

Coordinated Humanoid Robot Locomotion with Symmetry Equivariant Reinforcement Learning Policy

Buqing Nie^{*1}, Yang Zhang^{*1}, Rongjun Jin^{*1}, Zhanxiang Cao^{1,2},
Huangxuan Lin¹, Xiaokang Yang¹, Yue Gao^{†1,2}

¹MoE Key Lab of Artificial Intelligence and AI Institute, Shanghai Jiao Tong University, Shanghai, China

² Shanghai Innovation Institute, Shanghai, China

{niebuqing,zhangyang-sjtu-2022,k4rock,caozx1110,b3mylq,xkyang,yuegao}@sjtu.edu.cn

Abstract

The human nervous system exhibits bilateral symmetry, enabling coordinated and balanced movements. However, existing Deep Reinforcement Learning (DRL) methods for humanoid robots neglect morphological symmetry of the robot, leading to uncoordinated and suboptimal behaviors. Inspired by human motor control, we propose Symmetry Equivariant Policy (SE-Policy), a new DRL framework that embeds strict symmetry equivariance in the actor and symmetry invariance in the critic without additional hyperparameters. SE-Policy enforces consistent behaviors across symmetric observations, producing temporally and spatially coordinated motions with higher task performance. Extensive experiments on velocity tracking tasks, conducted in both simulation and real-world deployment with the Unitree G1 humanoid robot, demonstrate that SE-Policy improves tracking accuracy by up to 40% compared to state-of-the-art baselines, while achieving superior spatial-temporal coordination. These results demonstrate the effectiveness of SE-Policy and its broad applicability to humanoid robots.

Introduction

Recently, Deep Reinforcement Learning (DRL) has achieved remarkable achievements in robotics control tasks, including quadruped robots (Nahrendra, Yu, and Myung 2023), manipulators (Singh et al. 2019), bipedal robots (Li et al. 2025), and humanoid robots (He et al. 2024). DRL enables robots to acquire agile skills through interactions in simulation environments autonomously without extensive domain knowledge (Radosavovic et al. 2024).

However, existing DRL policies are typically black-box in nature, failing to leverage the inherent skill sharing offered by the symmetric morphology property of the humanoid robot (Zinkevich and Balch 2001; Van der Pol et al. 2020; Panangaden et al. 2024). Such policies exhibit inconsistent reactions to symmetrically equivalent observations, such as unequal movement styles of symmetric joints during locomotion, revealing a limited understanding to the robot morphology (Apraez et al. 2025; Ding and Gan 2024). Such

oversight leads to asymmetric and uncoordinated behaviors, consequently yielding unnatural and suboptimal policies that negatively impact user experience and diminish task performance (Su et al. 2024; Mittal et al. 2024).

In order to tackle this problem, previous works propose various novel methods to incorporate morphological symmetry into the training framework (Su et al. 2024; Mittal et al. 2024; Abdolhosseini et al. 2019; Wang et al. 2022). Some research improves symmetry performance from *temporal* perspective, i.e. encourages periodicity of robot motions (Lin et al. 2020; Lee et al. 2020; Gu, Wang, and Chen 2024; Ding and Gan 2024). For instance, some prior works (Gu, Wang, and Chen 2024; Ding and Gan 2024) introduce periodic phase signals into the observation and reward design, encouraging the periodical gait motions of the policy. Lee et al. (Lee et al. 2020) design a novel action space based on Central Pattern Generators (Bellegarda and Ijspeert 2022) to describe temporally symmetric motions.

Besides, other works aim to address this problem from a *spatial* perspective, i.e. output equivariant actions under symmetric states (Mittal et al. 2024; Wang et al. 2022; Su et al. 2024). Some works augment collected transitions with their symmetric copies to induce equivariance and invariance for the actor and critic correspondingly, which is effective in quadruped locomotion (Abdolhosseini et al. 2019; Mittal et al. 2024) and manipulation (Lin et al. 2020). Another approach is introducing regularization into the optimization objective of the actor and critic training, which has been widely utilized in RL-based robot methods (Gu, Wang, and Chen 2024; Abreu, Reis, and Lau 2025; Ben et al. 2025; Long et al. 2024; Xue et al. 2025). Prior works also explore enforcing equivariance on RL through introducing hard constraints on neural network-based policy architectures (Mondal, Nair, and Siddiqi 2020; Mondal et al. 2022). This approach is effective on classic control tasks (Van der Pol et al. 2020; Rezaei-Shoshtari et al. 2022), quadruped control (Su et al. 2024), and manipulations (Wang et al. 2022; Wang and Walters 2022).

Despite previous novel methods, the optimal approach to integrating symmetry equivariance into DRL-based robot policies still remains underexplored, particularly on real humanoid robots, which demand high agility and robustness in their motions (Zhuang, Yao, and Zhao 2024; Zhang

^{*}These authors contributed equally.

[†]Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

et al. 2025). Loosely equivariant methods, such as temporal symmetric policies and data augmentation, show moderate performance due to insufficient policy constraints, and are commonly employed as auxiliary techniques in robot tasks (Long et al. 2024; Xue et al. 2025). Loss regularization methods introduce additional hyperparameters requiring delicate tuning, and these terms may impede the optimization process of policies (Mittal et al. 2024; Su et al. 2024). Strict equivariant methods show promise predominantly in simulated classic control tasks, where their effectiveness on robots, especially real humanoid robots, remains under-explored (Van der Pol et al. 2020; Rezaei-Shoshtari et al. 2022; Wang and Walters 2022).

In this work, we propose a new DRL-based method for humanoid robot control tasks, called **Symmetry Equivariant Policy (SE-Policy)**. This method induces strict symmetry equivariance and invariance into the network architectures of the actor and critic respectively. This leads to more coordinated and natural motions without introducing additional hyperparameters. Experiments are conducted in both simulation and on a real humanoid robot through sim-to-real, demonstrating superior locomotion performance compared to previous methods.

The main contributions of this work can be summarized as follows:

- We propose a new method SE-Policy for humanoid robot control tasks, which incorporates strict symmetry equivariance property into actor-critic architecture without additional hyper-parameters.
- SE-Policy generates symmetric and natural motions, achieving higher control performance on tracking error and coordination compared to previous methods.
- Experiments are conducted in both the simulation environments and on the real humanoid robot via sim-to-real, demonstrating superior performance of our method.

Related Works

Equivariant Reinforcement Learning

Equivariance property is actively studied in DRL studies to improve sample efficiency and policy performance (Panangaden et al. 2024; Rezaei-Shoshtari et al. 2022; Wang and Walters 2022). Zinkevich et al. (Zinkevich and Balch 2001) formulate symmetric MDP and prove the equivariance/invariance for actor/critic in RL. Some works implement equivariant policy through data augmentation (Lin et al. 2020; Corrado and Hanna 2024). For example, Luo et al. (Luo, Chen, and Zhang 2024) propose to conduct data augmentation based on Euclidean symmetries, achieving improved data efficiency. Park et al. (Park et al. 2025) propose a novel equivariant RL method to tackle MDP with approximate equivariant structures. Some other works, such as Eqr (Mondal et al. 2022), utilize policy equivariance to improve representation learning, achieving higher sample efficiency. In addition, equivariance has demonstrated excellent performance across various task domains, including discrete action spaces (Van der Pol et al. 2020; Mondal, Nair, and Siddiqi 2020), continuous control (Rezaei-Shoshtari et al.

2022; Yarats et al. 2021), image observations (Nguyen et al. 2023), state observations (Wang and Walters 2022; Panangaden et al. 2024), and multi-agent RL domains (Chen and Zhang 2024; Bousias et al. 2025). In this work, we integrate equivariance into RL policy for humanoid robots, investigating its influence on humanoid robot control tasks.

Symmetry Equivariant Robot Policy

Soft equivariant policy has shown effectiveness in various robot tasks, including path planning (Theile et al. 2024), robot manipulation (Wang et al. 2022), grasping (Hu et al. 2025), quadruped locomotion (Su et al. 2024), and humanoid imitation learning (Long et al. 2024). Policy equivariance is effective to increase sample efficiency (Mittal et al. 2024), enhance motion coordination (Apraez et al. 2025), and improve policy performance (Ben et al. 2025). Current works enhance policy equivariance mainly through reward shaping (Ding and Gan 2024), data augmentation (Mittal et al. 2024), and loss regularization (Xue et al. 2025). These methods induce soft constraints on policies, resulting in loose equivariance with moderate performance. In addition, some methods such as loss regularization may hinder the training of the RL algorithm (Mittal et al. 2024; Su et al. 2024). Moreover, the influence of equivariance on humanoid robot has not been fully investigated.

Method

Problem Formulation

The interaction between the humanoid robot and the environment can be formulated as Markov decision process (MDP) $\mathcal{M} = \langle \mathcal{S}, \mathcal{O}, \mathcal{A}, P, R, \gamma \rangle$, where $s_t \in \mathcal{S}$, $o_t \in \mathcal{O}$, and $a_t \in \mathcal{A}$ denote the robot state, observation, and action at t -th timestep respectively. P denotes the transition probability of the environment. $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ denotes the reward function, and $\gamma \in [0, 1)$ denotes the discount factor. The agent takes action a_t according to its policy π at each step t given the observation, i.e. $a_t \sim \pi$. The objective of the robot is to maximize the cumulative episode return, which is formulated as follows:

$$\pi^* = \arg \max_{\pi} \mathbb{E}_{a_t \sim \pi, s_{t+1} \sim \mathcal{P}} \left[\sum_t \gamma^t R(s_t, a_t) \right]. \quad (1)$$

In order to conduct policy evaluation, we can define the state-action value function

$$Q(s_t, a_t) = R(s_t, a_t) + \gamma \mathbb{E}_{\pi, P} \left[\sum_{t'=t}^{\infty} \gamma^{t'-t} R(s_{t'}, a_{t'}) \right] \quad (2)$$

as the discounted return starting from s_t , given that a_t is taken and then π is followed. The value function $V(s_t) = \mathbb{E}_{a_t \sim \pi} [Q(s_t, a_t)]$ denotes the discounted return starting from s_t following π .

In this work, we utilize SE-Policy for the velocity tracking task of Unitree G1, a versatile humanoid robot frequently used in robotics research (Unitree 2024). Note that this method is a general training framework, thus can be applied to diverse control tasks and humanoid robots that possess morphological symmetry.

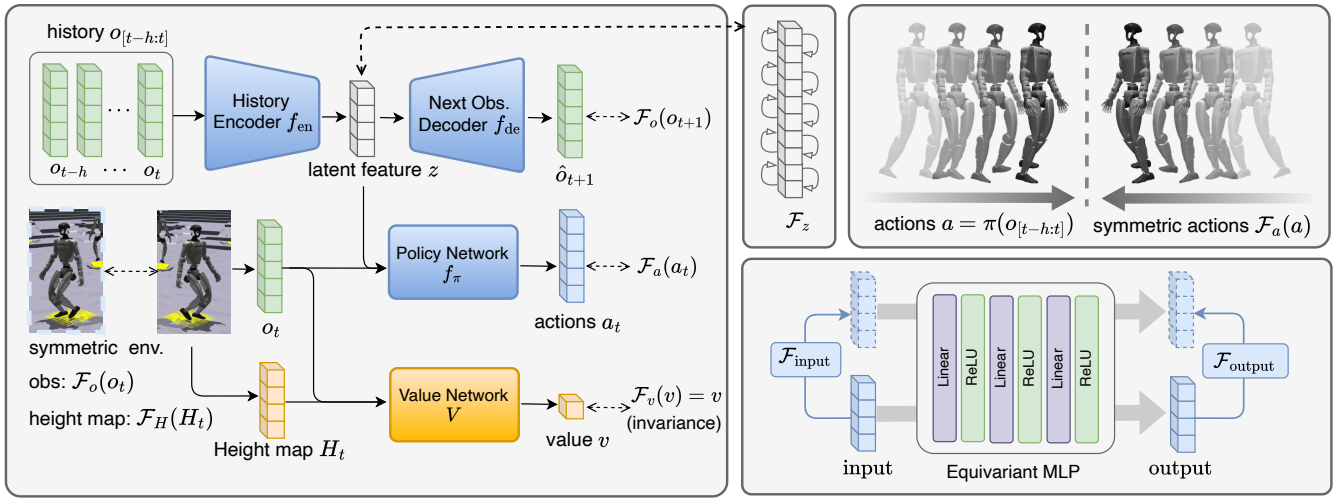


Figure 1: The overall architecture of *SE-Policy*. **(a) Left:** the architecture of the actor and critic model. **(b) upper right:** the visualization of \mathcal{F}_z , i.e. the symmetric transformation of z . The visualization of humanoid robot motions and corresponding symmetric motions. **(b) bottom right:** the description of equivariant MLP, which is widely utilized in this work.

Symmetry Equivariant Policy

Observation Space In this work, the observation of the humanoid robot $o_t \in \mathbb{R}^{96}$ is described in Table 1, i.e.

$$o_t = [\omega \quad g \quad c \quad \theta \quad v \quad a_{t-1} \quad \Phi]^T, \quad (3)$$

where the user-provided velocity command $c = (c_x, c_y, c_\omega)$ specifies the desired linear velocities along the x and y axes (c_x, c_y), and the angular velocity c_ω . The phase input Φ is a sinusoidal clock signal, which is employed to generate reference motion cycles. This design is widely used in RL-based control works to improve temporal symmetry (periodicity) of the gait. The height map H describes robot-centric terrain information and is utilized for critic training only.

Action Space This work employs position control for the humanoid robot. Accordingly, the action space $|\mathcal{A}| = 27$ denotes the target joint positions of each joint. The detailed description of the observation and action space is given in Appendix 1.3.

Symmetry in MDP As shown in Fig. 1, the humanoid robots are designed following humanoid morphology, thus exhibiting reflection symmetry in their structure. Consequently, the MDP exhibits a similar symmetry property described as follows.

As illustrated in Table 1, the *symmetry transformation* denotes the formulation after applying reflection symmetry. To facilitate reading, we denote the symmetric transformation function of the state, observation, and action as \mathcal{F}_s , \mathcal{F}_o and \mathcal{F}_a , respectively. For instance, o_t and $\mathcal{F}_o(o_t)$ represent the robot's observation at t -th step before and after reflection symmetry operations, respectively.

Given the MDP \mathcal{M} described in Sec. , we can find that:

- The transition probability P remains invariant under \mathcal{F}_o and \mathcal{F}_a transformations, i.e.

$$P(\mathcal{F}_s(s') | \mathcal{F}_s(s), \mathcal{F}_a(a)) = P(s' | s, a). \quad (4)$$

- The reward function R is also invariant:

$$R(\mathcal{F}_s(s), \mathcal{F}_a(a)) = R(s, a). \quad (5)$$

The above invariance properties are derived from the morphological symmetry of the humanoid robot structure. In other words, the robot takes symmetric actions under symmetric states correspondingly, it will obtain same rewards and reach symmetric states.

Based on Eq. (2) and the above formulations, we can obtain the invariance property of critic functions Q^* and V^* of the optimal policy π^* :

$$Q(\mathcal{F}_s(s), \mathcal{F}_a(a)) = Q(s, a), \quad V(\mathcal{F}_s(s)) = V(s). \quad (6)$$

This means symmetric states correspond to same values, i.e. same expected cumulative reward in future. Based on the optimization objective shown in Eq. (1), we can derive the symmetry equivariance of the optimal policy π^* :

$$\pi^*(\mathcal{F}_s(s)) = \mathcal{F}_a(\pi^*(s)), \quad (7)$$

i.e. the symmetric states correspond to symmetric optimal actions. This property is consistent with application practice empirically. For example, when the robot is balanced on its left foot with the right foot unsupported, its optimal action is to lower the right foot. Correspondingly, when the robot is in a symmetric state with its left foot unsupported, it needs to take symmetric actions, i.e. lower its left foot.

Model Architecture

As the overall architecture illustrated in Fig. 1, the new method is composed of a history encoder f_{en} , an observation decoder f_{de} , a policy network f_{π} as actor, and a value network V as critic. This architecture is inspired by DreamWaQ (Nahrendra, Yu, and Myung 2023), which is a state-of-the-art robot locomotion method based on Proximal Policy Gradient (PPO) (Schulman et al. 2017).

Component	Dim	Description	Symmetry Transformation \mathcal{F}
Base angular velocity ω	$o_t^{1:3}$	$(\omega_x, \omega_y, \omega_z)$	$(-\omega_x, \omega_y, -\omega_z)$
Projected gravity g	$o_t^{4:6}$	(g_x, g_y, g_z)	$(g_x, -g_y, g_z)$
Velocity commands c	$o_t^{7:10}$	(c_x, c_y, c_ω)	$(c_x, -c_y, -c_\omega)$
Joint positions θ	$o_t^{11:37}$	$(\theta_{\text{left}}^{\text{arm}}, \theta_{\text{right}}^{\text{arm}}, \theta_{\text{left}}^{\text{leg}}, \theta_{\text{right}}^{\text{leg}}, \theta_{\text{waist}})$	$(-\theta_{\text{right}}^{\text{arm}}, -\theta_{\text{left}}^{\text{arm}}, -\theta_{\text{right}}^{\text{leg}}, -\theta_{\text{left}}^{\text{leg}}, -\theta_{\text{waist}})$
Joint velocities v	$o_t^{38:65}$	$(v_{\text{left}}^{\text{arm}}, v_{\text{right}}^{\text{arm}}, v_{\text{left}}^{\text{leg}}, v_{\text{right}}^{\text{leg}}, v_{\text{waist}})$	$(-v_{\text{right}}^{\text{arm}}, -v_{\text{left}}^{\text{arm}}, -v_{\text{right}}^{\text{leg}}, -v_{\text{left}}^{\text{leg}}, -v_{\text{waist}})$
Previous action a_{t-1}	$o_t^{66:93}$	$(a_{\text{left}}^{\text{arm}}, a_{\text{right}}^{\text{arm}}, a_{\text{left}}^{\text{leg}}, a_{\text{right}}^{\text{leg}}, a_{\text{waist}})$	$(-a_{\text{right}}^{\text{arm}}, -a_{\text{left}}^{\text{arm}}, -a_{\text{right}}^{\text{leg}}, -a_{\text{left}}^{\text{leg}}, -a_{\text{waist}})$
Phase input Φ	$o_t^{94:96}$	$(\Phi_{\sin}, \Phi_{\cos})$	$(-\Phi_{\sin}, -\Phi_{\cos})$
Height map H	$H_t^{1:187}$	$(H_{\text{left}}, H_{\text{middle}}, H_{\text{right}})$	$(H_{\text{right}}, H_{\text{middle}}, H_{\text{left}})$
Action a_t	$a_t^{1:27}$	$(a_{\text{left}}^{\text{arm}}, a_{\text{right}}^{\text{arm}}, a_{\text{left}}^{\text{leg}}, a_{\text{right}}^{\text{leg}}, a_{\text{waist}})$	$(-a_{\text{right}}^{\text{arm}}, -a_{\text{left}}^{\text{arm}}, -a_{\text{right}}^{\text{leg}}, -a_{\text{left}}^{\text{leg}}, -a_{\text{waist}})$

Table 1: The description and dimensions for each observation and action component. The reflection transformation represents the formulation obtained after applying reflection symmetry. The height map H is utilized as critic input.

Symmetry Equivariant Actor In this work, the actor makes decisions based on temporal observation history $o_{[t-h:t]} = [o_{t-h} \ o_{t-h+1} \ \dots \ o_t]^T$, where h denotes the history length.

As shown in Fig. 1, the *history encoder* f_{en} inputs history $o_{[t-h:t]}$ and generates latent feature z . In order to train the history encoder f_{en} to extract appropriate features from the history, an *observation decoder* f_{de} is utilized to predict next observation o_{t+1} , thus is trained though

$$\mathcal{L}_{\text{AE}} = \text{MSE}(\hat{o}, o_{t+1}), \hat{o} = f_{\text{de}}(f_{\text{en}}(o_{[t-h:t]})). \quad (8)$$

Afterwards, the *policy network* f_{π} makes decisions based on the current observation o_t and the latent feature z , i.e. $a_t = f_{\pi}(o_t, z)$.

Symmetry Invariant Critic As shown in Fig. 1, the critic evaluates the policy using the value network $V(H_t, o_t)$, where height map H is privileged information. As described in Eq. (6), the critic is invariant to the symmetric transformation of the input observation.

Equivariant Neural Network In order to incorporate symmetry equivariance described in Eq. (7) and Eq. (6), all networks are constructed utilizing Linear layers and ReLU activations following ESCNN (Cesa, Lang, and Weiler 2022), which implements equivariance based on parameter sharing. As shown in Fig. 1 (bottom right), all equivariant MLPs are designed with symmetry equivariance, given symmetry transformations for inputs and outputs denoted as $\mathcal{F}_{\text{input}}$ and $\mathcal{F}_{\text{output}}$. Besides, we define the symmetry transformation of latent feature as \mathcal{F}_z shown as follows. Given latent feature z with even size, i.e. $|z| \bmod 2 = 0$, we define $\mathcal{F}_z(z)$:

$$[\mathcal{F}_z(z)]_i = \begin{cases} z_{i+1} & \text{if } i \text{ is odd,} \\ z_{i-1} & \text{if } i \text{ is even.} \end{cases} \quad (9)$$

Based on the symmetry transformation described in Eq. (9) and Table 1, we construct equivariant actor and critic shown in Fig. 1.

Training Framework

The training process is mainly consistent with the standard PPO algorithm. The critic is trained utilizing MSE loss

with ground truth obtained from the trajectory buffer \mathcal{D} , i.e. $\mathcal{L}_V = \text{MSE}(V(H_t, o_t), y)$, where y is the reward-to-go and is utilized as training labels.

The actor is trained utilizing \mathcal{L}_{AE} shown in Eq. (8) and \mathcal{L}_{PPO} shown as follows:

$$\mathcal{L}_{\text{PPO}} = \mathbb{E}_{\mathcal{D}} [\min(\rho_{\pi} A(s, a), g(\rho_{\pi}) A(s, a))], \quad (10)$$

$\rho_{\pi} = \frac{\pi(a|o_t, H)}{\pi_{\text{old}}(a|o_t, H)}$, and $g(\rho_{\pi}) = \text{clip}(\rho_{\pi}, 1 - \xi, 1 + \xi)$. ξ is a small hyper-parameter to limit the magnitude of the update, promoting stable and controlled updates. $A(s, a)$ denotes the advantage of taking a at state s , which is obtained utilizing Generalized Advantage Estimation (Schulman et al. 2015).

The reward functions employed during training are illustrated in Appendix 1.1, which consist of three key components: (1) tracking rewards for linear and angular velocity commands; (2) balance maintenance, such as penalties on velocity on z -axis to ensure stability; (3) regularization terms, such as penalties on action oscillations and torque exceedance to encourage smooth and reasonable actions.

In order to improve training efficiency and stability, curriculum learning is employed during the training process (Margolis et al. 2024). We set different difficulty levels for terrains, task commands, and sensor measurement noise. The terrains consist of flat, rough, discrete, and slope terrains. During training, the task difficulty progressively increases as policy performance improves. In addition, domain randomization is utilized to facilitate the real-world deployment of the policy. The detailed settings of domain randomization are described in Appendix 1.2. As shown in Table 2 in the appendix, we implement randomization encompassing ground friction, mass properties, center-of-mass positions, motor parameters, etc.

Experiment

In this section, we conduct experiments on the humanoid robot to investigate the following questions:

- (1) Can the proposed method be effectively integrated into current DRL framework for humanoid locomotion tasks?
- (2) Can the new method generate equivariant policy and improve task performance effectively?

- (3) Can the symmetry equivariance property contribute to more stable and coordinated robot motions?

Experimental Setup

In this work, we conduct a series of experiments on the Unitree G1 (Unitree 2024), which is a 27-DoF versatile humanoid robot commonly utilized in robot research. In this experiment, all the methods are trained utilizing the NVIDIA Isaac Gym simulator (Makoviychuk et al. 2021).

Baseline Methods In this experiment, the following three methods are used as baselines:

- (1) **DreamWaQ**: A state-of-the-art model-free DRL algorithm for legged robot locomotion (Nahendra, Yu, and Myung 2023).
- (2) **DreamWaQ-Regu**: Enhance DreamWaQ through introducing soft regularization term \mathcal{L}_{reg} in policy updates.

$$\mathcal{L}_{\text{reg}} = \left\| \pi(\mathcal{F}_o(o_{[t-h:t]})) - \mathcal{F}_a(\pi(o_{[t-h:t]})) \right\|^2. \quad (11)$$

This auxiliary term penalizes asymmetrical actuation patterns between bilateral joints, which has been widely used in existing works (Abreu, Reis, and Lau 2025; Ben et al. 2025; Long et al. 2024; Xue et al. 2025).

- (3) **SE-Policy (actor only)**: Replace critic network of SE-Policy with a vanilla critic based on MLP. This ablation study investigates the influence of the critic function’s invariance property on humanoid robot performance.

For a fair comparison, all hyper-parameters and environments are kept consistent across various methods. All methods are trained for over 5K iterations in approximately 4 hours utilizing NVIDIA RTX 4090 GPU devices. More details of the implementation are given in Appendix 2.1.

Evaluation Metrics In this experiment, the following five quantitative metrics are utilized to evaluate the performance of each method:

- (1) **Tracking Error (TE)** evaluates the velocity tracking performance of the policy. In detail, we analyze the tracking performance from the following three perspective:

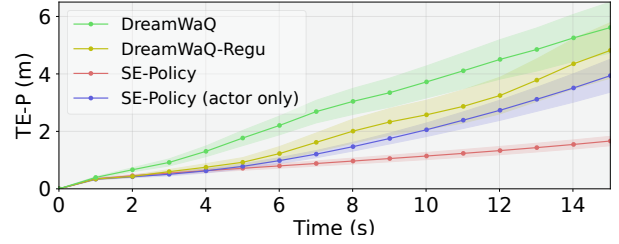
- **TE-Velocity** measures the average velocity difference between command and the real velocity vector:

$$\text{TE-V}(\pi) = \mathbb{E}_{\pi} [\|v_t - v_{\text{cmd}}\|].$$

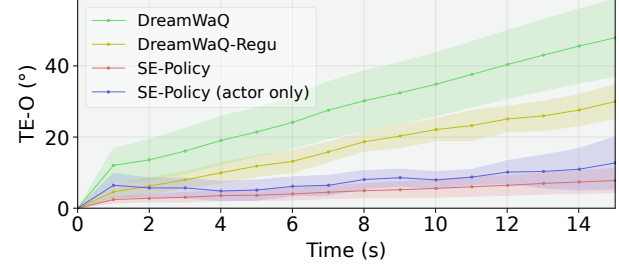
- **TE-Position** measures the positional error between the robot’s actual and desired positions.
- **TE-Orientation** measures the angle deviation between the robot’s actual and directions.

- (2) **Temporal Symmetry (Temp-S)** quantifies the difference between joint actions and their symmetric joints’ actions in the subsequent half period. This metric measures the temporal symmetric periodicity during the humanoid locomotion tasks, where lower values corresponding to higher motion coordination.

- (3) **Spatial Symmetry (Spat-S)** computes the difference between the action $\pi(o_{[t-h:t]})$ and the symmetric action of symmetric observation, i.e. $\mathcal{F}_a(\pi(\mathcal{F}_o(o_{[t-h:t]})))$.



(a) Tracking Error Position (TE-P)



(b) Tracking Error Orientation (TE-O)

Figure 2: The tracking errors in terms of position (TE-P) and orientation (TE-O) over locomotion time. Lines and shadow areas denote mean values and standard errors. SE-Policy (red) achieves lower TE-P and TE-O over time than other methods, validating the effectiveness of our method.

This metric measures symmetry equivariance property of the policy network directly, where lower values denote higher performance.

More details of these metrics are described in Appendix 2.2.

Experiment Result

Task Performance In this experiment, the robot is requested to track the random commands in the simulation environment, where velocity commands are randomly sampled with $|v_x| < 0.8m/s$, $|v_y| < 0.8m/s$, and $|\omega| < 0.5 \text{ rad/s}$.

As the results shown in Table 2, our method achieves superior tracking performance. Specifically, the TE-V of SE-Policy 9.85cm/s is 40.0% lower than DreamWaQ (16.43 cm/s) and 19.2% lower than DreamWaQ-Regu (12.19 cm/s), demonstrating the effectiveness of our method. Besides, the Spat-S of SE-Policy is 0.0, corresponding to the strict symmetry equivariance introduced in the previous section. The Temp-S of SE-Policy is also lower than other methods, indicating higher symmetric periodicity. For example, the motion of the left leg in the current phase is more consistent with the right leg in the next phase, which means a more coordinated locomotion style.

The results of Tracking Error-Position (TE-P) and Tracking Error-Orientation (TE-O) are given in Fig. 2. The x-axis and y-axis denote time and the error value, respectively. As shown in figures, both TE-P and TE-O increase over time because of the cumulative error. However, SE-Policy (red line) achieves lower position errors and orientation errors than other methods, because its symmetry equivariance property facilitates more accurate control on velocity and direction.

Metric	DreamWaq	DreamWaq-Regu	SE-Policy	SE-Policy (actor only)
TE-V (cm/s)	16.43 ± 9.54	13.91 ± 8.53	9.85 ± 1.54	11.06 ± 8.63
Temp-S (10^{-2} rad)	22.52 ± 2.70	16.58 ± 2.88	7.86 ± 1.44	9.20 ± 2.19
Spat-S (10^{-2} rad)	30.84 ± 5.201	8.18 ± 1.46	0.00 ± 0.00	0.00 ± 0.00

Table 2: The experiment results on Tracking Error-Velocity, Temporal Symmetry, and Spatial Symmetry scores. Lower values means higher performance. The bold scores in the table indicate the optimal results.

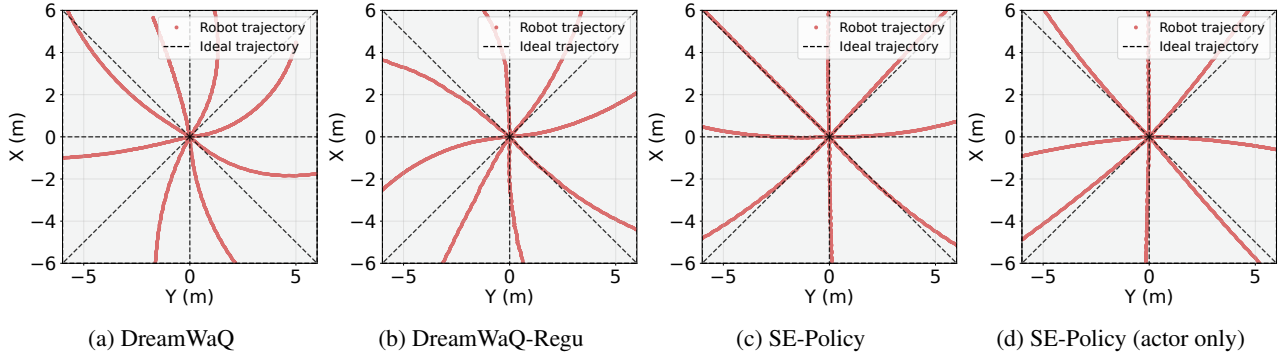


Figure 3: Visualization of each method’s locomotion trajectories. The robot is requested to move from the center to eight velocity directions, where dotted lines and red lines are ideal and real trajectories, respectively. Our method shown in Fig. 3c outperforms baseline methods on tracking accuracy and trajectory symmetry.

More details are described in the analysis of the motion visualization.

As illustrated in Table 2 and Fig. 2, DreamWaq-Regu (yellow line) generally achieves higher performance than vanilla DreamWaq. For instance, DrwamWaq-Regu obtains 13.91 TE-V value, 15.3% lower than vanilla DreamWaq (16.43), demonstrating the effectiveness of symmetry equivariance achieved by loss regularization shown in Eq. (11). However, as the Spat-S and Temp-S results shown in Table 2, the symmetry performance of DrwamWaq-Regu is weaker than SE-Policy. This is because the equivariance property of DrwamWaq-Regu is induced by loss regularization, which is soft and can be violated during inference, corresponding to nonzero Spat-S value in Table 2.

In addition, despite achieving superior performance over DreamWaq, SE-Policy (actor only) suffers a non-negligible performance reduction relative to the complete SE-Policy. For example, SE-Policy (actor only) obtains higher TE-V (12.2%) and Temp-S (17.0%) than complete SE-Policy, which means inferior tracking performance and temporal asymmetric motions. This suggests the necessity of the symmetry invariant critic during the optimization process of equivariant policies in SE-Policy.

Locomotion Trajectory Visualization In this section, in order to give a more intuitive explanation of the difference between SE-Policy and other methods, we visualize the trajectories of each policy in Fig. 3. As shown in the figures, the robot is placed at the center and commanded to move towards eight directions shown as dotted lines, where $v_x, v_y \in \{-0.5m/s, 0m/s, 0.5m/s\}$. The dotted lines and red lines denote ideal and real trajectories of the robot corre-

spondingly. The size of the plane is (12m, 12m). The results illustrated in Fig. 3 are analyzed as follows.

- (1) **Tracking Accuracy:** As shown in Fig. 3c, robot trajectories of SE-Policy are generally consistent with ideal path (dotted lines), corresponding to the low TE-P and TE-O values described in Fig. 2. However, the paths of DreamWaq in Fig. 3a deviate from the commanded routes, with cumulative deviation over time, resulting in high tracking errors shown in Table 2. Besides, DreamWaq-Regu (Fig. 3b) achieves better performance than DreamWaq, where the robot stay closer to the planned routes.
- (2) **Trajectory Symmetry:** The trajectories in Fig. 3c and Fig. 3d generally exhibit mirror symmetry with respect to the $y = 0$ m line, corresponding to the equivariance property of two policies. However, the paths of DreamWaq-Regu in Fig. 3b fails to exhibit this result, where paths deviates from commands with a tendency to turn left. This is because the equivariance regularization is quite loose, thus the policy cannot be constrained adequately.

Above all, the results in Fig. 3 demonstrate the effectiveness of strict equivariance induced in SE-Policy, and the necessity of symmetry invariance in critic training.

Motion Visualization and Analysis In this section, we analyze the policy performance through visualizing robot motions. As described in Fig. 4, we collect the height of robot feet during locomotion on the plane given constant velocity commands, where two feet are distinguished by two colors. As shown in Fig. 4c, SE-Policy enables two feet to maintain highly periodic motion with nearly constant phase differences, corresponding to the best Temp-S score given in

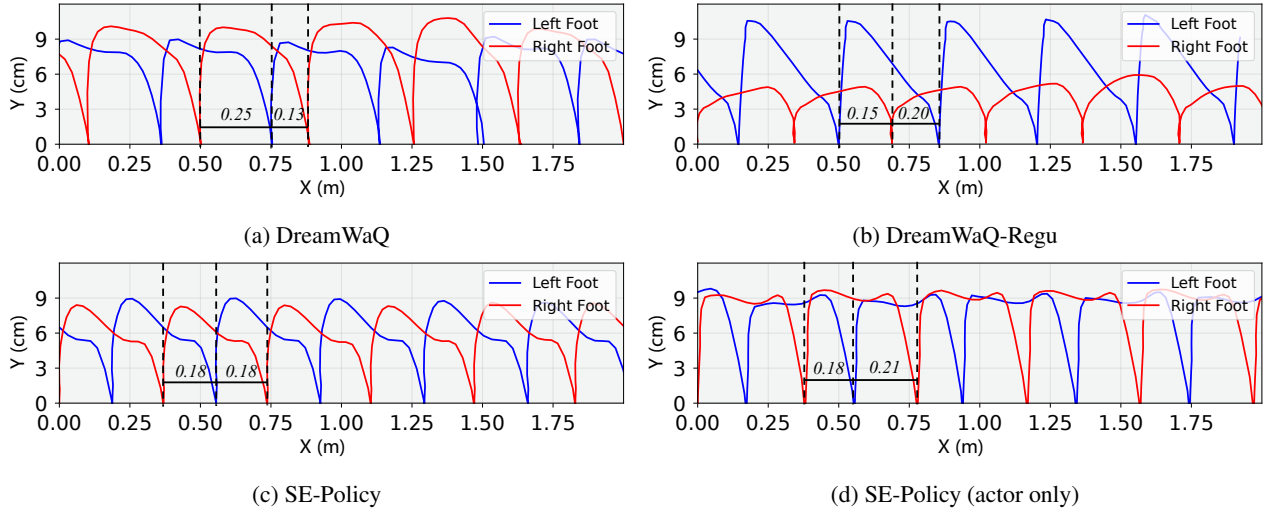


Figure 4: The visualization of foot movement during locomotion on the plane, where x -axis denotes moving distance of the robot torso, and y -axis denotes the height of two feet. Our method achieves consistent motions for two feet. The motions of two feet generated by SE-Policy are consistent with identical amplitudes, step sizes, and temporal coordination.

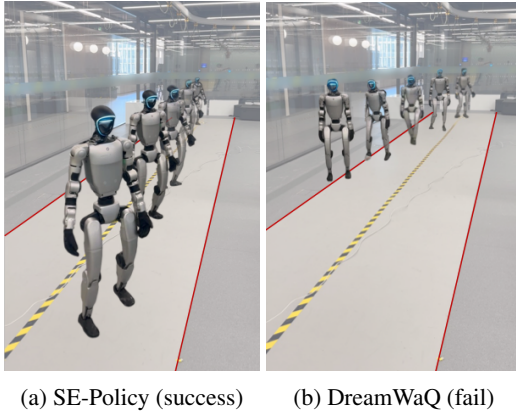


Figure 5: Real world experiments to validate the effectiveness of SE-Policy. The robot tracks given velocity (warning line) without stepping out of the boundary (red line). Please refer to the attached video for more details.

Table 2. Besides, two feet are controlled with identical amplitudes and step sizes, resulting in improved coordination stability during robotic movement.

Additionally, DreamWaQ’s movements are quite uncoordinated with inferior symmetric consistency. As shown in Fig. 4a, there exists non-negligible difference between the motion style of two feet, corresponding to unsatisfying temporal symmetry performance in Table 2. The mean step sizes of two feet are $0.13m$ and $0.25m$, leading to uncoordinated motions. As shown in Fig. 4b, DreamWaQ-Regu achieves more consistent step sizes through loss regularization, following the phase signal shown in Table 1. However, there exists significant difference between feet motions on styles and amplitudes, corresponding to inferior Temp-S performance compared to SE-Policy. This phenomenon suggests that loss

regularization described in Eq. (11) is effective to improve spatial symmetry, but may cause inconsistent motion styles of symmetric joints, leading to unsatisfying Temp-S performance shown in Table 2.

Real World Experiment

In this work, we also conduct experiments on the real humanoid robot through Sim-to-real. We conduct experiment shown in Fig. 3 on the real robot, where the robot is commanded to track multiple velocity for 10m without stepping out of the boundary (red line). Part of results are shown in Fig. 5. The result is generally consistent with the simulation experiment, where SE-Policy achieves higher tracking accuracy on velocity, position, and orientations, compared to baseline methods. Besides, we deploy policies on the real robot to evaluate its performance on different terrains, including grass, slope, sand, and stones. SE-Policy achieves traversability on common unstructured terrains, demonstrating its effectiveness in real applications. More experimental results can be found in the supplementary video.

Conclusions

In this work, we propose Symmetry Equivariant Policy (SE-Policy), a novel DRL method for humanoid robot control tasks. Inspired by the inherent symmetric morphology of humanoid robots, SE-Policy integrates strict symmetry equivariance and invariance into the actor and critic architecture respectively, without requiring additional hyperparameters during training. SE-Policy achieves more coordinated and natural robot motions with temporal and spatial symmetry properties. Multiple experiments are conducted in both simulation and on a real humanoid robot. The results reveal that SE-Policy consistently outperforms existing methods in terms of tracking accuracy on velocity, position, and orientation, demonstrating the effectiveness of our method.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 92248303 and No. 62373242), the Shanghai Municipal Science and Technology Major Project (Grant No. 2021SHZDZX0102), and the Fundamental Research Funds for the Central Universities.

References

- Abdolhosseini, F.; Ling, H. Y.; Xie, Z.; Peng, X. B.; and Van de Panne, M. 2019. On learning symmetric locomotion. In *Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games*, 1–10.
- Abreu, M.; Reis, L. P.; and Lau, N. 2025. Addressing imperfect symmetry: A novel symmetry-learning actor-critic extension. *Neurocomputing*, 614: 128771.
- Apraetz, D. O.; Turrisi, G.; Kostic, V.; Martin, M.; Agudo, A.; Moreno-Noguer, F.; Pontil, M.; Semini, C.; and Mastalli, C. 2025. Morphological symmetries in robotics. *The International Journal of Robotics Research*, 02783649241282422.
- Bellegarda, G.; and Ijspeert, A. 2022. CPG-RL: Learning central pattern generators for quadruped locomotion. *IEEE Robotics and Automation Letters*, 7(4): 12547–12554.
- Ben, Q.; Jia, F.; Zeng, J.; Dong, J.; Lin, D.; and Pang, J. 2025. Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit. *arXiv preprint arXiv:2502.13013*.
- Bousias, N.; Pertigkiozoglou, S.; Daniilidis, K.; and Pappas, G. 2025. Symmetries-enhanced Multi-Agent Reinforcement Learning. In *7th Annual Learning for Dynamics & Control Conference*, 999–1011. PMLR.
- Cesa, G.; Lang, L.; and Weiler, M. 2022. A program to build E (N)-equivariant steerable CNNs. In *International conference on learning representations*.
- Chen, D.; and Zhang, Q. 2024. E (3) -Equivariant Actor-Critic Methods for Cooperative Multi-Agent Reinforcement Learning. In *International Conference on Machine Learning*, 7640–7666. PMLR.
- Corrado, N.; and Hanna, J. P. 2024. Understanding when Dynamics-Invariant Data Augmentations Benefit Model-free Reinforcement Learning Updates. In *The Twelfth International Conference on Learning Representations*.
- Ding, J.; and Gan, Z. 2024. Breaking symmetries leads to diverse quadrupedal gaits. *IEEE Robotics and Automation Letters*.
- Gu, X.; Wang, Y.-J.; and Chen, J. 2024. Humanoid-gym: Reinforcement learning for humanoid robot with zero-shot sim2real transfer. *arXiv preprint arXiv:2404.05695*.
- He, T.; Luo, Z.; He, X.; Xiao, W.; Zhang, C.; Zhang, W.; Kitani, K. M.; Liu, C.; and Shi, G. 2024. OmniH2O: Universal and Dexterous Human-to-Humanoid Whole-Body Teleoperation and Learning. In *8th Annual Conference on Robot Learning*.
- Hu, B.; Zhu, X.; Wang, D.; Dong, Z.; Huang, H.; Wang, C.; Walters, R.; and Platt, R. 2025. OrbitGrasp: SE (3)-Equivariant Grasp Learning. In *Conference on Robot Learning*, 2456–2474. PMLR.
- Lee, J.; Hwangbo, J.; Wellhausen, L.; Koltun, V.; and Hutter, M. 2020. Learning quadrupedal locomotion over challenging terrain. *Science robotics*, 5(47): eabc5986.
- Li, Z.; Peng, X. B.; Abbeel, P.; Levine, S.; Berseth, G.; and Sreenath, K. 2025. Reinforcement learning for versatile, dynamic, and robust bipedal locomotion control. *The International Journal of Robotics Research*, 44(5): 840–888.
- Lin, Y.; Huang, J.; Zimmer, M.; Guan, Y.; Rojas, J.; and Weng, P. 2020. Invariant transform experience replay: Data augmentation for deep reinforcement learning. *IEEE Robotics and Automation Letters*, 5(4): 6615–6622.
- Long, J.; Ren, J.; Shi, M.; Wang, Z.; Huang, T.; Luo, P.; and Pang, J. 2024. Learning humanoid locomotion with perceptive internal model. *arXiv preprint arXiv:2411.14386*.
- Luo, J.; Chen, D.; and Zhang, Q. 2024. Reinforcement learning with euclidean data augmentation for state-based continuous control. *Advances in Neural Information Processing Systems*, 37: 90253–90276.
- Makoviychuk, V.; Wawrzyniak, L.; Guo, Y.; Lu, M.; Storey, K.; Macklin, M.; Hoeller, D.; Rudin, N.; Allshire, A.; Handa, A.; and State, G. 2021. Isaac Gym: High Performance GPU-Based Physics Simulation For Robot Learning. *arXiv:2108.10470*.
- Margolis, G. B.; Yang, G.; Paigwar, K.; Chen, T.; and Agrawal, P. 2024. Rapid locomotion via reinforcement learning. *The International Journal of Robotics Research*, 43(4): 572–587.
- Mittal, M.; Rudin, N.; Klemm, V.; Allshire, A.; and Hutter, M. 2024. Symmetry considerations for learning task symmetric robot policies. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 7433–7439. IEEE.
- Mondal, A. K.; Jain, V.; Siddiqi, K.; and Ravanbakhsh, S. 2022. Eqr: Equivariant representations for data-efficient reinforcement learning. In *International Conference on Machine Learning*, 15908–15926. PMLR.
- Mondal, A. K.; Nair, P.; and Siddiqi, K. 2020. Group equivariant deep reinforcement learning. *arXiv preprint arXiv:2007.03437*.
- Nahrendra, I. M. A.; Yu, B.; and Myung, H. 2023. Dreamwaq: Learning robust quadrupedal locomotion with implicit terrain imagination via deep reinforcement learning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 5078–5084. IEEE.
- Nguyen, H. H.; Baisero, A.; Klee, D.; Wang, D.; Platt, R.; and Amato, C. 2023. Equivariant reinforcement learning under partial observability. In *Conference on Robot Learning*, 3309–3320. PMLR.
- Panangaden, P.; Rezaei-Shoshtari, S.; Zhao, R.; Meger, D.; and Precup, D. 2024. Policy Gradient Methods in the Presence of Symmetries and State Abstractions. *Journal of Machine Learning Research*, 25(71): 1–57.
- Park, J. Y.; Bhatt, S.; Zeng, S.; Wong, L. L.; Koppel, A.; Ganesh, S.; and Walters, R. 2025. Approximate Equivariance in Reinforcement Learning. In *International Conference on Artificial Intelligence and Statistics*, 4177–4185. PMLR.

- Radosavovic, I.; Xiao, T.; Zhang, B.; Darrell, T.; Malik, J.; and Sreenath, K. 2024. Real-world humanoid locomotion with reinforcement learning. *Science Robotics*, 9(89): eadi9579.
- Rezaei-Shoshtari, S.; Zhao, R.; Panangaden, P.; Meger, D.; and Precup, D. 2022. Continuous mdp homomorphisms and homomorphic policy gradient. *Advances in Neural Information Processing Systems*, 35: 20189–20204.
- Schulman, J.; Moritz, P.; Levine, S.; Jordan, M.; and Abbeel, P. 2015. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.
- Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Singh, A.; Yang, L.; Hartikainen, K.; Finn, C.; and Levine, S. 2019. End-to-end robotic reinforcement learning without reward engineering. *arXiv preprint arXiv:1904.07854*.
- Su, Z.; Huang, X.; Ordoñez-Apraéz, D.; Li, Y.; Li, Z.; Liao, Q.; Turrisi, G.; Pontil, M.; Semini, C.; Wu, Y.; et al. 2024. Leveraging symmetry in rl-based legged locomotion control. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 6899–6906. IEEE.
- Theile, M.; Cao, H.; Caccamo, M.; and Sangiovanni-Vincentelli, A. L. 2024. Equivariant ensembles and regularization for reinforcement learning in map-based path planning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 14164–14171. IEEE.
- Unitree. 2024. Humanoid robot G1_Humanoid Robot Functions_Humanoid Robot Price — Unitree Robotics — unitree.com. <https://www.unitree.com/g1>.
- Van der Pol, E.; Worrall, D.; van Hoof, H.; Oliehoek, F.; and Welling, M. 2020. Mdp homomorphic networks: Group symmetries in reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 4199–4210.
- Wang, D.; Jia, M.; Zhu, X.; Walters, R.; and Platt, R. 2022. On-robot learning with equivariant models. *arXiv preprint arXiv:2203.04923*.
- Wang, D.; and Walters, R. 2022. SO (2) Equivariant Reinforcement Learning. In *International Conference on Learning Representations*.
- Xue, Y.; Dong, W.; Liu, M.; Zhang, W.; and Pang, J. 2025. A Unified and General Humanoid Whole-Body Controller for Fine-Grained Locomotion. *arXiv preprint arXiv:2502.03206*.
- Yarats, D.; Zhang, A.; Kostrikov, I.; Amos, B.; Pineau, J.; and Fergus, R. 2021. Improving sample efficiency in model-free reinforcement learning from images. In *Proceedings of the aaai conference on artificial intelligence*, volume 35, 10674–10681.
- Zhang, H.; Zhang, L.; Chen, Z.; Chen, L.; Wang, Y.; and Xiong, R. 2025. Natural Humanoid Robot Locomotion with Generative Motion Prior. *arXiv preprint arXiv:2503.09015*.
- Zhuang, Z.; Yao, S.; and Zhao, H. 2024. Humanoid Parkour Learning. In *8th Annual Conference on Robot Learning*.
- Zinkevich, M.; and Balch, T. R. 2001. Symmetry in markov decision processes and its implications for single agent and multiagent learning. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 632.