Hybrid Learning-based Control for Reliable Robotic Arm Manipulation

Keya Sinha, Mandar Joshi, Avani Baranwal, Agniva Banerjee and Arijit Sen

***Abstract*— This study presents a hybrid learning-based control framework for robotic arm manipulation that handles sparse rewards and kinematic constraints by integrating Deep Reinforcement Learning (DRL) with Hindsight Experience Replay (HER) and Model Predictive Control (MPC) separately. The method pairs HER's sample-efficient exploration with MPC's trajectory optimization and safety within standard RL algorithms. Simulation results using the PandaReach-v3 environment demonstrate that the hybrid approach achieves a higher success rate and maintains constraint satisfaction with improved sample efficiency for robotic arm manipulators.**

I. Introduction

Robotic manipulators mainly involves reaching, gripping and pick-and-place tasks [1], which are used in industrial automation [2], logistics [3] and assistive robotics [4]. These tasks require continuous, precise control across different degrees of freedom [5]. Reinforcement learning (RL) is introduced as a promising approach to tackle the uncertainty in the environment by enabling sequential decision-making, which makes it well-suited for learning through interaction [6]. Traditional RL methods, such as Deep Q-learning (DQN), fail in continuous spaces because they must evaluate all actions in the discrete domain [7]. Continuous-control methods such as Deep Deterministic Policy Gradient (DDPG) map states to joint actions but suffer from instability and overestimation [8]. Twin Delayed DDPG (TD3) improves stability with twin critics and delayed updates [9]. Similarly, soft actor-critic (SAC) promotes exploration through entropy maximization [10]. These methods learn continuous control but struggle in sparse reward settings where binary success signals limit learning. HER converts failed attempts into useful feedback by replacing desired goals with achieved states [11]. Real-world deployment presents additional challenges, as large-scale exploration is often unsafe for physical robots. However, MPC helps by optimizing short-horizon trajectories that maintain safety and respect constraints [12].

Inspired by the existing challenges in robotic arm manipulations, this work presents a hybrid approach that integrates continuous-control RL algorithms with complementary components, such as HER and MPC, to enhance overall performance. In which HER provides dense learning signals, whereas MPC adds safety and smooth trajectories. Simulations in the PandaReach-v3 environment (Fig. 1) demonstrate that these hybrid approaches offer strong theoretical and practical validation, showing effective learning performance suitable for real-world deployment.

Department of Electrical Engineering and Computer Science, IISER Bhopal, India, keya23, mandar23, avani23, agniva24, ajsen@iiserb.ac.in

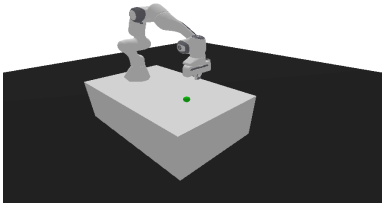


Fig. 1: PandaReach-v3 environment showing the robotic arm reaching toward a target goal point.

II. Methodology

Two complementary hybridization strategies are used to address the sparse rewards and safety constraints in robotic manipulation. Off-policy RL algorithms utilize a universal value function approximator, where the state vector concatenates the observation, achieved goal, and desired goal [13]. Training emphasizes efficient data reuse through HER, which relabels failed attempts by substituting achieved states as alternative goals. Both raw and relabeled transitions are used to fill the replay buffer, and multiple optimization steps are performed per episode, improving sample efficiency. Task performance is measured through the win percentage defined by the end-effector reaching within the PandaReach-v3 goal tolerance. The win percentage is computed as the average success rate over the last 100 episodes. After each episode, the environment provides a binary output (either success or failure) [14]. The win percentage at episode t is calculated as:

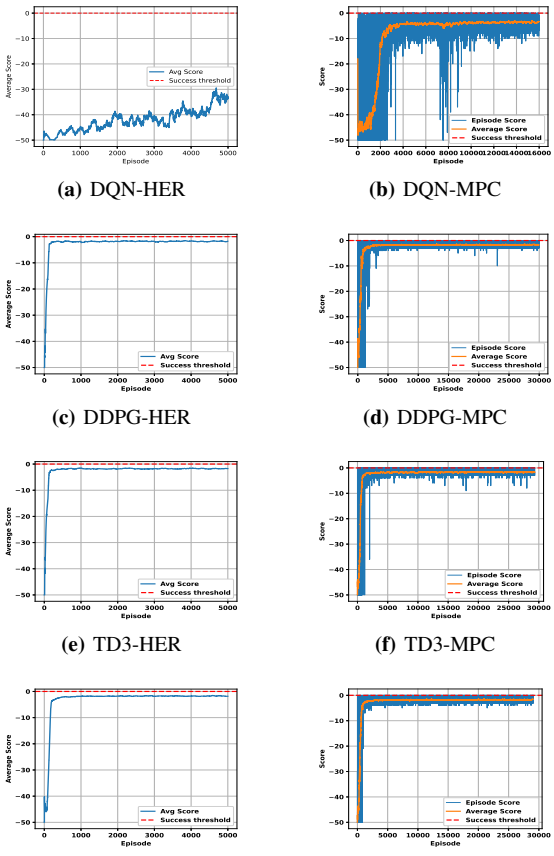
$$\text{Win}_t = \frac{1}{\min(t,100)} \sum_{i=\max(0,t-100)}^{t} \mathbf{1}_{\text{success}(i)} \tag{1}$$

In (1) $\mathbf{1}_{\text{success}}(i)$ represents the binary success indicator for episode i. For MPC hybridization, DRL agents act as long-horizon planners while MPC refines actions over short horizons. The DRL policy first proposes an action, which is used to initialize the MPC controller. The MPC then solves a five-step optimization problem: it predicts future states via Euler integration, evaluates costs associated with goal distance, action magnitude, and action smoothness, optimizes the resulting action sequence using L-BFGS-B [15], and finally executes only the first optimized action while enforcing system dynamics and constraint terms. The trained RL agents supply stable action estimates, which the MPC subsequently refines to guarantee safe and precise execution.

III. Results and Discussion

Performance evaluation on PandaReach-v3 compares baseline DRL algorithms augmented with HER and MPC frameworks across average score, win percentage, and best

score metrics for DQN, DDPG, TD3, and SAC. The models are run for 250,000 timesteps, corresponding to a total of 5,000 episodes. The negative episodes that arise from the reward structure of the PandaReach environment occur when the agent incurs a continuous penalty until the goal is reached. Consequently, during early training with largely uninformative exploratory actions, the agent accumulates a substantial amount of negative reward before developing a goal-directed policy. Referring to Table 2, HER hybridization reveals clear performance disparities where DQN-HER achieves only 40% win percentage with an average score of -33.3, validating that discretization-based approaches cannot handle high-dimensional continuous action spaces despite goal relabelling. DDPG-HER and SAC-HER both achieve 100% win percentages with average scores of -1.8 and -1.7, respectively. At the same time, TD3-HER shows a comparable average score of -1.7 but only a 2% win percentage. This substantial reliability difference indicates TD3 occasionally fails, unfortunately, despite strong average performance when successful, suggesting vulnerability to value estimation errors that HER cannot mitigate.



(a) DQN-HER **(b)** DQN-MPC

(c) DDPG-HER **(d)** DDPG-MPC

(e) TD3-HER **(f)** TD3-MPC

(g) SAC-HER **(h)** SAC-MPC

Fig. 2: Comparison of HER and MPC performance across DQN, DDPG, TD3, and SAC using score percentage metrics.

Similarly, MPC hybridization demonstrates universal improvement, with all algorithms achieving a 100% win percentage (see Table I). DQN-MPC demonstrates substantial improvement over DQN-HER, reducing the average score from -33.3 to -3.31, which validates MPC's ability to refine even suboptimal policies through trajectory optimization. However, DQN-MPC still underperforms continuous-control methods that maintain average scores between -1.57 and -1.85. TD3-MPC achieves the best average score of -1.57, suggesting MPC's refinement particularly benefits deterministic policies by smoothing trajectories and preventing the catastrophic failures observed in TD3-HER. DDPG and SAC maintain consistent performance across frameworks with minimal variations, indicating that both algorithms benefit equally, albeit through different mechanisms. HER provides dense learning signals, while MPC ensures the quality of the trajectory. Fig. 2 compares training curves of HER and MPC hybridizations across DQN, DDPG, TD3, and SAC, showing that MPC variants achieve rapid score improvement and consistent 100% win percentage stabilization while HER variants exhibit gradual learning with algorithm-dependent convergence patterns. Moreover, the win percentage metrics over the episode are given in Appendix A.

Method	Variant	Average Score	Win (%)	Best Score
DQN	HER	-33.3	40	-29.4
	MPC	-3.31	100	-3.09
DDPG	HER	-1.8	100	-1.5
	MPC	-1.73	100	-1.42
TD3	HER	-1.7	2	-1.5
	MPC	-1.57	100	-1.46
SAC	HER	-1.7	100	-1.6
	MPC	-1.85	100	-1.47

TABLE I: Unified performance comparison of HER and MPC hybridizations across DRL methods.

The results validate three critical insights about hybrid DRL architectures for robotic arm manipulation. Algorithm selection for continuous control proves essential, as DQN consistently underperforms due to its reliance on discretization. Among continuous-control methods, DDPG-HER and SAC-HER achieve strong performance with 100% win percentage and high average scores, while TD3-MPC shows the best overall average score and perfect reliability. HER offers efficient learning in sparse-reward settings, but its effectiveness varies across algorithms, as seen by TD3-HER's low success rate despite good scores. In contrast, MPC guarantees consistent success and smooth trajectories, especially for deterministic policies. Framework choice should therefore match the application's need for sample efficiency and safety.

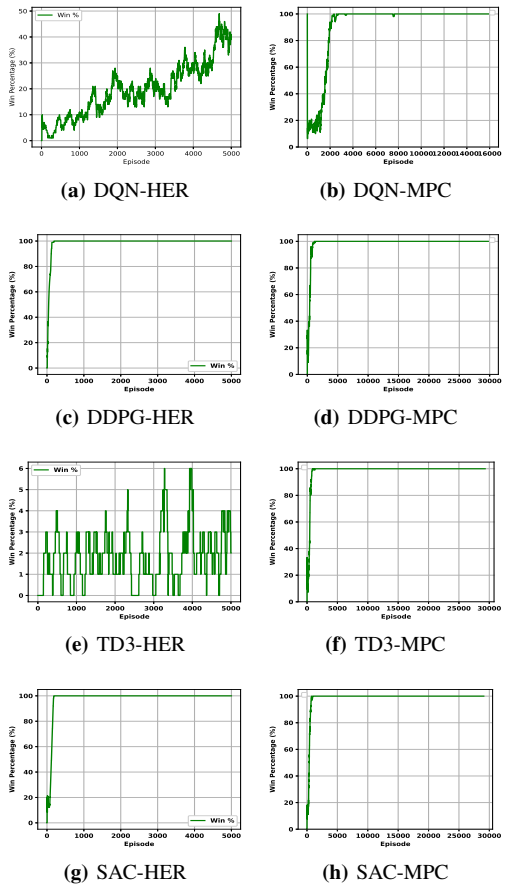
IV. Conclusion

This study demonstrates that hybrid learning-based control frameworks address various challenges in robotic manipulation. DDPG-HER and SAC-HER achieve 100% win rates through reward relabeling, whereas TD3-MPC provides stable and reliable control. HER improves sample efficiency but does not fix exploration issues, whereas MPC enforces safety and consistency. Thus, HER suits fast learning and MPC suits safety-critical tasks. Future work should integrate both within a hierarchical design and test on real robots for stronger validation.

APPENDIX

A. Win Percentage

This appendix provides detailed win percentage convergence patterns across training episodes to complement the aggregate performance metrics presented in the main results. Fig. 3 illustrates win percentage convergence across training episodes for HER and MPC hybridizations. DQN-HER shows gradual improvement, reaching approximately 40% success rate, while DDPG-HER and SAC-HER achieve rapid convergence to 100% within early episodes. TD3-HER exhibits persistent instability, characterized by highly volatile win percentages throughout training, which oscillate erratically despite achieving competitive average scores. In contrast, all MPC variants, including DQN-MPC, DDPG-MPC, TD3-MPC, and SAC-MPC, demonstrate immediate stabilization at a 100% win percentage from the initial episodes, validating MPC's ability to provide consistent reliability through trajectory optimization and constraint satisfaction regardless of the underlying policy quality.



(a) DQN-HER **(b)** DQN-MPC

(c) DDPG-HER **(d)** DDPG-MPC

(e) TD3-HER **(f)** TD3-MPC

(g) SAC-HER **(h)** SAC-MPC

Fig. 3: Comparison of HER and MPC performance across DQN, DDPG, TD3, and SAC using win percentage metrics.

REFERENCES

[1] L. Sciavicco and B. Siciliano, *Modelling and control of robot manipulators*. Springer Science & Business Media, 2012.

[2] T. Al Khawli, M. Anwar, A. Sunda-Meya, and S. Islam, "A calibration method for laser guided robotic manipulation for industrial automation," in *IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2018, pp. 2489–2495.

[3] K. Benali, J.-F. Brethé, F. Guérin, and M. Gorka, "Dual arm robot manipulator for grasping boxes of different dimensions in a logistics warehouse," in *2018 IEEE International Conference on Industrial Technology (ICIT)*. IEEE, 2018, pp. 147–152.

[4] Z. Pan, A. Zeng, Y. Li, J. Yu, and K. Hauser, "Algorithms and systems for manipulating multiple objects," *IEEE Transactions on Robotics*, vol. 39, no. 1, pp. 2–20, 2022.

[5] R. Li and H. Qiao, "A survey of methods and strategies for high-precision robotic grasping and assembly tasks—some new trends," *IEEE/ASME Transactions on Mechatronics*, vol. 24, no. 6, pp. 2718–2732, 2019.

[6] S. E. Li, "Reinforcement learning for sequential decision and optimal control," 2023.

[7] A. Malik, Y. Lischuk, T. Henderson, and R. Prazenica, "A deep reinforcement-learning approach for inverse kinematics solution of a high degree of freedom robotic manipulator," *Robotics*, vol. 11, no. 2, p. 44, 2022.

[8] Y. Hou, L. Liu, Q. Wei, X. Xu, and C. Chen, "A novel ddpg method with prioritized experience replay," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2017, pp. 316–321.

[9] H. Kumar, A. Koppel, and A. Ribeiro, "On the sample complexity of actor-critic method for reinforcement learning with function approximation," *Machine Learning*, vol. 112, no. 7, pp. 2433–2467, 2023.

[10] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," 2018. [Online]. Available: https://arxiv.org/abs/1801.01290

[11] Y. Luo, Y. Wang, K. Dong, Q. Zhang, E. Cheng, Z. Sun, and B. Song, "Relay hindsight experience replay: Self-guided continual reinforcement learning for sequential object manipulation tasks with sparse rewards," *Neurocomputing*, vol. 557, p. 126620, 2023.

[12] S. M. Tahamipour-Z, G. R. Petrovic, and J. Mattila, "Robust model predictive control for robot manipulators," in *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*. IEEE, 2022, pp. 420–426.

[13] H. R. Maei, C. Szepesvári, S. Bhatnagar, and R. S. Sutton, "Toward off-policy learning control with function approximation." in *ICML*, vol. 10, 2010, pp. 719–726.

[14] Q. Gallouédec, N. Cazin, E. Dellandréa, and L. Chen, "panda-gym: Open-source goal-conditioned environments for robotic learning," 2021. [Online]. Available: https://arxiv.org/abs/2106.13687

[15] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on scientific computing*, vol. 16, no. 5, pp. 1190–1208, 1995.