2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Unlocking the Potential of Soft Actor–Critic for Imitation Learning

Nayari Marie Lessa*[1,2], Melya Boukheddimi[1], and Frank Kirchner[1,2]

Abstract— **Learning-based methods have enabled robots to acquire bio-inspired movements with increasing levels of naturalness and adaptability. Among these, Imitation Learning (IL) has proven effective in transferring complex motion patterns from animals to robotic systems. However, current state-of-the-art frameworks predominantly rely on Proximal Policy Optimization (PPO), an on-policy algorithm that prioritizes stability over sample efficiency and policy generalization. This paper proposes a novel IL framework that combines Adversarial Motion Priors (AMP) with the off-policy Soft Actor-Critic (SAC) algorithm to overcome these limitations. This integration leverage replay-driven learning and entropy-regularized exploration, enabling naturalistic behavior and task execution improving data efficiency and robustness. We evaluate the proposed approach (AMP+SAC) on quadruped gaits involving multiple reference motions and diverse terrains. Experimental results demonstrate that the proposed framework not only maintains stable task execution but also achieves higher imitation rewards compared to the widely used AMP+PPO method. These findings highlight the potential of an off-policy IL formulations for advancing motion generation in robotics. Code and supplementary material are available at: `https://github.com/nayariml/AMP_SAC.git`**

I. Introduction

The past decade has witnessed remarkable progress in learning-based algorithms for robotics, with applications spanning a wide range of domains. These advances aim to endow robots with greater intelligence and adaptability, enabling them to perform tasks with smooth, natural, and human or animal–like motions [1], [2], [3]. Among the many research directions, bio-inspired robotics has emerged as particularly impactful, influencing both mechanical design and motion generation [4]. Within this context, IL plays a pivotal role in generating bio-inspired behaviors, enabling robots to reproduce complex motion patterns during task execution [5]. IL has been successfully applied across diverse areas of robotics. For example, in medical applications robotic arms replicate the precise motions of expert physicians [6]. While in animal-inspired locomotion, quadruped robots have learned gaits from natural counterparts [7]. A variety of IL approaches have been explored. For instance, natural motion generation is proposed by [8] an approach that learns a CPG-based controller for humanoid gait generation. Similarly, model-based strategies have been developed. [9] Integrated Model Predictive Control (MPC) with Multi-Layer Perceptron (MLP) experts to generate diverse quadruped gaits, while [10] combined MPC with deep reinforcement learning (RL) to improve locomotion. Although effective, these methods mainly rely on optimal trajectory tracking and often neglect the dynamic variability inherent in animal-inspired motion. To address these limitations, recent research has combined IL with RL, often leveraging real animal motion dataset as references. For example, [11] proposed "in-between" an IL approach that generated dynamically consistent gaits with precise velocity tracking on quadruped robots. Similarly, [12] introduced an RL-based IL framework trained on dog motion data, where a trajectory encoderdecoder was coupled with PPO [13], enabling the reproduction of highly dynamic behaviors on real hardware. Among RL algorithms, PPO has become the dominant on-policy method in IL, primarily due to its stability and efficiency [14]. While stable, PPO is limited in sample efficiency and policy generalization. A wide range of PPO-based IL frameworks have been proposed to transfer animal-inspired behaviors to quadruped robots. Early work demonstrated the feasibility of this paradigm, such as [15], which trained dog-inspired policies and successfully deployed them on physical systems. Building on this, [16] introduced CASSI, an unsupervised adversarial imitation framework coupled with PPO that enabled quadrupeds to acquire a diverse repertoire of behaviors. To further enhance robustness, [17] augmented PPO with a reference motion encoder, improving locomotion on challenging terrains. [18] proposed a hierarchical RL framework in which animal motions were represented as discrete latent embeddings, enabling successful deployment of highly dynamic behaviors on quadruped robots. More recently, [19] leveraged domain randomization during PPO training to obtain robust policies capable of adapting to di-

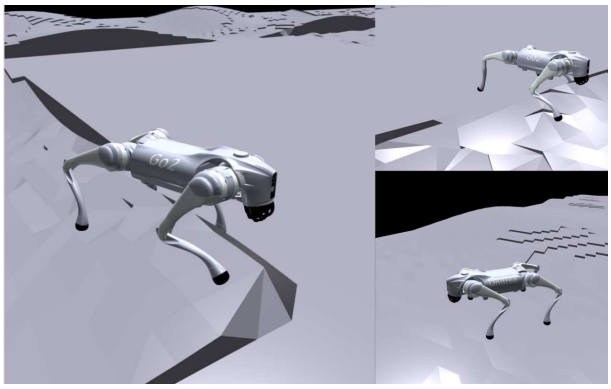

Fig. 1: Snapshots of an obtained AMP+SAC imitation learning policy performing multiple gaits across different terrains.

*Corresponding author: lessa@uni-bremen.de
[1]Robotics Innovation Center, DFKI GmbH, 28359 Bremen, Germany.
[2]University of Bremen, 28359 Bremen, Germany.

This work has been supported by the German Federal Ministry of Research, Technology and Space (BMFTR) under the Robotics Institute Germany (RIG), German Aerospace Center (DLR) within the ActGPT project (Grant number 01IW25002).

verse motions and terrains. Despite this remarkable progress, most efforts remain focused on performance optimization and sim-to-real transfer, with comparatively less attention given to exploring alternative algorithmic formulations and design choices within IL frameworks. In this work, we aim to address this gap. Specifically, we propose a framework that replaces the commonly used on-policy PPO with the off-policy SAC algorithm [20], while combining it with AMP [21], [22] to provide structured imitation guidance. Unlike PPO, SAC benefits from its off-policy nature, allowing more efficient exploration and broader policy generalization. These properties are particularly important for robots that must operate in unconstrained environments, adapting to diverse motions and terrains while maintaining smooth, natural behaviors. As a use case, we evaluate quadruped locomotion across multiple motions and terrains. Including a performance benchmarking against a baseline AMP+PPO implementation.

Contributions: The main contributions of this paper are:

- A novel IL framework that integrates AMP with the off-policy SAC algorithm, enabling robust task execution while preserving natural, animal-inspired motion.
- Extensive evaluations on quadruped locomotion tasks involving multi-motions and varying terrains, demonstrating the frameworks ability to generalize beyond reference trajectories.
- Performance comparison against a widely used AMP+PPO baseline, showing that our approach achieves superior imitation rewards.

Organization: The remainder of this paper is organized as follows. Section II introduces the proposed methodology. Section III describes the experimentation setup. Section IV presents the results and comparisons with the baseline. Finally, Section V summarizes the findings and limitations and outlines future research directions.

II. Methodology

A. Synopsis

In this work, we aim to train a robot to imitate animal behaviors using a combination of reinforcement learning and adversarial imitation. The overview of the proposed methodology is illustrated in Fig. 2. The pipeline is categorized in two sequential stages: (1) motion capture processing for extracting reference kinematic features, and (2) adversarial imitation learning. The latter stage is itself composed of two components: training a discriminator to differentiate expert from agent motions, and optimizing a policy via SAC.

B. Reference Data Processing

In order to perform our training, a data processing step is necessary, for that reason, the raw dataset is pre-processed using affine transformations to account for scale differences and foot offsets. To ensure kinematic feasibility, we converted the 3D joint positions into joint angles via inverse kinematics and computed joint velocities as well as base linear and angular velocities. This procedure yields a sequence of joint angle vectors $q_t{}_{t=1}^{T}$, where T is the number of frames

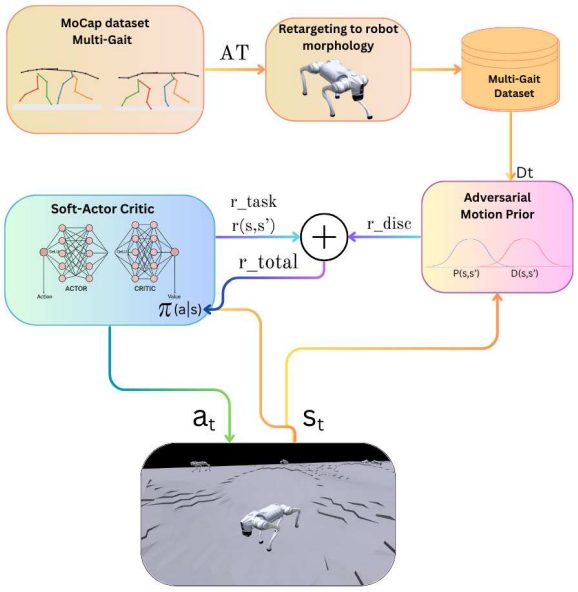

Fig. 2: Pipeline of the AMP+SAC imitation learning.

in the motion clip. The final reference motion dataset $\mathcal{D}$ consists of root position and orientation, joint angles, joint velocities, and foot positions, all expressed in the local frame of the robot.

C. Discriminator-Guided Imitation with Soft Actor Critic

We formulate the locomotion imitation learning problem as a Markov Decision Process (MDP) defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P(s'|s,a)$ denotes the transition dynamics, $r(s,a,s')$ is the reward function, and $\gamma \in [0,1]$ is the discount factor. The goal of Reinforcement learning is to find a policy $\pi_\theta(a|s)$ parameterized by θ that maximizes the expected cumulative discounted reward $J(\pi) = \mathbb{E}_\pi[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)]$ over a finite horizon T. To solve the locomotion control problem, we adopt the SAC algorithm [20], an off-policy actor-critic method that augments the reward maximization objective with an entropy regularization term. This encourages exploration and improves robustness of the learned policy. The SAC objective is defined as:

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim\pi}\big[r(s_t,a_t) + \alpha\,\mathcal{H}(\pi(\cdot|s_t))\big], \qquad (1)$$

where $\mathcal{H}(\pi(\cdot|s))$ denotes the entropy of the policy π_θ, and α is the temperature coefficient that controls the trade-off between maximizing the expected return and encouraging exploration. In our framework, the actor $\pi_\theta(a|s)$ is parameterized by a multi-layer perceptron (MLP). The network receives normalized states at time t and outputs the mean and log-standard deviation of a Gaussian distribution. An action is then sampled from this distribution and passed through a $\tanh$ squashing function to ensure outputs lie within valid bounds, a process implemented using the re-parameterization trick to enable end-to-end gradient back-propagation. The

actor outputs continuous joint commands for the robot. The critic consists of two independently initialized Q-networks, $Q_{\psi_1}(s,a)$ and $Q_{\psi_2}(s,a)$, each implemented as an MLP. These are trained to estimate the expected soft Q-value, which includes the expected return plus an entropy bonus, for state-action pairs. A discriminator network $\mathcal{D}_\phi$ is trained to distinguish between state transitions (s,s') generated by the policy π and those from the expert dataset $\mathcal{D}$. We employ Adversarial Motion Priors (AMP) [21] to train the discriminator objective:

$$\begin{aligned}\arg\min_{\phi} \mathbb{E}_{(s,s')\sim\mathcal{D}}\left[\left(D_\phi(s,s')-1\right)^2\right] \\ + \mathbb{E}_{(s,s')\sim\pi_\theta(s,a)}\left[\left(D_\phi(s,s')+1\right)^2\right] \\ + \frac{w^{\text{gp}}}{2}\,\mathbb{E}_{(s,s')\sim\mathcal{D}}\left[\left\|\nabla_\phi D_\phi(s,s')\right\|^2\right],\end{aligned} \tag{2}$$

where the first two terms form a least-squares adversarial loss objective to differ the states transitions, and the last term is the gradient penalty regularization controlled by the weight w^{gp} to prevent over-fitting to the reference dataset, improving stability during adversarial training.

1) Actor objective with AMP shaping: In SAC, the actor objective $\mathcal{J}_\pi(\theta)$ optimizes the policy π_θ to maximize expected future return while simultaneously enforcing entropy maximization. Formally, the actor is trained by minimizing the KullbackLeibler divergence between the entropy-augmented reward, weighted by temperature α, and the state-action value estimate $Q_\psi(s,a)$:

$$\mathcal{J}_\pi(\theta) = \mathbb{E}_{s\sim\mathcal{D}}\left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\alpha\log\pi_\theta(a|s) - Q_\psi(s,a)\right]\right]. \tag{3}$$

To integrate AMP with SAC the actor loss function is augmented with adversarial objectives:

$$\mathcal{J}\pi^{\text{AMP}}(\theta) = \mathcal{J}\pi(\theta) + \lambda_{\text{AMP}}\cdot\mathcal{L}_{\text{AMP}} + \lambda_{\text{grad}}\cdot\mathcal{L}_{\text{grad}}, \tag{4}$$

where $\mathcal{L}_{\text{AMP}}$ is the adversarial loss from the AMP discriminator, encouraging policy-generated state transitions to resemble expert demonstrations, and $\mathcal{L}_{\text{grad}}$ is the gradient penalty loss that ensures training stability for the discriminator. The coefficients λ_{AMP} and $\mathcal{L}_{\text{grad}}$ control the weighting of these additional terms. This formulation enables the policy to acquire complex skills through imitation while preserving the robustness and exploration guarantees of maximum entropy RL.

D. Task and Imitation Reward Formulation

The adversarial framework minimizes reward engineering by replacing explicit trajectory tracking with a learned reward signal. A discriminator D_ϕ provides this signal, guiding the policy to imitate the expert by rewarding it for generating states that D_ϕ classifies as real. This AMP reward is defined as:

$$r(s_t,s_t') = \max\left[1 - 0.25(D_\phi([s,s'])-1)^2\right], \tag{5}$$

The transformation $(D_\phi(\cdot)-1)^2$ and the $\max$ operation serve to bound the reward within the range [0,1] to stable learning. Our goal is to evaluate the locomotion imitation while promoting a task-oriented and physically feasible behaviors on the robot. Therefore, the adversarial reward is combined with a separate locomotion reward $r_{\text{task}}(s,a,s')$, composing by a forward command velocity $\vec{v}_{\text{t}} = [\vec{v}_{\text{t}}^{x}]$ specified in the base frame, and additional terms that penalize undesired motions such as excessive lateral drift, high base angular velocity around the x,y,z-axis, and position limit violations. The total reward r_t at each time-step is therefore a weighted sum of these two components:

$$r(s_t,a_t) = \mathcal{W}\,r_{\text{task}}(s,a,s'), + \mathcal{W}_{\text{AMP}}\,r_(s_t,s_t'), \tag{6}$$

where $\mathcal{W}_{\text{AMP}}$ is a scalar coefficient that balances the importance of imitation against the task reward.

III. Experimental implementation

In this section, we present the experimental setup employed to evaluate the proposed imitation learning framework. We detail the training configuration and evaluation protocol used to benchmark SAC and PPO within AMP framework.

A. Reference Dataset

The motion capture dataset used in this work [23] contains clips of a German Shepherd dog performing various motion behaviors. For this work, we selected walking and trotting gaits, each clipped to one gait cycle. The resulting data have a total duration of 1.95 seconds (95 frames), with 1.134 seconds (55 frames) for walking and 0.82 seconds (40 frames) for trotting. The raw data consist of 3D joint positions and orientations, which were pre-processed using the methodology detailed in Sec.II.b.

B. Multi-Gait Imitation Learning

We evaluate the performance of our method against the established benchmark of PPO within the AMP framework. Both algorithms were trained under identical conditions, with 5 random seeds each, using the same reward function parameterization and domain randomization settings. Our SAC-based implementation uses separate Actor and Critic networks, each comprising two hidden layers with 1024 and 512 units, and GELU activation functions. The AMP discriminator network also consists of two hidden layers but uses ReLU activation. We choose AdamW to optimize α and actor. The Actor is optimized with the policy parameters coupled with discriminator's encoder and readout layer, for alignment. The discriminator is optimized within a supervised learning objective defined in Eq. 2. Training is performed using a replay buffer with a capacity of 10^6 experience tuples. For each training iteration, a batch of 16.384 transitions is sampled for the SAC policy update over 8 epochs, while a separate batch of 8.192 transitions is used for a single discriminator update epoch. Additionally, adopted automatic α tuning to leverage the trade-off between exploration and exploitation based on the entropy level of the policy. The complete hyper-parameter configuration for our method is detailed in Table I and Table II. For the PPO+AMP

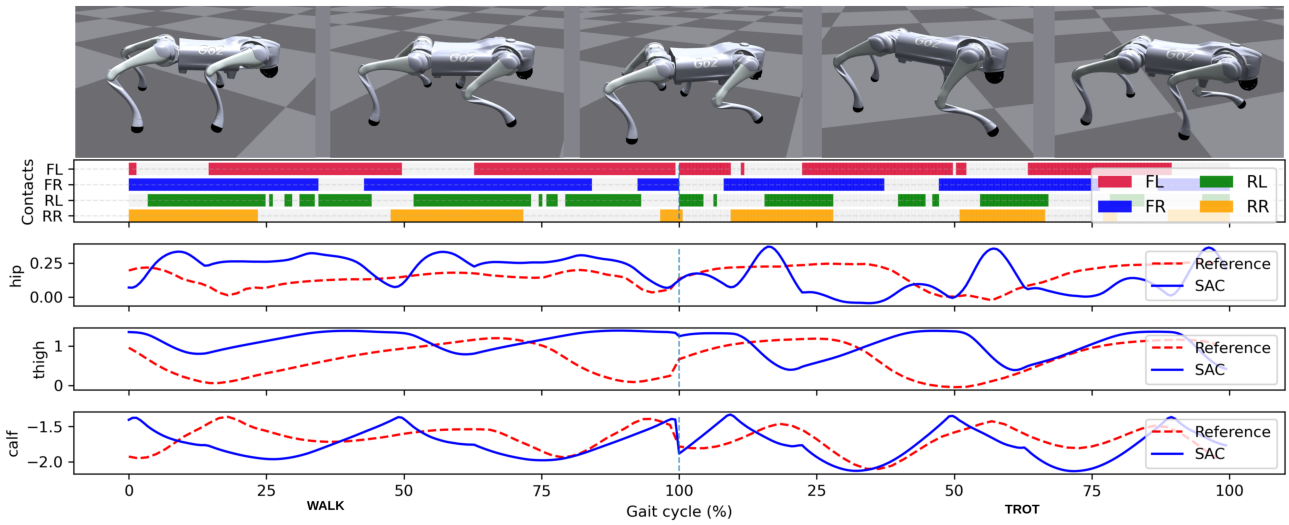

Fig. 3: Walk–trot transition over normalized gait cycles (0–100% walk, 0–100% trot). Top: robot frames showing foot contacts. Bottom: front-left hip, thigh, and calf joint rotations (rad); AMP+SAC (blue) outputs compared with retargeted reference (red dotted) on flat terrain.

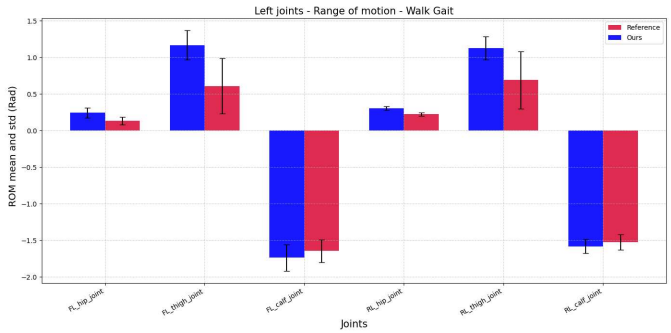

Fig. 4: Comparison of joint ROM (mean ± std) during walk (left side) between the retargeted reference (red) and our AMP+SAC (blue) implementation on flat terrain.

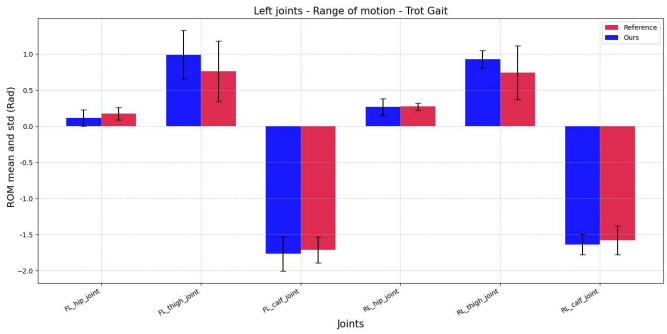

Fig. 5: Comparison of joint ROM (mean ± std) during trot (left side) between the retargeted reference (red) and our AMP+SAC (blue) implementation on flat terrain.

baseline, we adopted the setup and hyper-parameters from [22].

C. Progressive Terrain Curriculum

To evaluate the robustness and generalization capabilities of the learned policies, we employed a structured terrain curriculum. This curriculum progresses from flat ground to continuous wave-like terrains characterized by a maximum undulation height of 3 cm, with varying amplitude and frequency. Each policy was warm-started from a pre-trained multi-gait model previously optimized for flat terrain. This initialization strategy accelerates learning on novel terrains by transferring the prior kinematic knowledge encoded in the policy's weights. Specifically, the weights from the pre-trained model were used to initialize the actor network before proceeding with the multi-gait training regimen on the new terrain. The core training configuration detailed in Table I and Table II was maintained from the warm-start model. However, to enhance training stability, we increased the number of discriminator update steps per iteration to four. The terrain difficulty automatically advanced to the next level once the policy consistently achieved a performance threshold of at least 60% of the maximum possible reward for velocity tracking on the current terrain.

D. Training

We conducted the imitation learning experiments using 4096 parallel environments per GPU, with policies evaluated at 200 Hz. All experiments were performed in simulation using Isaac Gym [24], running on a workstation equipped with an NVIDIA RTX A6000 GPU and 80 GB RAM. The learned policies were evaluated on a Unitree Go2 quadruped robot with 12 actuated joints. At the beginning of each training episode, the robot is initialized with a random motion sample from the reference dataset $\mathcal{D}$, ensuring diverse starting conditions. The state representation $s \in \mathcal{S}$ includes the

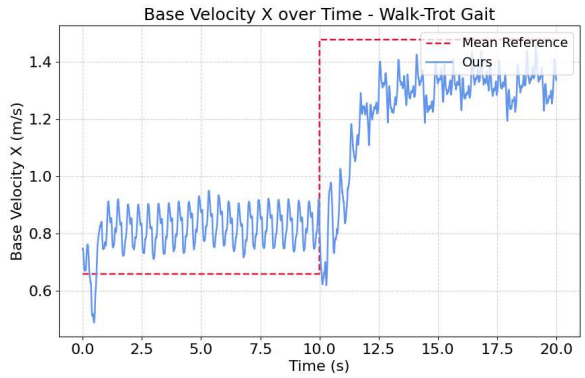

Fig. 6: Base x-velocity (blue) over time for walk and trot; the dotted red line denotes the mean reference velocity, highlighting the AMP+SAC transition on flat terrain with multiple motions.

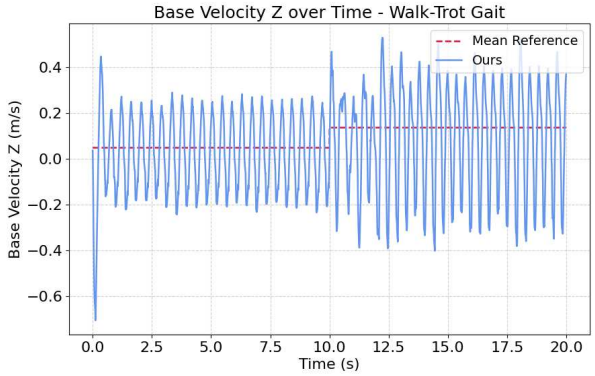

Fig. 7: Base z-velocity (blue) over time for walk and trot; the dotted red line denotes the mean reference velocity, highlighting the AMP+SAC transition on flat terrain with multiple motions.

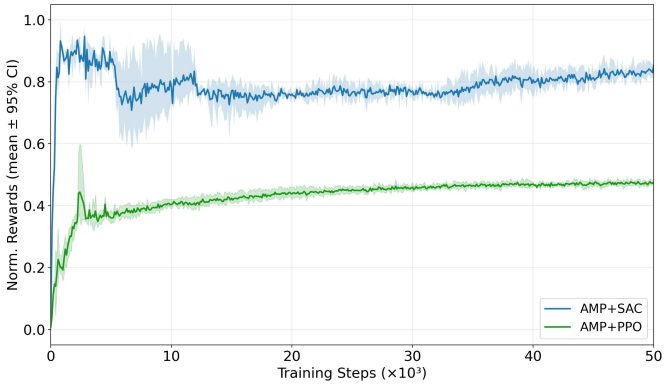

Fig. 8: AMP discriminator reward during training over 50k episodes for AMP+SAC (ours, blue) and AMP+PPO (green) on flat terrain with multiple motion skills. Solid lines show the mean across seeds; shaded regions indicate 95% bootstrapped confidence intervals.

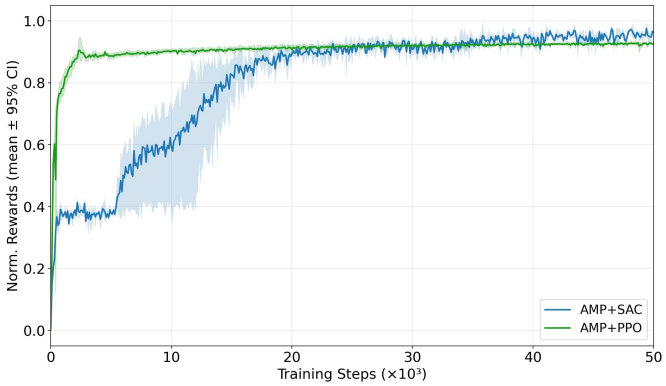

Fig. 9: Learning curves over 50k training episodes for AMP+SAC (ours, blue) and AMP+PPO (green) on flat terrain with multiple motion skills. Solid lines show the mean return across seeds; shaded regions indicate 95% bootstrapped confidence intervals.

command velocity target, joint positions and velocities, and base orientation. The action $a \in \mathcal{A}$ corresponds to continuous joint position targets. For terrain curriculum experiments, the terrain parameters are also included in the observation space. For all the experiments, the reward composition is defined by weight for the imitation task $\mathcal{W}_{\mathcal{AMP}}$ set to 0.6 and for the reward task set to 0.4. The detailed locomotion support reward terms and their scales are listed in the Table III. In the complex terrain setting, the penalty for approaching joint limits was decreased to -2. Specifically, in the complex terrain experiment, we decreased the penalization to -2 to boost exploration and adaptability in the uneven surfaces. To improve policy robustness, we applied domain randomization over the ranges specified in Table IV, varying the robots base mass, the motor gain multipliers of the PD controllers, and the terrain friction coefficients. The data collection steps for the multi-gait experiment was approximately 6 billion over 40 hours, corresponding to the equivalent of 1.9 years of real-world experience. The terrain curriculum experiment involved roughly 6 million steps completed in 48 hours, equivalent to 0.69 years of real-world interaction. Training on terrain curricula proved more computationally demanding, requiring significantly longer simulation time per step.

E. Comparison Metrics

1) Imitation metric : We evaluate imitation learning performance using the *task reward* (environment return) and the *discriminator reward* from AMP. For fair comparison between AMP+SAC and AMP+PPO, each seed trajectory is linearly interpolated onto a common grid every 100 training steps and then globally normalized. Given N random seeds, the mean performance at training step t is computed as

$$\mu_t = \frac{1}{N} \sum_{i=1}^{N} R_{i,t},$$

where $R_{i,t}$ denotes the normalized metric value (task or discriminator reward) from seed i at step t. Shaded regions in the plots represent 95% bootstrapped confidence intervals of the mean computed across seeds.

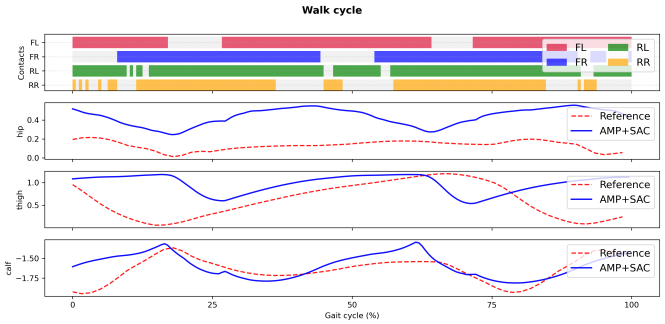

Fig. 10: Foot contact timings (top) and joint rotations (front-left hip, thigh, calf; rad) during a normalized walk cycle (0–100%). Blue lines show AMP+SAC outputs, compared with the retargeted reference (red dotted). Experiments on wave terrain (max undulation 3 cm)

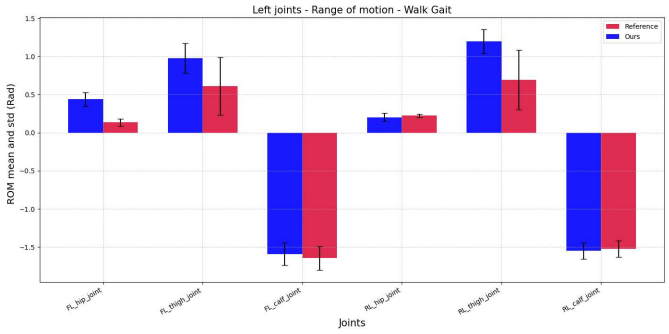

Fig. 11: Comparison of joint ROM (mean ± std) over one walk cycle (left side) between the retargeted reference (red) and our method (AMP+SAC, blue) on wave terrain (max undulation 3 cm).

2) Bio-Mechanical Parameters: We analyze locomotion performance using three representative parameters: foot contacts, which quantify the stance and swing phases of each limb; the base velocity, reflecting forward progression and gait stability; and the joint range of motion (ROM) of the front-left hip, which indicates how closely the robot reproduces the kinematic envelope of the reference motion. This multi-faceted approach allows us to evaluate performance across contact dynamics, whole-body trajectory, and detailed kinematic reproduction.

IV. Results

In this section, we present the experimental results, compare them against the selected baseline, and highlight key insights. The reported outcomes are averaged over six independent runs with different random seeds in order to account for variability in training. Only one side of the robot is represented in the figures for brevity.

A. Evaluation across multiple motions

In this part, we summarize the obtained performance across the tasks considered in our study, using the implementations introduced above. Fig. 3 presents the evolution

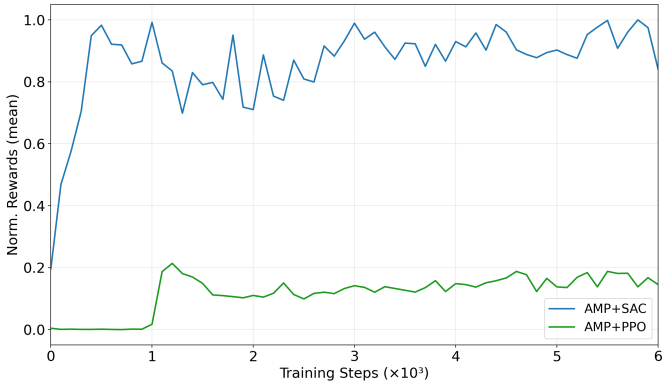

Fig. 12: AMP discriminator reward over 6k training episodes for AMP+SAC (ours, blue) and AMP+PPO (green) on wave terrain (max undulation 3 cm). Solid lines show the mean episode return.

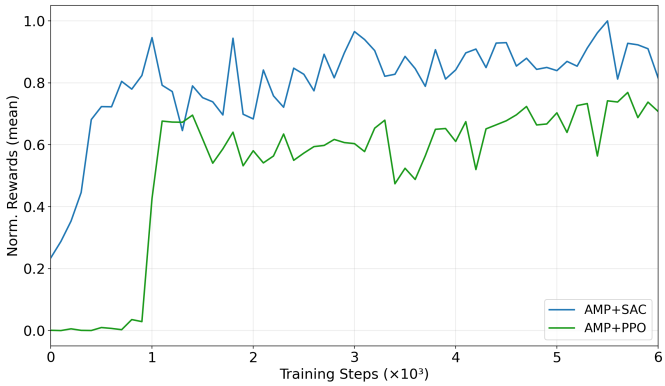

Fig. 13: Learning curves over 6k training episodes for AMP+SAC (ours, blue) and AMP+PPO (green) on wave terrain (max undulation 3 cm). Solid lines show the mean episode return.

of the front-left hip, thigh, and calf joint rotations along with the corresponding foot contact timings during normalized walk and trot cycles. The obtained trajectories with the proposed AMP+SAC framework closely track the retargeted dog's reference motions, both in amplitude and in phase. From the contact patterns, it can be observed that the policy has learned to reproduce the alternating stanceswing transitions that characterize walk and trot, while performing the reward's tasks of walking forward. Fig. 4 and Fig. 5 present a comparison of the obtained ROM, including standard deviations (in rad), for the walk and trot motions. These results are evaluated against the corresponding retargeted ROM derived from dog reference data. The figures demonstrate that the imitated motions (walk and trot) achieve joint rotations with ROM closely matching those of the reference. This indicates that the robot not only performs the task of walking straight but also effectively reproduces the underlying gait dynamics of the reference motions. Overall, the results validate that the proposed AMP+SAC framework enables the robot to acquire natural and smooth motion

TABLE I: SAC hyper-parameters used in training.

Parameter	Value
Replay memory size	1.0×10^7
Batch size	16384
Updates per step	8
n-step return	3
Target smoothing coefficient τ	0.05
Discount factor γ	0.99
Actor learning rate α_π	0.001
Entropy coefficient α	automatic
Actor hidden layers	[1024, 512]
Actor activation	GELU
Critic hidden layers	[1024, 512]
Critic activation	GELU
Warm-up steps	100

TABLE II: AMP hyper-parameters for imitation learning.

Parameter	Value
AMP loss coefficient λ_{AMP}	0.1
Gradient penalty coefficient λ_{GP}	0.01
AMP batch size	8192
AMP batch count per update	1
AMP reward coefficient	2.0
Pre-loaded motion transitions	2.0×10^7
Discriminator hidden layers	[1024, 512]
Minimum normalized std.	[0.01, 0.01, 0.01]

TABLE III: Reward terms used in the locomotion task.

Reward term	Value
Tracking lin. vel. x	240.0
Lin. vel. z	-15.0
Angular vel. xy	-1.0
Angular vel. z	-5.0
Base height	-2.0
Collision	-1.0
Feet contact forces	-2.0
DoF acceleration	-2.5×10^{-7}
DoF pos. limits	-4.0
Torques	-5.0×10^{-5}

TABLE IV: Domain randomization parameters used during training.

Parameter	Range
Base mass (m)	[-1., 1.]
PD gains factor	[0.9, 1.1]
Terrain friction coefficient (μ)	[0.25, 1.75]

patterns across both walk and trot gaits. Fig. 6, Fig. 7 illustrates the evolution of the robots base velocity in the x and z directions, obtained with the proposed method. The policy regulates the forward velocity consistently across walk and trot motions. The results indicate that the learned policy is capable of imitating the reference, leading to stable and natural motion. Combined with the joint rotation and foot contact analysis presented above, these findings demonstrate that the proposed framework successfully learns coordinated gait dynamics that closely imitate animal (dog) motion while performing the task of walking forward.

The results presented in Fig. 8 and 9 demonstrate that the combination of AMP+SAC offers notable advantages over AMP+PPO. In terms of imitation performance, AMP+SAC achieves higher average discriminator rewards, suggesting a closer match to expert behaviors. The discriminator loss figure, available in the supplementary material, shows that this outcome is achieved while AMP+SAC maintains moderate discriminator loss levels (~ 1.3). This finding implies that the policy continues to match the expert state distribution within a balanced minmax game, unlike the near-zero loss observed with AMP+PPO (~ 0.3). The overall task reward further indicates that AMP+SAC exhibits superior long-term sample efficiency, outperforming AMP+PPO.

B. Evaluation across multi-terrains

In the second experiments, we assess the robustness of the proposed approach when transitioning from flat to wave terrain. The objective is to determine whether AMP+SAC can perform the task of walking forward, while imitating the reference under terrain perturbations. Figure 10 illustrates the joint rotations (in rad) and foot contact timings over one walk cycle, on wave terrain with a maximum undulation amplitude of 3 cm. This figure shows the expected periodic walking pattern and joint trajectories with similar variations as the reference. Figure 11 compares the obtained joints ROM on the left side of the robot with the re-targeted reference, demonstrating that the trajectories remain closely aligned with the reference ROM.

A similar trend to the previous experiment appears in Fig. 12 and Fig. 13 where AMP+SAC consistently achieves higher discriminator reward values and mean task rewards compared to AMP+PPO. The discriminator loss curves in the supplementary material show that AMP+SAC remains stable (~ 2.2) within a balanced game, whereas AMP+PPO exhibits stagnation (~ 0.1).The consistent performance across both flat and uneven terrain indicates that AMP+SAC reliably discovers complex motions that fulfill both task objectives and stylistic authenticity compared to AMP+PPO.

V. Discussion and Outlooks

The presented results highlight the use cases achieved in order to assess the usability of the proposed implementation. A core strength of SAC is its replay buffer, which retains short- to medium-horizon interaction histories and enables efficient reuse of past experience, leading to more effective exploration and higher sample efficiency. Although this off-policy algorithm is a promising RL approach, SAC performance is sensitive to hyperparameters, which has historically constrained its adoption. In this work, we proposed, to the best of our knowledge, the first implementation of an AMP+SAC imitation learning framework. We tested our implementation in two experiments: (i) learning two motions, and (ii) continuing imitation while the environment, in our case the terrain, is modified. The first experiment tested the capacity to learn and imitate two motions with higher mean imitation reward. The comparison was done against the re-targeted motions from a dog and the AMP+PPO [22]. The second experiment evaluated the robustness and

adaptability of our learning policy, in an uneven terrain. Under these conditions, SAC demonstrated sustained optimization of the expected return across successive epochs, even when presented with novel, unseen observations. We observed that AMP+SAC require more training time to learn a good policy than its counterpart AMP+PPO. The reason is well established, SAC fundamentally is made for exploration. For high-dimensional tasks, this exploratory priority leads to increased sample complexity and extended training periods before policy improvement is observed. Therefore, the main limitation of our approach is training time, especially compared to PPO. Another limitation is the buffer. The combination of SAC buffer plus AMP buffer requires higher computational resources. This trade-off is decisive for real-world applications. The purpose of this work was to propose to the community a new IL implementation, adding more options to our toolbox. The full implementation will be open-sourced and made available for further experiments and optimization of the current implementation. At a time when robots are expected to evolve more and more into our everyday lives, systems must become more generalist and less specialized in a single skill or behavior. For this, an off-policy algorithm like SAC is crucial in enabling the learning of multiple skills or motions. Furthermore, the development of new and versatile IL algorithms is essential to produce smoother and more natural motions, thereby allowing robots to interact more effectively with humans in environments designed for human movements and capacities. Based on these very promising results, our future work will explore the integration of an additional small storage to bias SAC replay toward recent data for reflects the current policy and changing AMP reward, while keeping the older transitions to avoid forgetting past experience. With this change, we aims to speed up learning and stabilize training under non-stationary conditions. Additionally, we aim to extend our framework capabilities by enabling the multi-skill learning using hierarchical learning in combination with continual learning technique to add new skills without forgetting, and expands its repertory to skills generalization. Finally, we plan to perform real-world deployment, especially in legged robots, to test the robustness and generalization of our framework addressing sim-to-real challenges.

REFERENCES

[1] Y. Tong, H. Liu, and Z. Zhang, "Advancements in humanoid robots: A comprehensive review and future prospects," *IEEE/CAA Journal of Automatica Sinica*, vol. 11, no. 2, pp. 301–328, 2024.

[2] X. Qin, X. Ma, Y. Qi, Q. Liu, C. Xue, N. Gui, Q. Dong, J. Yang, and B. Liang, "Integrating diffusion-based multi-task learning with online reinforcement learning for robust quadruped robot control," *arXiv preprint arXiv:2507.05674*, 2025.

[3] O. Eren Akgün, N. Cuevas, M. Farias, and D. Garces, "Tiny reinforcement learning for quadruped locomotion using decision transformers," *arXiv e-prints*, pp. arXiv–2402, 2024.

[4] Z. Zhang, T. Liu, L. Ding, H. Wang, P. Xu, H. Yang, H. Gao, Z. Deng, and J. Pajarinen, "Imitation-enhanced reinforcement learning with privileged smooth transition for hexapod locomotion," *IEEE Robotics and Automation Letters*, 2024.

[5] B. W. Budiarto, M. Syahputra, B. S. Nurpriyanto, D. Dwiyanto, and U. Y. Oktiawati, "Design and analysis of bio-inspired robotic systems for search and rescue operations," *The Journal of Academic Science*, vol. 1, no. 4, pp. 408–416, 2024.

[6] X. Jian, Y. Song, D. Liu, Y. Wang, X. Guo, B. Wu, and N. Zhang, "Motion planning and control of active robot in orthopedic surgery by cdmp-based imitation learning and constrained optimization," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 12 197–12 212, 2025.

[7] Z. Zhang, T. Liu, L. Ding, H. Wang, P. Xu, H. Yang, H. Gao, Z. Deng, and J. Pajarinen, "Imitation-enhanced reinforcement learning with privileged smooth transition for hexapod locomotion," *IEEE Robotics and Automation Letters*, vol. 10, no. 1, pp. 350–357, 2025.

[8] G. Li, A. Ijspeert, and M. Hayashibe, "Ai-cpg: Adaptive imitated central pattern generators for bipedal locomotion learned through reinforced reflex neural networks," *IEEE Robotics and Automation Letters*, vol. 9, no. 6, pp. 5190–5197, 2024.

[9] A. Reske, J. Carius, Y. Ma, F. Farshidian, and M. Hutter, "Imitation learning from mpc for quadrupedal multi-gait control," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 5014–5020.

[10] D. Youm, H. Jung, H. Kim, J. Hwangbo, H.-W. Park, and S. Ha, "Imitating and finetuning model predictive control for robust and symmetric quadrupedal locomotion," *IEEE Robotics and Automation Letters*, vol. 8, no. 11, pp. 7799–7806, 2023.

[11] Y. Chen, L. Zhao, J. Ma, and P. Lu, "In-between motion generation based multi-style quadruped robot locomotion," *arXiv preprint arXiv:2507.23053*, 2025.

[12] C. Zhang, J. Sheng, T. Li, H. Zhang, C. Zhou, Q. Zhu, R. Zhao, Y. Zhang, and L. Han, "Learning highly dynamic behaviors for quadrupedal robots," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 9183–9189.

[13] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017.

[14] C. Luo, Q. Zhang, S. Li, H. Chai, W. Xu, and K. Wang, "Behavior generation approach for quadruped robots based on 3d action design and proximal policy optimization," in *2024 IEEE International Conference on Unmanned Systems (ICUS)*. IEEE, 2024, pp. 1088–1093.

[15] X. B. Peng, E. Coumans, T. Zhang, T.-W. Lee, J. Tan, and S. Levine, "Learning agile robotic locomotion skills by imitating animals," *arXiv preprint arXiv:2004.00784*, 2020.

[16] C. Li, S. Blaes, P. Kolev, M. Vlastelica, J. Frey, and G. Martius, "Versatile skill control via self-supervised adversarial imitation of unlabeled mixed motions," *arXiv preprint arXiv:2209.07899*, 2022.

[17] T. Li, Y. Zhang, C. Zhang, Q. Zhu, J. Sheng, W. Chi, C. Zhou, and L. Han, "Learning terrain-adaptive locomotion with agile behaviors by imitating animals," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 339–345.

[18] L. Han, Q. Zhu, J. Sheng, C. Zhang, T. Li, Y. Zhang, H. Zhang, Y. Liu, C. Zhou, R. Zhao, *et al.*, "Lifelike agility and play in quadrupedal robots using reinforcement learning and generative pre-trained models," *Nature Machine Intelligence*, vol. 6, no. 7, pp. 787–798, 2024.

[19] E. Xiao, Y. Dong, J. Ma, and P. Lu, "Stable imitation of multigait and bipedal motions for quadrupedal robots over uneven terrains," *Advanced Robotics Research*, p. 202500036, 2025.

[20] 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.

[21] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa, "Amp: Adversarial motion priors for stylized physics-based character control," *ACM Transactions on Graphics (ToG)*, vol. 40, no. 4, pp. 1–20, 2021.

[22] A. Escontrela, X. 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.

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

[24] J. Liang, V. Makoviychuk, A. Handa, N. Chentanez, M. Macklin, and D. Fox, "Gpu-accelerated robotic simulation for distributed reinforcement learning," in *Conference on Robot Learning*. PMLR, 2018, pp. 270–282.