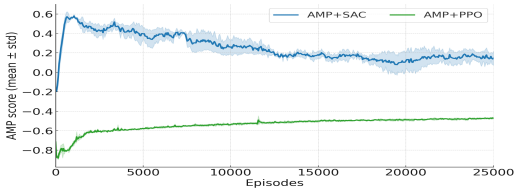
Adversarial Priors Unleashed with Soft Actor Critic

Anonymous Author(s)

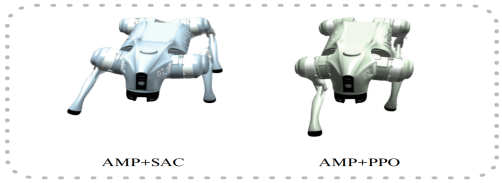
Bio–inspired robotics motion has emerged as a critical research direction. A widely adopted paradigm is imitation learning, where robots acquire motor skills by mimicking animal behaviors under reward-driven mechanisms. While substantial progress has been achieved, existing work has often emphasized performance tuning often to the detriment of a more in-depth analysis of the underlying methodological design. In this work, we address this gap by proposing a novel integration of Adversarial Motion Priors (AMP) with the Soft Actor–Critic (SAC) [2] algorithm that leverages entropy–regularized off–policy learning. Whereas prior studies have primarily combined AMP with Proximal Policy Optimization (PPO) [1], our design enables policies to more effectively incorporate discriminator feedback by reusing past trajectories, in contrast to PPO's reliance on short-horizon, freshly collected rollouts. For the experiments, we trained the Unitree Go2 quadruped in Isaac Gym with 4096 parallel environments to imitate walking and trotting gaits under both AMP+SAC and AMP+PPO, evaluating across three random seeds. The reward objective function is defined in Eq.1:

$$r(s,a,s') \;=\; \lambda_{\text{AMP}}\,\max\Big[1-0.25,(D_\phi([s,s'])-1)^2\Big] \;+\; \lambda\, r_{\text{task}}(s,a,s'), \tag{1}$$

where, $\lambda_{\text{task}} = 0.3$ and $\lambda_{\text{AMP}} = 0.7$. The discriminator D_ϕ distinguishes expert transitions re–targeted of the dog motion–capture dataset [3] from those generated by the robot policy. This adversarial signal reflects how indistinguishable the policy's motion is from the expert reference, and its average over time, "AMP score" serves as the key metric we propose in this work to measure imitation fidelity. Fig.1.(a) shows that AMP+SAC achieves higher and more stable imitation scores than AMP+PPO. Fig.1.(b) highlights the stance–swing transitions under both algorithms. AMP+SAC maintains coordinated fore–hind limbs synchronization and consistent step height closer to a dog posture. AMP+PPO, in contrast, displays inconsistencies such as asynchronous hind limbs movement, uneven swing clearance, and phase drifts that break natural dog's gait symmetry. Our preliminary findings



(a) Comparison of mean ± standard deviation of AMP score.



(b) Snapshots comparing body poses at stance–swing transitions.

Figure 1: Comparison between the proposed AMP+SAC approach and the baseline AMP+PPO.

show that the off–policy nature of SAC synergizes with AMP, enabling better use of discriminator signals and more natural quadruped motion compared to PPO. This suggests that discriminator–driven imitation learning benefits significantly from algorithms capable of reusing experience.

References

[1] A Escontrela, X.B Peng, W Yu, T Zhang, A Iscen, K Goldberg, and P Abbeel. Adversarial motion priors make good substitutes for complex reward functions. In *IEEE/RSJ IROS*, 2022.

[2] T Haarnoja, 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*, 2018.

[3] H Zhang, S Starke, T Komura, and J Saito. Mode-adaptive neural networks for quadruped motion control. *ACM Transactions on Graphics*, 2018.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025).