X, \mathbf{a}_i^m) \mid i = 1 \dots N\}$  for the motion parameter of each point  $\mathbf{a}_i^m$ . Both networks share a common encoder  $f(X) = \{f_i \mid i = 1 \dots N\}$  which predicts a per-point feature map. Based on  $Q$ , a location policy  $\pi_{loc}$  then selects a discrete point  $x_i$  and, in that way, also the corresponding continuous  $\mathbf{a}_i^m$ . Feldman et. al. [6] add a max-entropy term for exploring the continuous motion prediction space, but ignore the exploration on discrete location space. HACMan [7] uses TD3 which resorts to a simple  $\epsilon$ -greedy strategy for exploring both locations and motion parameters. It computes the probability of a point being selected as the contact location by

$$\pi_{loc}(x_i | X_{obj}) = \frac{\exp(\beta Q_i)}{\sum_{k=1 \dots N} \exp(\beta Q_k)}, \quad (3)$$

where  $\beta$  is the softmax’s temperature, and  $N$  is the number of points on the object point cloud  $X_{obj}$ <sup>1</sup>.

## IV. METHODOLOGY

### A. Problem Formulation

The goal of this work is to develop policies for non-prehensile manipulation tasks, specifically targeting 6D object pose alignment. This task requires the policy to process

<sup>1</sup>Points on the background point cloud  $X_b$  determined through separate segmentation component are not considered.

a point cloud  $X$  as input, where each point is represented by its 3D coordinates, a 1D segmentation mask, and a 3D goal flow vector. To solve this problem, the policy must handle both discrete actions, such as selecting contact points, and continuous actions, like generating motion primitive vectors.

We formulate this problem in a principled manner as an online off-policy maximum entropy reinforcement learning task. This framework is chosen to encourage exploration and diversity in the learned behaviors. To represent the policies, we leverage Diffusion Probabilistic Models (DDPMs) [12] and Consistency Models (CMs) [37], which are well-suited for capturing diverse behaviors. Specifically, we first introduce a principled formulation to integrate diffusion policies into the Soft Actor-Critic (SAC) algorithm, enabling the continuous policies to capture multi-modalities. We then extend this approach to support both discrete and continuous actions within a hybrid RL framework, leading to our Hybrid Diffusion Policy (HyDo). This principled formulation enhances exploration across both action spaces, ensuring the development of diverse and robust manipulation policies.

### B. Soft Actor-Critic with Diffusion Policy

Given a policy  $\pi$  parameterized by a diffusion model defined in Eq. 2, we propose to incorporate a Diffusion Policy that optimizes an objective similar to SAC, i.e. the entropy-regularized cumulative return:

$$J_\pi(\theta) = \sum_t \mathbb{E}_{\mathbf{s}_t, \mathbf{a}_t^{0:K} \sim \pi_\theta} \left[ r(\mathbf{s}_t, \mathbf{a}_t^0) - \alpha \sum_{k=0}^K \log \pi_\theta(\mathbf{a}_t^{k-1} | \mathbf{a}_t^k, k, \mathbf{s}_t) \right], \quad (4)$$

where  $\alpha$  is a hyperparameter. The entropy term in Eq. 4 can also be interpreted as  $-\log p(\mathbf{a}_t | \mathbf{s}_t)$  of the whole sampling action diffusion path instead of only  $-\log p(\mathbf{a}_t^0 | \mathbf{s}_t)$  (with the true RL action) as in the standard SAC's objective, because it is intractable to compute the density of diffusion models. We follow the derivation of the structured variational inference [40] to prove that  $J_\pi(\theta)$  is the lower-bound of the maximum reward likelihood,

$$\log p(O_{1:T}) \geq \mathbb{E}_{\mathbf{s}_{1:T}, \mathbf{a}_1^{0:K}, \dots, \mathbf{a}_T^{0:K} \sim q(\mathbf{s}_{1:T}, \mathbf{a}_1^{0:K}, \dots, \mathbf{a}_T^{0:K})} \left[ \sum_{t=1}^T r(\mathbf{s}_t, \mathbf{a}_t^0) - \alpha \sum_{k=0}^K \log q(\mathbf{a}_t^{k-1} | \mathbf{a}_t^k, k, \mathbf{s}_t) \right].$$

where the binary random variable  $O$  denotes if time step  $t$  is optimal or not, and  $q$  is the variational distribution, in which the distribution over  $O$  is  $p(O_t | \mathbf{s}_t, \mathbf{a}_t) = \exp(\frac{1}{\alpha} r(\mathbf{s}_t, \mathbf{a}_t))$ . The proof applies Jensen's inequality as follows,

$$\begin{aligned} \log p(O_{1:T}) &= \log \int p(O_{1:T}, \tau) d\tau \\ &\geq \mathbb{E}_{\tau \sim q(\tau)} [\log p(O_{1:T}, \tau) - \log q(\tau)] \\ &= \mathbb{E}_{\tau \sim q(\tau)} \left[ \sum_{t=1}^T \left( r(\mathbf{s}_t, \mathbf{a}_t) - \alpha \sum_{k=1}^K \log q(\mathbf{a}_t^{k-1} | \mathbf{a}_t^k, \mathbf{s}_t) \right) \right]. \end{aligned}$$

where the variational policy distribution  $q(\mathbf{a}_t^{k-1} | \mathbf{a}_t^k, \mathbf{s}_t)$  is parameterized as  $\pi_\theta(\mathbf{a}_t^{k-1} | \mathbf{a}_t^k, k, \mathbf{s}_t)$ .

As a result, we obtain the policy evaluation step of the soft policy iteration with the modified soft Bellman backup operator  $T^\pi$

$$T^\pi Q(\mathbf{s}_t, \mathbf{a}_t^0) = r(\mathbf{s}_t, \mathbf{a}_t^0) + \gamma \mathbb{E}_{\mathbf{a}_{t+1}^{0:K} \sim \pi, \mathbf{s}_{t+1} \sim P} \left[ Q(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}^0) - \alpha \sum_{k=0}^K \log p_\theta(\mathbf{a}_{t+1}^{k-1} | \mathbf{a}_{t+1}^k, k, \mathbf{s}_{t+1}) \right]. \quad (5)$$

The policy improvement step updates the policy with the same objective as SAC, i.e., we minimize  $L(\theta) = \mathbb{E}_{\mathbf{s}_t \sim D} \left[ \text{D}_{\text{KL}} \left( \pi'(\cdot | \mathbf{s}_t) \parallel \frac{\exp(Q_\theta(\mathbf{s}_t, \cdot))}{Z_\theta(\mathbf{s}_t)} \right) \right]$ . Using DDPMs or CMs [13], each probability  $p_\theta(\mathbf{a}_t^{k-1} | \mathbf{a}_t^k, k, \mathbf{s}_t)$  is a Gaussian and benefits from the reparameterization trick using this transformation  $\mathbf{a}_t^{k-1} = f_\theta(\epsilon_t^{k-1}; \mathbf{a}_t^k, k, \mathbf{s}_t)$ . Thus the gradient of the actor loss can be approximated as

$$\begin{aligned} \nabla_\theta L(\theta) &= -\nabla_{\mathbf{a}_t^0} Q_\theta(\mathbf{s}_t, \mathbf{a}_t^0) \frac{\partial \mathbf{a}_t^0}{\partial \theta} \\ &\quad + \sum_{k=1}^K \nabla_{\mathbf{a}_t^{k-1}} \log p_\theta(\mathbf{a}_t^{k-1} | \mathbf{a}_t^k, k, \mathbf{s}_t) \nabla_\theta f_\theta(\epsilon_t^{k-1}; \mathbf{a}_t^k, k, \mathbf{s}_t), \end{aligned} \quad (6)$$

where the term  $\frac{\partial \mathbf{a}_t^0}{\partial \theta}$  is also computed with reparameterization trick as already used in previous methods as direct policy optimization [16], [38].

### C. Hybrid Diffusion Policy

To address the challenge of non-prehensile manipulation, we build on the formulation presented in Subsection IV-B by extending it to a hybrid setting, where the policy must handle both discrete and continuous actions. To achieve this, we first augment the HACMan [7] objective with entropy regularization terms for both location and motion policies:

$$J_i(\theta) = -Q_\phi(f_i, \mathbf{a}_i^m) + \alpha \log \pi_{\theta,i}(x_i, \mathbf{a}_i^m | \mathbf{s}),$$

where  $\log \pi_{\theta,i}(x_i, \mathbf{a}_i^m | \mathbf{s})$  includes both location entropy  $\log \pi_i^{\text{loc}}(x_i | \mathbf{s})$  and motion parameter's entropy  $\log \pi_i^m(\mathbf{a}_i^m | \mathbf{s})$ . As a result, the total objective of the actor is  $J_\theta(\theta) = \sum_i^N \pi_i^{\text{loc}}(x_i | \mathbf{s}) J_i(\theta)$ . Similarly, the critic loss is defined in Eq.1 with an addition of the maximum entropy term to the target

$$y_t = r_t + \gamma \mathbb{E}_{x_i \sim \pi_i^{\text{loc}}, \mathbf{a}_i^m \sim \pi_i^m} \left[ Q_\phi(f_i(\mathbf{s}_{t+1}, \mathbf{a}_i^m)) - \alpha \log \pi_{\theta,i}(x_i, \mathbf{a}_i^m | \mathbf{s}) \right].$$

We then introduce Hybrid Diffusion Policy (HyDo), which models the motion parameter policy  $\pi_m$  using diffusion models. In particular, diffusion policy  $\pi_m$  predicts action map  $\mathbf{a}^m$  as a denoising process

$$\begin{aligned} \pi^m(\mathbf{a}^m | \mathbf{s}) &= p_\theta(\mathbf{a}^{m,0:K} | \mathbf{s}) \\ &= N(\mathbf{a}^{m,K}; \mathbf{0}, \mathbf{I}) \prod_{k=1}^K p_\theta(\mathbf{a}^{m,k-1} | \mathbf{a}^{m,k}, \mathbf{s}), \end{aligned}$$

where we denote  $\mathbf{a}^{m,k}$  is an action map at denoising step  $k$ . As a result of applying HyDo in IV-B, the per-point loss of the actor is rewritten as

$$J_i(\theta) = -Q_\phi(f_i, \mathbf{a}^{m,0}) + \alpha_1 \log \pi_i^{\text{loc}}(x_i | \mathbf{s}) \\ + \alpha_2 \sum_{k=0}^K \log p_\theta(\mathbf{a}_i^{k-1} | \mathbf{a}_i^k, k, \mathbf{s}),$$

for  $i = 1, \dots, N$ . Thus, the total objective of the actor is  $J_\pi(\theta) = \sum_i^N \pi^{\text{loc}}(x_i | \mathbf{s}) J_i(\theta)$ , and its gradient is computed using the chain rule through the softmax of  $Q$  of the location policy in Eq. 3 and the gradient in Eq. 6. Finally, the critic is updated using a standard update in Eq. 1 with the following entropy maximization term in the target as

$$y_t = r_t + \gamma \mathbb{E} \left[ Q_\phi(f_i(\mathbf{s}_{t+1}, \mathbf{a}_i^{m,0})) - \alpha \log \pi_i^{\text{loc}}(x_i | \mathbf{s}_{t+1}) \right. \\ \left. - \alpha \sum_{k=1}^K \log p_\theta(\mathbf{a}_i^{k-1} | \mathbf{a}_i^k, k, \mathbf{s}_{t+1}) \right],$$

where the expectation is taken over both the location and motion policies,  $x_i \sim \pi^{\text{loc}}(\cdot | \mathbf{s}_{t+1})$ ,  $\mathbf{a}_i^{m,0:K} \sim \pi^m(\cdot | \mathbf{s}_{t+1})$ .

In addition, the diffusion policy can be replaced by a consistency model  $\pi^m(\mathbf{a}^m | \mathbf{s}) = \text{ConsistencyModel}(\mathbf{s}; f_\theta)$  without changing the underlying optimization procedure (as shown in Algo. 1). We name this variant, Hybrid Diffusion Policy with Consistency Models (HyDo + CM). Both variants are summarized with pseudo-code in Alg. 2.

---

#### Algorithm 1 Consistency Model Action Sampling

---

- 1: **Given:** state  $\mathbf{s}$ ,  $f_\theta$ , a sub-sequence of time points  $\{\tau_n\}_{n \in [N]}$ , diffusion steps  $K$ .
  - 2: Initialize mean  $\mathbf{a}_N = \mathbf{0}$ , variance  $\Sigma_N = K^2 \mathbf{I}$
  - 3: **for**  $n = N$  to 1 **do**
  - 4:   Sample  $\hat{\mathbf{a}}$  from  $N(\mathbf{a}_n, \Sigma_n)$
  - 5:   Compute  $c_{\text{skip}} = 0.25 / ((\tau_n - \varepsilon)^2 + 0.25)$  and  $c_{\text{out}} = 0.5(\tau_n - \varepsilon) / \sqrt{(\tau_n^2 + 0.25)}$
  - 6:   Compute  $\mathbf{a}_n, \Sigma_n = f_\theta(\mathbf{s}, \hat{\mathbf{a}}, \tau_n)$
  - 7:   Compute output  $\mathbf{a}_n = c_{\text{skip}} \cdot \hat{\mathbf{a}} + c_{\text{out}} \cdot \mathbf{a}_n$
  - 8: **end for**
  - 9: **return**  $\mathbf{a}_1$
- 

## V. EXPERIMENTS

In this section, we evaluate our proposed algorithms, HyDo and HyDo + CM, with the main baseline HACMan and its variants, HACMan + Diff, and HACMan + CM. In addition, we also investigate HyDo (w/o diffusion). All methods are evaluated on a set of simulated and real-world tasks. Our primary goal is to assess their performance in terms of success rate, behavior diversity, and generalization ability across different task settings. We use the same training settings as HACMan [7]. The input is a 4D point cloud obtained by concatenating 3D goal flow vectors with a 1D segmentation mask which indicates if the point belongs to the target object or the background. In simulation, ground-truth object masks are used as segmentation labels. In the real world robot experiments labels are obtained by background subtraction.

---

#### Algorithm 2 HyDo: Hybrid Diffusion Policy

---

- 1: Initialize policy  $\pi_\theta$  and critic networks  $Q_{\phi_1}$  and  $Q_{\phi_2}$
  - 2: Initialize the target networks:  $Q_{\phi_1'}^*$  and  $Q_{\phi_2'}^*$
  - 3: Initialize replay buffer:  $D = \emptyset$
  - 4: **while** not converge **do**
  - 5:   Forward the encoder to compute features  $f = f(\mathbf{s}_t)$
  - 6:   Sample action map  $\mathbf{a}_t^m$  from diffusion policy as in Eq. 3, e.g. DDPM or CM sampling presented in *Algo. 1*
  - 7:   Compute Q-value map  $Q = Q_\phi(f, \mathbf{a}_t^m)$
  - 8:   Select contact point  $x_i$  using location policy Eq. 4
  - 9:   Select corresponding action's motion parameter  $\mathbf{a}_{t,i}^m$
  - 10:   Execute action  $(x_i, \mathbf{a}_{t,i}^{m,0})$ , observe  $r_t$  and  $\mathbf{s}_{t+1}$
  - 11:   Add sample  $\{\mathbf{s}_t, (x_i, \mathbf{a}_{t,i}^{m,0}), r_t, \mathbf{s}_{t+1}\}$  to replay buffer  $D$
  - 12:   Sample a minibatch  $\{\mathbf{s}, (x_i, \mathbf{a}_{t,i}^0), r_t, \mathbf{s}'\}$  from  $D$
  - 13:   Critic update as with target  $y_t$  defined in IV-C.
  - 14:   Actor update with loss  $J_\pi(\theta)$  as defined in IV-C
  - 15:   Adjust temperature  $\alpha$
  - 16:   Update the target networks like SAC.
  - 17: **end while**
  - 18: **return** final policy  $\pi_\theta$ .
- 

#### A. Experimental Setup

We validate our method through comparisons and ablation studies performed on the 6D object pose alignment task and translation task introduced in HACMan [7]. This task requires diverse non-prehensile manipulations like pushing and flipping to achieve a specified goal pose. **Simulation setting:** The environment is built using Robosuite [41] and MuJoCo [42] and provides 44 different objects for alignment. The dataset is split into training (32 objects), unseen instances (7 objects), and unseen categories (5 objects). Success is defined by a mean distance of less than 3 cm between the corresponding points of the object and the goal. More details can be found in HACMan [7]. **Real robot setting:** The robot system for sim2real evaluations is equipped with a 7DoF Franka Panda arm and three static Realsense cameras. **Evaluations:** We evaluate the algorithms on training objects set. Subsequently, we test these two different configurations: i) Planar goals, where the object starts at a fixed pose and with a randomized planar translation goal pose; and ii) 6D goals, where both the initial and goal poses are stable SE(3) poses that are randomized. Specifically, for both real-world and simulation evaluations, the goals are sampled from SE(3). Initially, object poses are sampled from SE(3) in the air above the center of the bin. The object is dropped into the bin and allowed to settle into a stable position, which is recorded as the goal pose. In simulation, 100 stable poses are collected for each object, and at the start of each episode, a goal is randomly selected from these poses. The location of the selected stable pose is then randomized within the bin.

#### B. Experimental Results

1) *Simulation Results:* Tab. I presents the evaluation results of the experiments conducted in simulation for 6D pose tasks on unseen category, unseen instance objects, as well as

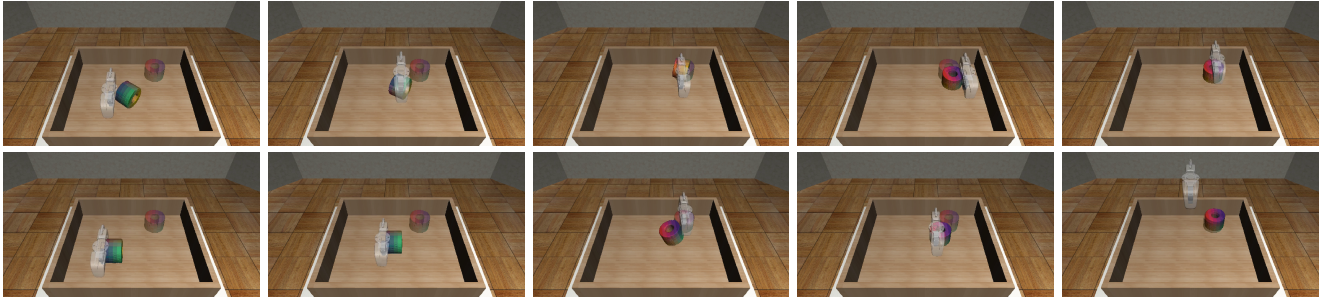


Fig. 2: A simulation task showcases the multi-modalities of action sequences, **(top)** Push→ Push → Push→ Flip→ Push; **(bottom)** Push→ Push → Flip→ Push→ Push. In this task, we fixed the goal and initial pose and generated the action sequences with two different random seeds.

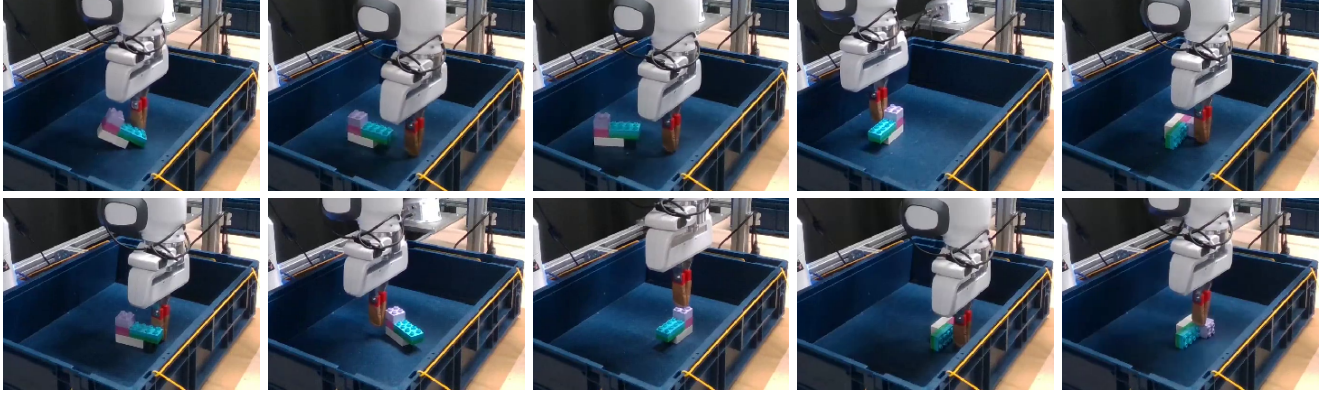


Fig. 3: A real robot task showcases the multi-modalities of action sequences, **(top)** Push→ Push → Push→ Push→ Flip; **(bottom)** Push→ Push → Push→ Flip→ Push. In this task, we fixed the goal and initial pose and generated the action sequences with two different random seeds.

the objects in the training set. In this evaluation, we report the interquartile mean (IQM) [43] success rate over 10 seeds and using the best checkpoint for each method.

TABLE I: Generalization performance on the 6D pose simulation task (using IQM with 95% confidence intervals).

Method	Unseen Category	Unseen Instance	Training Objects
HACMan	0.760 ± 0.042	0.818 ± 0.049	0.769 ± 0.062
HACMan + Diff	0.728 ± 0.041	0.780 ± 0.028	0.712 ± 0.047
HACMan + CM	0.671 ± 0.123	0.703 ± 0.130	0.632 ± 0.128
HyDo (w/o Diff)	0.816 ± 0.026	0.848 ± 0.032	0.794 ± 0.028
HyDo	<b>0.843 ± 0.043</b>	<b>0.884 ± 0.046</b>	<b>0.814 ± 0.044</b>
HyDo + CM	0.827 ± 0.034	0.861 ± 0.030	0.794 ± 0.038

Overall, these results indicate a substantial performance improvement with the introduction of the entropy term as a regularization. Specifically, HyDo and its variants consistently outperform the HACMan baselines on the *Unseen Category* and *Unseen Instance* object sets. The difference in performance is particularly highlighted in diffusion-based policies (+ Diff or + CM), where HyDo and HyDo + CM with the additional entropy regularization terms yields up to a 10% and 15% improvement compared to HACMan + Diff and HACMan + CM, respectively. In addition, either with or without the entropy terms, the diffusion policies achieve slightly better performance than consistency models across all evaluation sets.

These findings demonstrate the effectiveness of entropy regularization and underscore its critical role in optimizing complex policies, such as diffusion and consistency models, for improved generalization in unseen scenarios.

2) *Real Robot Results:* We evaluate the trained policies on the "All Objects + 6D Goals" simulation task on a real-world robot for which we use the same real-world setup, pose randomization, and success criteria as in HACMan (see Fig. 4).



(a) Lego (b) Lotion (c) Milk (d) Soja (e) Cube

Fig. 4: Set of 5 objects used for real robot evaluations.

TABLE II: Real Robot Experiments for Planar (left) and 6D (right) Goals.

Object	HACMan		HyDo (w/o Diff)		HyDo		HyDo + CM	
Lego	6/10	6/10	8/10	5/10	8/10	6/10	9/10	7/10
Lotion	6/10	5/10	7/10	6/10	8/10	7/10	7/10	7/10
Milk	4/10	6/10	8/10	6/10	7/10	7/10	8/10	7/10
Soja	5/10	5/10	6/10	4/10	6/10	6/10	7/10	5/10
Cube	5/10	5/10	8/10	6/10	8/10	5/10	9/10	6/10
<b>Total</b>	<b>26/50</b>	<b>27/50</b>	<b>37/50</b>	<b>27/50</b>	<b>37/50</b>	<b>31/50</b>	<b>40/50</b>	<b>32/50</b>

Tab. II reports the evaluation results on both the planar and 6D tasks. We run each method with ten object trials and randomize the initial and target poses. To make it fair, we keep them similar across evaluations for all methods. The average success rates of the four methods are 53% for HACMan, 64% HyDo (w/o Diffusion), 68% for HyDo, and 72% for HyDo + CM. These results demonstrate that all methods are also capable of effectively generalizing to unseen objects in real world. We do, however, observe a gap between non-diffusion and diffusion policy methods which is larger than in simulation. The gap between simulation and real world scenes is generally larger than the one for simulated training to simulated evaluation scenarios. Our hypothesis is that the diverse behaviors of the learned multi-modal diffusion policies allow for better generalization to such new environments. To further assess this diversity, we evaluate multi-modal action sequences under the same environment conditions, differing only by action sampling seeds. Fig. 2 and Fig. 3 show two distinct action sequences achieving the same goal pose in both simulation and real robot experiments, demonstrating how the order of flips and pushes can vary, achieving the same outcome through diverse execution actions.

### C. Qualitative Policy Diversity Analysis

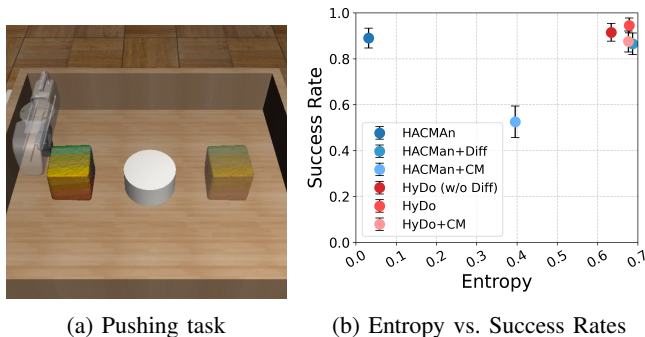


Fig. 5: Diverse behaviors of different diffusion-based policies on a simple pushing task (a). In (b), we show the Pareto plot between Entropy and Success Rates. Overall, methods without entropy regularization (HACMan) result in lower entropy compared to its counterparts. Note that we disable the entropy term on the location action to solely investigate the effect of the entropy on the continuous policies.

We follow the proposal to analyze the policy diversity as defined in [10]. They approximate behavior entropy of diffusion policies  $\pi$  with the following function  $H(\pi(\beta)) = -\sum_{\beta \in B} \pi(\beta) \log_{|\beta|} \pi(\beta)$ , where  $B$  is defined as the set of task-specific behavior descriptors. We run simulations using the final policy of each method to estimate  $\pi(\beta)$ . We design a controlled task with  $|\beta| = 2$  where the robot can push the target object (on the left) along the upper or below path around the white static obstacle object to the target pose (on the right) as depicted in Fig. 5a. If the entropy  $H(\pi(\beta)) = 0$ , it indicates that the policy has collapsed to a single solution or a single path (either upper or lower).

We qualitatively evaluate only methods with policies represented by diffusion and consistency models. More specifically, we train HyDo, HyDo + CM, HACMan + Diffusion, and HACMan + CM until convergence with random initial poses. Then, we evaluate them for 200 trials with the same initial pose depicted in Fig. 5a. In this experiment, to highlight the effect of the entropy on the continuous motion parameter, we disable the entropy term for the location action. As shown in Fig. 5b, methods without entropy regularization such as HACMan, and HACMan + CM result in less diverse behaviors (low entropy) and/or non-optimal actions (low success rate).

### D. Computational Efficiency Analysis of Consistency and Diffusion Models

To assess the computational speed of consistency and diffusion models with varying denoising steps, we run an experiment measuring inference times for  $K \in \{2, 5, 10, 20, 50\}$ , using a single-object "Hammer" training setup. As shown in Tab. III, the consistency model [37] achieves similar performance to the diffusion model with fewer denoising steps, demonstrating its efficiency advantage. Both models reach a performance plateau around  $K = 5$ , with further increases in  $K = 50$  leading to a drop in performance.

TABLE III: Inference Time in milliseconds per sample for each method and diffusion step. (using IQM with 95% confidence intervals)

Method	$K$	Inference Time (in ms)	Success Rate
HyDo	2	$3.54 \pm 0.60$	$0.631 \pm 0.041$
HyDo	5	$7.90 \pm 1.06$	$0.684 \pm 0.038$
HyDo	10	$17.01 \pm 2.66$	$0.646 \pm 0.059$
HyDo	20	$29.27 \pm 3.67$	$0.677 \pm 0.045$
HyDo	50	$74.23 \pm 10.25$	$0.644 \pm 0.025$
HyDo + CM	2	$3.23 \pm 0.69$	$0.684 \pm 0.043$
HyDo + CM	5	$7.51 \pm 1.14$	$0.787 \pm 0.077$
HyDo + CM	10	$14.89 \pm 1.77$	$0.731 \pm 0.007$
HyDo + CM	20	$26.72 \pm 3.65$	$0.713 \pm 0.058$
HyDo + CM	50	$67.70 \pm 9.93$	$0.575 \pm 0.009$

## VI. CONCLUSIONS

We presented Hybrid Diffusion Policy (HyDo), an online diffusion-based off-policy maximum entropy RL algorithm for 6D non-prehensile manipulation. We derived a principled objective, i.e. the maximum entropy regularization, that considers diffusion policies as a class of stochastic policies. We showed that treating the stochastic diffusion policy with a principled objective significantly improves its performance in RL applications. Our qualitative results indicated that online RL is hard for learning multi-modal policy distributions with diffusion models, as it can make diffusion policies converge to uni-modal quickly. Therefore, stochastic diffusion-based and entropy maximizing RL algorithms can be a promising combination for improved exploration strategies and learning more diversity behaviors. For future work, we envision

extending our approach to more complex, dynamic environments, such as closed-loop settings and tasks requiring continuous adaptation [44].

## REFERENCES

- [1] K.-T. Yu, M. Bauza, N. Fazeli, and A. Rodríguez, “More than a million ways to be pushed. a high-fidelity experimental dataset of planar pushing,” in *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2016, pp. 30–37.
- [2] Z. Xu, Z. He, and S. Song, “Universal manipulation policy network for articulated objects,” *IEEE robotics and automation letters*, vol. 7, no. 2, pp. 2447–2454, 2022.
- [3] X. Cheng, E. Huang, Y. Hou, and M. T. Mason, “Contact mode guided motion planning for quasidynamic dexterous manipulation in 3d,” in *2022 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2730–2736.
- [4] Y. Hou and M. T. Mason, “Robust execution of contact-rich motion plans by hybrid force-velocity control,” in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.
- [5] K. Mo, L. J. Guibas, M. Mukadam, A. Gupta, and S. Tulsiani, “Where2act: From pixels to actions for articulated 3d objects,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6813–6823.
- [6] Z. Feldman, H. Ziesche, N. A. Vien, and D. D. Castro, “A hybrid approach for learning to shift and grasp with elaborate motion primitives,” in *2022 International Conference on Robotics and Automation, ICRA*. IEEE, 2022, pp. 6365–6371.
- [7] W. Zhou, B. Jiang, F. Yang, C. Paxton, and D. Held, “HACMan: Learning hybrid actor-critic maps for 6d non-prehensile manipulation,” in *Conference on Robot Learning (CoRL)*, vol. 229. PMLR, 2023.
- [8] J. Merel, L. Hasenclever, A. Galashov, A. Ahuja, V. Pham, G. Wayne, Y. W. Teh, and N. Heess, “Neural probabilistic motor primitives for humanoid control,” in *International Conference on Learning Representations*, 2018.
- [9] S. Kumar, A. Kumar, S. Levine, and C. Finn, “One solution is not all you need: Few-shot extrapolation via structured maxent rl,” *Advances in Neural Information Processing Systems*, 2020.
- [10] X. Jia, D. Blessing, X. Jiang, M. Reuss, A. Donat, R. Lioutikov, and G. Neumann, “Towards diverse behaviors: A benchmark for imitation learning with human demonstrations,” in *The Twelfth International Conference on Learning Representations*, 2024.
- [11] Y. Song and S. Ermon, “Generative modeling by estimating gradients of the data distribution,” *Advances in neural information processing systems*, vol. 32, 2019.
- [12] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, 2020.
- [13] Z. Ding and C. Jin, “Consistency models as a rich and efficient policy class for reinforcement learning,” in *The Twelfth International Conference on Learning Representations*, 2023.
- [14] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, “Diffusion policy: Visuomotor policy learning via action diffusion,” *arXiv preprint arXiv:2303.04137*, 2023.
- [15] T. Haaroja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,” in *International conference on machine learning*. PMLR, 2018, pp. 1861–1870.
- [16] Z. Wang, J. J. Hunt, and M. Zhou, “Diffusion policies as an expressive policy class for offline reinforcement learning,” in *The Eleventh International Conference on Learning Representations*, 2023.
- [17] P. Hansen-Estruch, I. Kostrikov, M. Janner, J. G. Kuba, and S. Levine, “Idql: Implicit q-learning as an actor-critic method with diffusion policies,” *arXiv preprint arXiv:2304.10573*, 2023.
- [18] C. Lu, H. Chen, J. Chen, H. Su, C. Li, and J. Zhu, “Contrastive energy prediction for exact energy-guided diffusion sampling in offline reinforcement learning,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 22825–22855.
- [19] S. Venkatraman, S. Khaitan, R. T. Akella, J. Dolan, J. Schneider, and G. Berseth, “Reasoning with latent diffusion in offline reinforcement learning,” in *The Twelfth International Conference on Learning Representations*, 2024.
- [20] Z. Li, R. Krohn, T. Chen, A. Ajay, P. Agrawal, and G. Chaitzaki, “Learning multimodal behaviors from scratch with diffusion policy gradient,” in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [21] M. Reuss, M. Li, X. Jia, and R. Lioutikov, “Goal-conditioned imitation learning using score-based diffusion policies,” 2023.
- [22] T. Pearce, T. Rashid, A. Kanervisto, D. Bignell, M. Sun, R. Georgescu, S. V. Macua, S. Z. Tan, I. Momennejad, K. Hofmann *et al.*, “Imitating human behaviour with diffusion models,” *arXiv preprint arXiv:2301.10677*, 2023.
- [23] H. Chen, C. Lu, C. Ying, H. Su, and J. Zhu, “Offline reinforcement learning via high-fidelity generative behavior modeling,” in *The Eleventh International Conference on Learning Representations*, 2022.
- [24] L. He, L. Zhang, J. Tan, and X. Wang, “Diffcps: Diffusion model based constrained policy search for offline reinforcement learning,” *arXiv preprint arXiv:2310.05333*, 2023.
- [25] J. Peters and S. Schaal, “Reinforcement learning by reward-weighted regression for operational space control,” in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 745–750.
- [26] K. Lee, H. Liu, M. Ryu, O. Watkins, Y. Du, C. Boutilier, P. Abbeel, M. Ghavamzadeh, and S. S. Gu, “Aligning text-to-image models using human feedback,” *arXiv preprint arXiv:2302.12192*, 2023.
- [27] K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine, “Training diffusion models with reinforcement learning,” in *The Twelfth International Conference on Learning Representations*, 2024.
- [28] M. Uehara, Y. Zhao, K. Black, E. Hajiramezani, G. Scalia, N. L. Diamant, A. M. Tseng, S. Levine, and T. Biancalani, “Feedback efficient online fine-tuning of diffusion models,” in *International Conference on Machine Learning (ICML)*, 2024.
- [29] M. Uehara, Y. Zhao, K. Black, E. Hajiramezani, G. Scalia, N. L. Diamant, A. M. Tseng, T. Biancalani, and S. Levine, “Fine-tuning of continuous-time diffusion models as entropy-regularized control,” *CoRR*, vol. abs/2402.15194, 2024.
- [30] M. Psenka, A. Escontrela, P. Abbeel, and Y. Ma, “Learning a diffusion model policy from rewards via q-score matching,” *arXiv preprint arXiv:2312.11752*, 2023.
- [31] W. Zhou and D. Held, “Learning to grasp the ungraspable with emergent extrinsic dexterity,” in *Conference on Robot Learning*. PMLR, 2023, pp. 150–160.
- [32] B. Jiang, Y. Wu, W. Zhou, C. Paxton, and D. Held, “Hacman++: Spatially-grounded motion primitives for manipulation,” 2024.
- [33] P. M. Scheikl, N. Schreiber, C. Haas, N. Freymuth, G. Neumann, R. Lioutikov, and F. Mathis-Ullrich, “Movement primitive diffusion: Learning gentle robotic manipulation of deformable objects,” *CoRR*, vol. abs/2312.10008, 2023.
- [34] G. Li, Z. Jin, M. Volpp, F. Otto, R. Lioutikov, and G. Neumann, “Prodmp: A unified perspective on dynamic and probabilistic movement primitives,” *IEEE Robotics and Automation Letters*, 2023.
- [35] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [36] S. Fujimoto, H. Hoof, and D. Meger, “Addressing function approximation error in actor-critic methods,” in *International conference on machine learning*. PMLR, 2018, pp. 1587–1596.
- [37] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, “Consistency models,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 32211–32252.
- [38] B. Kang, X. Ma, C. Du, T. Pang, and S. Yan, “Efficient diffusion policies for offline reinforcement learning,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [39] S. Fujimoto and S. S. Gu, “A minimalist approach to offline reinforcement learning,” *Advances in neural information processing systems*, vol. 34, pp. 20132–20145, 2021.
- [40] S. Levine, “Reinforcement learning and control as probabilistic inference: Tutorial and review,” *arXiv preprint arXiv:1805.00909*, 2018.
- [41] Y. Zhu, J. Wong, A. Mandelkar, R. Martín-Martín, A. Joshi, S. Nasiriany, and Y. Zhu, “robosuite: A modular simulation framework and benchmark for robot learning,” *arXiv preprint arXiv:2009.12293*, 2020.
- [42] E. Todorov, T. Erez, and Y. Tassa, “Mujoco: A physics engine for model-based control,” in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 5026–5033.
- [43] R. Agarwal, M. Schwarzer, P. S. Castro, A. C. Courville, and M. Bellemare, “Deep reinforcement learning at the edge of the statistical precipice,” *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [44] Y. Cho, J. Han, Y. Cho, and B. Kim, “CORN: Contact-based Object Representation for Nonprehensile Manipulation of General Unseen Objects,” in *International Conference on Learning Representations (ICLR)*, 2024.