Diff-HySAC: Diffusion-Based Hybrid Soft Actor-Critic for 6D Non-Prehensile Manipulation

Huy Le * [1,2] **Miroslav Gabriel**[1] **Tai Hoang**[2] **Gerhard Neumann**[2] **Ngo Anh Vien** [1]
[1] Bosch Center for Artificial Intelligence
[2] Autonomous Learning Robots, Karlsruhe Institute of Technology

Abstract: Learning diverse policies for non-prehensile manipulation of objects can potentially improve skill transfer and generalization to out-of-distribution scenarios and unseen objects. In this work, we propose an innovative approach to learning versatile 6D non-prehensile manipulation policies by introducing a new objective function based on entropy maximization terms. This allows for simultaneous exploration of discrete and continuous action spaces, such as contact location and motion parameter spaces. To further enhance the diversity of the agent's policy, we represent a continuous motion parameter policy as a diffusion model and derive the maximum entropy objective for optimizing diffusion policies as the lower bound of the maximum reward likelihood using structured variational inference. As a result, we introduce the hybrid soft actor-critic with diffusion policy algorithm (Diff-HySAC). We evaluate the benefit of adding maximum entropy regularization and diffusion on both simulation and zero-shot sim2real tasks. Results show that this combination helps learn more diverse behavior policies. The largest improvements we obtain are for zero-shot sim2real transfer on a 6D object pose alignment task where the success rate increases from 53% to 72%.

Keywords: diffusion policies, non-prehensile manipulation, soft actor-critic

1 Introduction

The ability to manipulate objects beyond simple grasping is critical to human dexterity and essential for tasks ranging from daily activities to complex industrial processes. Enabling robots to achieve such dexterity remains a significant challenge in robotics [1, 2]. While previous work has made advances, challenges in object generalization and complex motion persist [3, 4, 5]. Motion primitives (MPs) and object-centric action representations are often employed to simplify action representations and reduce sample complexity, with Reinforcement Learning (RL) used to learn these skills, especially within hybrid action spaces combining discrete contact points and continuous MPs [6, 7]. To improve generalization and skill transfer, we introduce HySAC, an off-policy hybrid soft actor-critic algorithm with maximum entropy objectives across discrete and continuous actions.

Diffusion models have shown promise for learning multi-modal policies in RL by generating actions through structured denoising steps, which encourage diverse behavior [8]. Diffusion Probabilistic Models (DDPMs) excel at modeling complex, multi-modal distributions, valuable for RL tasks with diverse action spaces. While they've been explored in offline RL [9, 10], their application in online RL remains limited. In this work, we propose an approach to learning versatile and diverse manipulation policies by introducing a maximum entropy objective for both discrete and continuous actions, and a novel off-policy hybrid soft actor-critic algorithm, HySAC. Our second contribution is DiffSAC, a soft actor-critic algorithm optimized for diffusion-based policies, incorporating entropy maximization as a lower bound through structured variational inference. Building on this,

[*] correspondence to baohuy.le@de.bosch.com

8th Conference on Robot Learning (CoRL 2024), Munich, Germany.

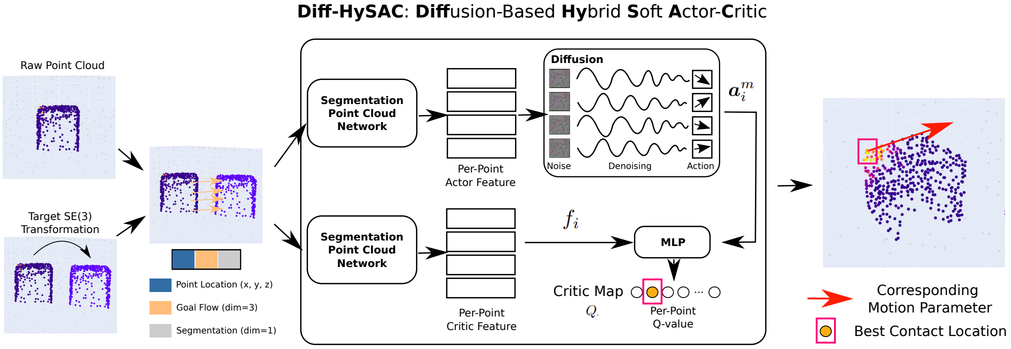

Figure 1: Diff-HySAC overview. The input to the policy is a point cloud where each point includes a 3D location, a 1D segmentation mask and a 3D goal flow vector. Diff-HySAC outputs motion primitive parameters along with the optimal contact location.

we propose Diff-HySAC, a hybrid diffusion-based soft actor-critic algorithm for 6D non-prehensile manipulation, as illustrated in Fig. 1.

2 Methodology

Problem Statement We define the non-prehensile manipulation task as a Markov decision process (MDP) represented by the tuple $\{\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma, \mathcal{P}_0\}$ [11], where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $\mathcal{R}$ the reward function, $\mathcal{P}$ the transition probabilities, γ the discount factor, and $\mathcal{P}_0$ the initial state distribution. The goal is to maximize the cumulative reward $R_t = \sum_{i=0}^{\infty} \gamma^i r(s_{t+i}, a_{t+i})$.

The policy receives input as a point cloud X, where each point consists of 3D location, a 1D segmentation mask, and a 3D goal flow vector. This setup allows hybrid action spaces that combine discrete contact locations with continuous motion parameters, which we address using our proposed HySAC and Diff-HySAC algorithms for effective exploration and diverse policy learning.

2.1 Soft Actor-Critic with Diffusion Policy

We propose DiffSAC, a variant of SAC designed to optimize policies parameterized by diffusion models. The diffusion policy π_θ is represented as a diffusion process $\pi_\theta(a|s)$. DiffSAC optimizes the following objective: $J_\pi(\theta) = \sum_t \mathbb{E}_{\boldsymbol{s}_t, \boldsymbol{a}_t^{0:K} \sim \pi_\theta}\left[r(\boldsymbol{s}_t, \boldsymbol{a}_t^0) - \alpha \sum_{k=0}^{K} \log \pi_\theta(\boldsymbol{a}_t^{k-1}|\boldsymbol{a}_t^k, k, \boldsymbol{s}_t)\right]$ where α is the entropy regularization coefficient, a^0 is the action executed by the agent (sampled at step 0), and the diffusion chain includes K steps. The entropy term can be interpreted as $-\log p(\boldsymbol{a}_t|\boldsymbol{s}_t)$ for the entire diffusion-sampled action path, instead of just $-\log p(\boldsymbol{a}_t^0|\boldsymbol{s}_t)$ as in standard SAC, because computing the density of diffusion models is intractable. We follow the derivation of structured variational inference [12] to show that $J_\pi(\theta)$ forms a lower bound on the maximum reward likelihood (proof is provided in Appendix A). Using DDPMs or Consistency models [13], each probability $p_\theta(\boldsymbol{a}_t^{k-1}|\boldsymbol{a}_t^k, k, \boldsymbol{s}_t)$ is a Gaussian and benefits from the reparameterization trick: $\boldsymbol{a}_t^{k-1} = f_\theta(\epsilon_t^{k-1}; \boldsymbol{a}_t^k, k, \boldsymbol{s}_t)$. Thus, the gradient of the actor loss is approximated as:

$$\nabla_\theta \mathcal{L}(\theta) = -\nabla_{\boldsymbol{a}_t^0} Q_\phi(\boldsymbol{s}_t, \boldsymbol{a}_t^0) \frac{\partial \boldsymbol{a}_t^0}{\partial \theta} + \sum_{k=1}^{K} \nabla_{\boldsymbol{a}_t^{k-1}} \log p_\theta(\boldsymbol{a}_t^{k-1}|\boldsymbol{a}_t^k, k, \boldsymbol{s}_t) \nabla_\theta f_\theta(\epsilon_t^k; \boldsymbol{a}_t^k, k, \boldsymbol{s}_t),$$

where the term $\frac{\partial \boldsymbol{a}_t^0}{\partial \theta}$ is also computed using the reparameterization trick, as in previous direct policy optimization methods [9, 10].

2.2 Hybrid Soft Actor-Critic with Standard Policy and Diffusion Policy

We introduce two hybrid actor-critic algorithms for non-prehensile manipulation: Hybrid Soft Actor-Critic with a standard policy representation (HySAC) and Hybrid Soft Actor-Critic with a

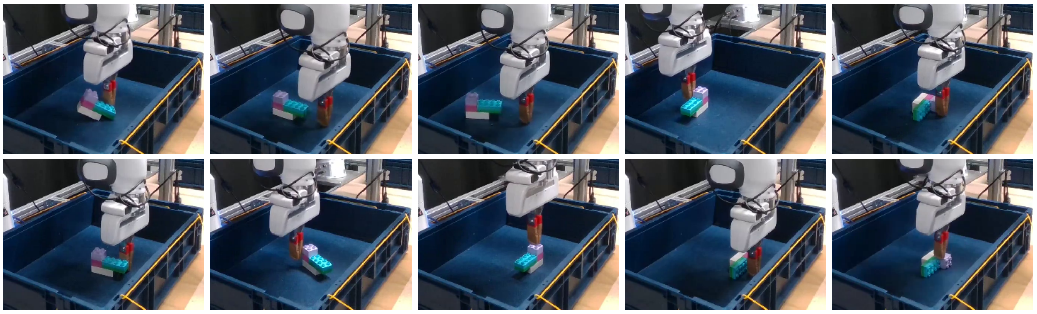

Figure 2: A real robot task showcases the multi-modalities of action sequences, **(top)** Push→ Push → Push→ Push→ Flip; **(bottom)** Push→ Push → Push→ Flip→ Push. In this task, we fixed the goal and initial pose and generated the action sequences with two different random seeds.

diffusion policy (Diff-HySAC). HySAC extends SAC to handle hybrid action spaces, while Diff-HySAC builds on DiffSAC to incorporate diffusion-based policies (see Section 2.1).

HySAC: HySAC adapts SAC by incorporating both discrete and continuous actions. The actor loss includes entropy terms for both location and motion policies:

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

where $\log \pi_{\theta,i}(x_i, \boldsymbol{a}_i^m|\boldsymbol{s})$ includes both location entropy $\log \pi_i^{\mathrm{loc}}(x_i|\boldsymbol{s})$ and motion entropy $\log \pi_i^{\mathrm{m}}(\boldsymbol{a}_i^m|\boldsymbol{s})$. The critic update is: $y_t = r_t + \gamma \mathbb{E}_{x_i \sim \pi_{\mathrm{loc}}, \boldsymbol{a}_i^m \sim \pi_{\mathrm{m}}}[Q_\phi(f_i(\boldsymbol{s}_{t+1}, \boldsymbol{a}_i^m)) - \alpha \log \pi_{\theta,i}(x_i, \boldsymbol{a}_i^m|\boldsymbol{s})]$.

Diff-HySAC: Diff-HySAC uses a diffusion model to generate action maps $\boldsymbol{a}^m$ through a denoising process $\pi^m(\boldsymbol{a}^m|\boldsymbol{s})$, α_1 and α_2 are hyperparameter. The actor loss for diffusion models is:

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

This approach can incorporate consistency models, as proposed by [13], to improve inference speed, leading to Con-HySAC. The diffusion policy can be replaced by a consistency model $\pi^m(\boldsymbol{a}^m|\boldsymbol{s}) =$ `Consistency_Model`$(\boldsymbol{s}; f_\theta)$ without changing the optimization process. Both algorithm variants are shown in Appendix B.

3 Experiments

3.1 Experimental Setup

We evaluate HySAC, Diff-HySAC, and Con-HySAC against HACMan [7], along with two baselines: HybridDiff-TD3 and HybridCon-TD3. These baselines use TD3 but replace the motion parameter policy with a diffusion or consistency model, respectively. All methods share the same training setup, using a 4D point cloud input composed of 3D goal flow vectors and a 1D segmentation mask, which labels points as part of the target object or background. In simulations, ground-truth masks are used, while real-world experiments employ background subtraction for segmentation.

Table 1: Simulation experiment: Generalization to unseen objects, reported as mean *success rate* and std. Results are averaged over 200 runs for unseen category and instance evaluations.

Method	Unseen Category	Unseen Instance
HACMan	0.80 ± 0.06	0.84 ± 0.05
HySAC	0.76 ± 0.06	0.86 ± 0.05
HybridDiff-TD3	0.78 ± 0.06	0.81 ± 0.06
Diff-HySAC	0.81 ± 0.05	0.89 ± 0.04
HybridCon-TD3	0.75 ± 0.06	0.76 ± 0.06
Con-HySAC	**0.86 ± 0.05**	**0.93 ± 0.04**

Task Setting: We validate our method on the 6D object pose alignment task introduced in HACMan [7], which requires non-prehensile manipulations such as pushing and flipping. The simulation environment, built with Robosuite [14] and MuJoCo [15], consists of 44 objects split into 32 training objects, 7 unseen instances, and 5 unseen categories. Success is measured by achieving a mean distance of less than 3 cm between corresponding points of the object and the goal.

Policy Diversity: To assess the diversity of the learned policies, we evaluate Con-HySAC in terms of action sequence variety. We keep the initial pose and goal the same across two runs with identical environment settings, such as point cloud sampling. As shown in Fig. 2, Con-HySAC produces two distinct action sequences to reach the goal pose. The order of flipping and pushing actions differs between the two, demonstrating policy flexibility in action.

Real Robot Setting: For sim-to-real transfer, we evaluate the trained policies on a 7DoF Franka Panda arm with three static Realsense cameras. We test two configurations: (i) Planar goals, where the object starts in a fixed pose and the goal involves a planar translation, and (ii) 6D goals, where both the initial and goal poses are stable $SE(3)$ poses.

3.2 Experimental Results

Simulation Results: Table 1 presents results for 6D tasks, including unseen categories, and unseen instances. Diff-HySAC and Con-HySAC outperform HACMan and HySAC, in unseen category and instance evaluations. HybridDiff-TD3 and HybridCon-TD3, lacking entropy regularization, suffer from mode collapse and perform worse than HACMan.

Real Robot Results: We evaluated the trained policies on the "All Objects + 6D Goals" simulation task using a real-world robot with the same setup, pose randomization, and success criteria as HACMan. The evaluation includes 5 objects: Lego, Lotion, Milk, Soja, and Cube. Diff-HySAC and Con-HySAC achieved success rates of 68% and 72%, respectively, outperforming HACMan (53%) and HySAC (64%). The performance gap between diffusion-based methods and non-diffusion methods was more pronounced in real-world experiments compared to simulations, suggesting that diffusion-based policies generalize better in real-world settings.

Table 2: Results for real robot experiments on Planar goal (left) and 6D goal (right) tasks.

Object	HACMan		HySAC (ours)		Diff-HySAC (ours)		Con-HySAC (ours)	
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
Total	26/50	27/50	37/50	27/50	37/50	31/50	**40/50**	**32/50**

4 Conclusion

We present Hybrid Soft Actor-Critic with Diffusion Policy (Diff-HySAC), an online diffusion-based off-policy maximum entropy RL algorithm for 6D non-prehensile manipulation. We derive a principled objective, the maximum entropy regularization, that treats diffusion policies as a class of stochastic policies. Our results show that incorporating this objective improves the performance of diffusion-based policies in RL applications. Our qualitative analysis highlights that online RL can be challenging for learning multi-modal policy distributions, as diffusion policies tend to converge to uni-modal solutions. Therefore, combining stochastic diffusion-based methods with entropy-maximizing RL algorithms holds promise for better exploration and learning more diverse behaviors. Future work will focus on enhancing the expressiveness and multi-modal capabilities of diffusion models for exploration strategies and on improving entropy estimation for diffusion model densities to further advance off-policy learning techniques.

References

[1] K.-T. Yu, M. Bauza, N. Fazeli, and A. Rodriguez. 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)*, pages 30–37. IEEE, 2016.

[2] Z. Xu, Z. He, and S. Song. Universal manipulation policy network for articulated objects. *IEEE robotics and automation letters*, 7(2):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)*, pages 2730–2736. IEEE, 2022.

[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)*, pages 1933–1939. 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*, pages 6813–6823, 2021.

[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 2022, Philadelphia, PA, USA, May 23-27, 2022*, pages 6365–6371. IEEE, 2022.

[7] W. Zhou, B. Jiang, F. Yang, C. Paxton, and D. Held. HACMan: Learning hybrid actor-critic maps for 6d non-prehensile manipulation. In J. Tan, M. Toussaint, and K. Darvish, editors, *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, volume 229 of *Proceedings of Machine Learning Research*, pages 241–265. PMLR, 2023. URL https://proceedings.mlr.press/v229/zhou23a.html.

[8] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[9] 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.

[10] B. Kang, X. Ma, C. Du, T. Pang, and S. Yan. Efficient diffusion policies for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[11] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[12] S. Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

[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] Y. Zhu, J. Wong, A. Mandlekar, 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.

[15] 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*, pages 5026–5033. IEEE, 2012.