Keep on Going: Learning Robust Humanoid Motion Skills via Selective Adversarial Training

Yang Zhang[1], Zhanxiang Cao[1,2], Buqing Nie[1], Haoyang Li[1,2], Zhong Jiangwei[3], Qiao Sun[3], Xiaoyi Hu[3], Xiaokang Yang[1], Yue Gao[1,2*]

[1]MoE Key Lab of Artificial Intelligence and AI Institute, Shanghai Jiao Tong University
[2]Shanghai Innovation Institute, Shanghai, China
[3]Lenovo Research
{zhangyang-sjtu-2022, yuegao}@sjtu.edu.cn

Abstract

Humanoid robots are expected to operate reliably over long horizons while executing versatile whole-body skills. Yet Reinforcement Learning (RL) motion policies typically lose stability under prolonged operation, sensor/actuator noise, and real world disturbances. In this work, we propose a **S**elective **A**dversarial **A**ttack for **R**obust **T**raining (**SA2RT**) to enhance the robustness of motion skills. The adversary is learned to identify and sparsely perturb the most vulnerable states and actions under an attack-budget constraint, thereby exposing true weakness without inducing conservative overfitting. The resulting non-zero sum, alternating optimization continually strengthens the motion policy against the strongest discovered attacks. We validate our approach on the Unitree G1 humanoid robot across perceptive locomotion and whole-body control tasks. Experimental results show that adversarially trained policies improve the terrain traversal success rate by 40%, reduce the trajectory tracking error by 32%, and maintain long horizon mobility and tracking performance. Together, these results demonstrate that selective adversarial attacks are an effective driver for learning robust, long horizon humanoid motion skills.

Extended version — https://arxiv.org/abs/2507.08303

Figure 1: Snapshots of the humanoid robot executing whole-body trajectory tracking. WBC-SAP can track challenging dynamic trajectories over an extended duration, demonstrating that the SA2RT significantly improves the robustness of motion policies.

Introduction

Humanoid robots, with their human-like morphology and versatile mobility (Gu et al. 2025), have the potential to replace humans performing various tasks in daily life (Tong, Liu, and Zhang 2024), demonstrating significant potential in fields such as domestic service (Mende et al. 2019), industrial production (Malik, Masood, and Brem 2023), and healthcare (Mukherjee et al. 2022). Recent advances in Reinforcement Learning (RL)-based motion control enable robots to autonomously learn optimal policies through simulated environmental interactions, achieving complex motion skills (Long et al. 2024; Chen et al. 2024; van Marum et al. 2024; Ji et al. 2024; Radosavovic et al. 2024). However, due to the differences between the simulation environment and the real world (Zhang et al. 2024), including environmental variations (Fujimoto et al. 2024), sensor noise (Liang et al. 2022), and external disturbances (Peng et al. 2018), neural controllers encounter domain distribution shift (Fujimoto et al. 2024), leading to the sim-to-real transfer problem of motion policies (Moos et al. 2022). In addition, the inherent sensitivity of RL-based neural controllers to perturbations (Shi et al. 2025) and the lack of systematic robustness design (Barbara, Wang, and Manchester 2024) further exacerbate the instability of deploying these policies in real-world applications (Kobayashi 2022), resulting in the inability of humanoid robots to complete long horizon motion tasks in complex environments.

To enhance the robustness of motion policies, previous works employ Domain Randomization (DR) to introduce perturbations (Wei et al. 2023) such as observation noise and environmental variations during training to simulate the distribution gap between training and deployment, improving the robustness of policies to environmental uncertainty (Gu, Wang, and Chen 2024). Other works utilize reg-

*Corresponding author.

ularization constraints to indirectly improve the robustness of motion policies (Chen et al. 2024; Yan et al. 2024). Lipschitz regularization limits the output variation of the policy under small input perturbations, improving action smoothness and robustness to small perturbations (Barbara, Wang, and Manchester 2024). Symmetry regularization leverages the structural symmetry of the robot to constrain the exploration space of the policy, guiding the learning of the optimal policy with symmetry, and improves the coordination and robustness of the robot's motion (Su et al. 2024). To address the dynamic mismatch between simulation and reality, recent studies integrate residual action models trained on real-world deployment data (He et al. 2025). These models can compensate for unmodeled dynamic variations, improve simulation fidelity, and further fine-tune motion policies to improve sim-to-real transfer performance.

Although these methods have achieved remarkable robustness, DR cannot specifically perturb policy vulnerabilities (Tang, Tan, and Harada 2020; Long et al. 2024). Regularization constraints require a trade-off between policy exploration and robust constraints (Shi et al. 2024). Residual models are dependent on real-world data, resulting in low implementation efficiency (Salvato et al. 2021). Furthermore, the approach of improving robustness by accurately identifying vulnerabilities in the policy network and applying targeted perturbations remains underexplored (Chen et al. 2024), especially in humanoid robots, where high-dimensional observations and degrees of freedom complicate motion policies.

In this work, we propose a novel **S**elective **A**dversarial **A**ttack for **R**obust **T**raining (**SA2RT**) to enhance humanoid motion policy robustness. A learnable adversary network identifies policy vulnerabilities and generates targeted perturbations to destabilize the robot with a minimal attack budget. Through alternating optimization of attack and motion policies, our method continually strengthens the motion policy against the strongest discovered attacks. We conduct extensive experiments on the Unitree G1 humanoid robot (Unitree G1 2025) on various challenging terrains and agile trajectory tracking. The results demonstrate that the SA2RT significantly improves the motion performance of humanoid robots in real-world environments.

In summary, the contributions of this work are as follows:

- A novel selective adversarial attack for robust training is proposed for humanoid robots. By introducing adversarial attack policies to identify vulnerability of motion policies, effective adversarial samples are generated to enhance the robustness of motion policies.
- The Selective Attack Policy (SAP) constrained by an attack budget, enhances the vulnerability mitigation of motion policy without inducing conservative degradation.
- Extensive experimental results on real robots demonstrate that our method significantly improves the long horizon mobility and tracking performance.

Related Works

Learning-Based Humanoid Motion Control

In recent years, RL-based approaches have achieved impressive results in humanoid robot motion control (Ren et al. 2025; Wei et al. 2023; Radosavovic et al. 2024; Cui et al. 2024; Gu et al. 2024), especially in complex terrain traversal (Zhuang, Yao, and Zhao 2024; Long et al. 2024) and whole-body trajectory imitation (Cheng et al. 2024; He et al. 2024a). Radosavovic et al. (Radosavovic et al. 2024) use causal transformers trained via autoregressive modeling of observations and actions on diverse motion datasets, achieving stable long-distance walking in outdoor environments. Sun et al. (Sun et al. 2025) extract terrain features around the robot based on visual perception information to enhance the robot's locomotion capabilities in challenging unstructured terrains (Zhuang, Yao, and Zhao 2024). Furthermore, by remapping human motion trajectories from teleoperation and motion capture (MoCap) to humanoid robots (Ji et al. 2024; He et al. 2024b), whole-body imitation learning frameworks have been developed to replicate expressive human-like movements (Ze et al. 2025). He et al. (He et al. 2025) compensate for dynamic mismatches by collecting data in real robots to train a residual action model and incorporate it into simulation to fine-tune the policy to align with real-world dynamics. However, the inherent sensitivity of neural control policies to sensor/actuator noise significantly impairs the stability of humanoid robots during high-dynamic motions (Salvato et al. 2021), failing to maintain long horizon robust mobility and tracking performance.

Robust Reinforcement Learning

Robust adversarial training introduces adversarial attacks during training (Moos et al. 2022), allowing the agent to learn under uncertain perturbations and improve its robustness in real environments (Huang et al. 2017). Adversarial attacks can impose perturbations in the observation space (Zhang et al. 2021), action space (Tessler, Efroni, and Mannor 2019), reward space (Wang, Liu, and Li 2020), and environmental state space (Kuang et al. 2022), effectively simulating various uncertainties that the task policy may encounter in deployment scenarios. Shi et al. (Shi et al. 2024) use adversarial attacks to evaluate the robustness of RL-based locomotion policies for quadrupedal robots. (Shi et al. 2025) decouples the upper-body and lower-body control of humanoid robots as two independent agents, proposing an adversarial motion and imitation learning framework. On the other hand, considering the high sensitivity of neural networks to small input perturbations (Zhang, Nie, and Gao 2024), some studies have investigated the use of Lipschitz-constrained policy networks (Mysore et al. 2021) to improve robustness against input noise and small perturbations (Kobayashi 2022). By regularizing or employing specialized network architectures to constrain the Lipschitz constant of neural networks (Cho and Kim 2024), the output actions of the policy can be effectively smoothed (Barbara, Wang, and Manchester 2024), thereby improving the robustness of the policy to perturbations (Chen et al. 2024).

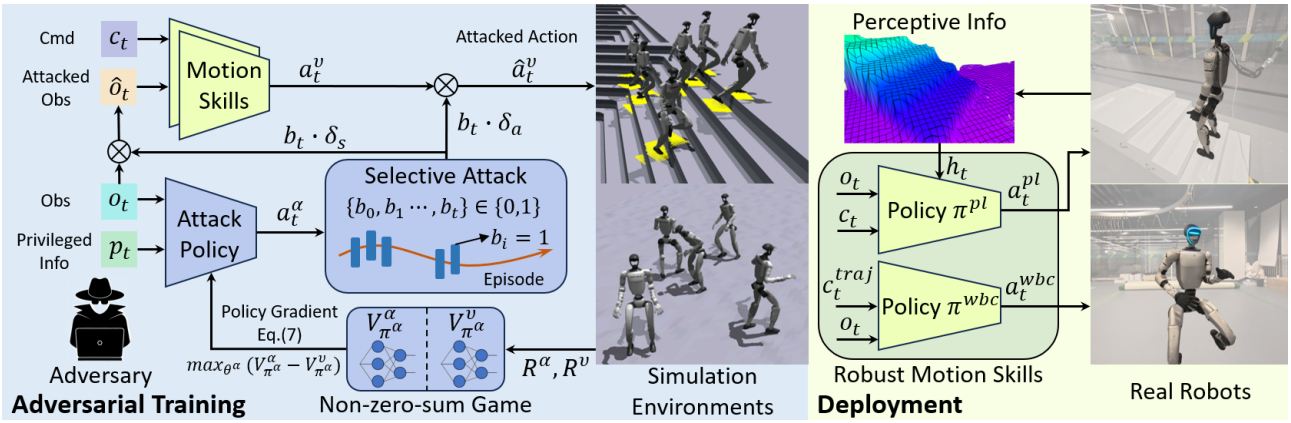


Figure 2: Overview of the SA2RT. The SAP identifies vulnerabilities in motion states and generates adversarial samples by applying perturbations in both state spaces and action spaces. Through alternating adversarial training under non-zero-sum game, the motion policy continuously addresses its own vulnerabilities using adversarial samples, enhancing its robustness against perturbations. During deployment, the robust motion skills are deployed to real robots without requiring the SAP, enabling robust whole-body motion control for humanoid robots.

Preliminaries

Reinforcement Learning

The RL-based control problem is typically formulated as a Markov Decision Process (MDP), where an agent π dynamically interacts with the environment through trial-and-error mechanisms to maximize the task-specific objective function. The MDP is defined as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the transition probability, $R(s, a)$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor. At each timestep t, the agent π receives an observation s_t and chooses an action $a_t \in \pi(s_t)$ to obtain the reward $R(s_t, a_t)$. The agent's goal is to maximize the expected cumulative return:

$$J(\pi) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right]. \tag{1}$$

Two-Player Markov Games

A two-player Markov game (Guo et al. 2021) can be formally represented as $(\mathcal{N}, \mathcal{S}, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, \mathcal{T}, \{R^i\}_{i\in\mathcal{N}}, \gamma)$, where $\mathcal{N} = \{\pi^\alpha, \pi^\upsilon\}$ represents the set of agents with adversary π^α and victim π^υ. $\mathcal{S}$ represents the state space of the global environment observations. $\mathcal{A}^i$ represents the action space of the agent i. $\mathcal{T} : \mathcal{S} \times \mathcal{A}^\alpha \times \mathcal{A}^\upsilon \to \mathbb{R}$ represents the transition probability for any joint actions of both agents. $R^i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function for the agent i. γ is the discount factor. The state-value function for each agent is defined as a function of the joint policy $\pi(a|s) = \prod_{\pi^i \in \mathcal{N}} \pi^i(a^i|s)$

$$V^i_\pi(s) = \mathbb{E}_{a\sim\pi(a|s)}\left[R^i(s, a) + \gamma \mathbb{E}_{s'\sim\mathcal{T}}[V^i_\pi(s')]\right]. \tag{2}$$

In this Markov game, the victim π^υ improves its defense capability by maximizing its state value function $V^\upsilon_\pi(s)$, while the adversary π^α enhances its attack effect by maximizing its state value function V^α_π, thereby enhancing the adversarial robustness.

Methods

In this section, we present the SA2RT in detail, which aims to enhance the robustness of humanoid motion policies via selective adversarial attacks. As shown in Fig. 2, the framework integrates two learnable modules: motion policy and attack policy. The motion policy optimizes task performance under perturbations from the attack policy, improving perturbation resistance. Conversely, the attack policy identifies vulnerable states and selectively applies adversarial perturbations to induce falls. The two modules are alternately optimized in a game-theoretic manner.

Selective Adversarial Attacks

When deploying RL-based neural controllers, observations are susceptible to natural noise and outliers, causing state estimation errors that lead to suboptimal actions. Additionally, actuator latency and mechanical wear introduce execution deviations, further degrading motion performance. To holistically assess policy vulnerabilities from these observation-actuator error cascades, we define the adversarial attack policy's action space as the joint space of the motion policy's state and action spaces. Let $\mathcal{B}_s = \{s+\delta_s | \delta_s \in \mathcal{D}_s\}$ and $\mathcal{B}_a = \{a+\delta_a | \delta_a \in \mathcal{D}_a\}$ be bounded perturbation sets for states and actions, respectively, where $\|\delta_s\| \le \epsilon_s$, $\|\delta_a\| \le \epsilon_a$. Given a pre-trained motion policy π^m on MDP $\mathcal{M} = (S, A, P, R, \gamma)$, we construct an adversarial MDP $\hat{\mathcal{M}} = (\mathcal{S}, \hat{\mathcal{A}}, \hat{\mathcal{P}}, \hat{\mathcal{R}}, \gamma)$ where $\hat{A} = \mathcal{S} \times \mathcal{A}$ represents the joint state-action attack space, $\hat{\mathcal{P}}(s'|s, \hat{a}) = \mathbb{E}_{a\sim\pi^m(\cdot|s+\delta_s)}[\mathcal{P}(s'|s, a + \delta_a)]$, and $\hat{R}(s, a)$ encodes attack objectives. The Persistent Attack Policy (PAP) in $\hat{\mathcal{M}}$ maximizes attack efficacy while adhering to perturbation bounds $\mathcal{B}_s$, $\mathcal{B}_a$. The temporal homogeneity of the PAP prevents it from distinguishing state-specific vulnerability differences and renders it easily detectable, leading to excessive conservatism in the victim policy.

To further improve the stealth and efficiency of attacks, we propose a Selective Attack Policy (SAP). By introducing

an attack-budget constraint, SAP not only determines how to generate optimal perturbations but also identifies vulnerable states where attacks are most effective. This approach aims to significantly degrade the performance of the motion policy using the fewest attack steps, thereby maximizing impact while minimizing detectability. In an episode, the motion policy produces a state-action sequence $\{s_t, a_t\}_{t=0}^{T}$. The SAP strategically selects a subset of timesteps $\mathcal{T}_{adv} \subset \{1, \cdots, T\}$ (where $|\mathcal{T}_{adv}| \ll T$) to perturb the robot, rather than attacking all timesteps. Let $\{\delta_1, \cdots, \delta_T\}$ denote a sequence of perturbations sampled from an arbitrary attack policy π^{adv}, and N_a be the attack budget. The SAP can be formulated as the following optimization problem:

$$\begin{aligned} \max_{b_0,\cdots,b_T,\, \pi^{adv}} \quad & \mathbb{E}[\sum_{t=0}^{T} \hat{R}(s_t, a_t)] \\ s.t. \quad & [\delta_s, \delta_a] \sim \pi^{adv}(s_t), \\ & a_t \sim \pi(s_t + b_t\delta_{s,t}) + b_t\delta_{a,t}, \\ & s_{t+1} \sim \mathcal{P}(s_t, a_t), \\ & \sum_{t=0}^{T} b_t \le N_a. \end{aligned} \tag{3}$$

The binary variable b_t indicates whether the perturbation $\pi^{adv}(s_t)$ is added to the state s_t, and step t is a critical step found by the SAP if $b_t = 1$. To address the attack-budget constraint in problem (3), we introduce a Lagrange multiplier λ to transform the hard budget constraint into a soft penalty term:

$$\max_{b_0,\cdots,b_T,\, \pi^{adv}} \quad \mathbb{E}\left[\sum_{t=0}^{T} \hat{R}(s_t, a_t)\right] - \lambda \sum_{t=0}^{T} b_t. \tag{4}$$

The λ is a hyperparameter that controls the penalty for each attack. A higher penalty parameter λ corresponds to a lower budget constraint N_a. We use RL to train SAP in another MDP $\hat{\mathcal{M}}' = (\mathcal{S}, \hat{\mathcal{A}}', \hat{\mathcal{P}}', \hat{R}', \gamma')$, where $\hat{\mathcal{A}}' = \{0, 1\} \times \hat{A}$, $\hat{\mathcal{P}}'(s'|s, \hat{a}') = \mathbb{E}_{a\sim\pi^m(\cdot|s+b\delta_s)}[\mathcal{P}(s'|s, a + b\delta_a)]$, $\hat{R}' = \hat{R} - \lambda b$, and $\gamma' = 1$.

Non-Zero-Sum Adversarial Training

When the reward function $R^\alpha = -R^\upsilon$, adversarial training establishes a zero-sum game between the victim π^υ and the adversary π^α, where both policy updates follow an iterative competitive process to maximize and minimize the shared value function, respectively. As demonstrated by Perolat et al. (Perolat et al. 2015), the optimal equilibrium value function ensures the equivalence between the minimax equilibrium and the Nash equilibrium in such games:

$$V_\pi^{\alpha*}(s) = \min_{\pi^\upsilon} \max_{\pi^\alpha} V_\pi^\alpha(s) = \max_{\pi^\alpha} \min_{\pi^\upsilon} V_\pi^\alpha(s). \tag{5}$$

However, previous analyses of the minimax formulation in adversarial training shows that the attacker learned under the zero-sum game framework does not guarantee continuous improvement in attack performance. This results in a weak adversary dilemma, which in turn leads to a robust overfitting of the victim.

Algorithm 1: Alternating Adversarial Training

Input: The motion policy π^m with pre-trained parameters θ_m, the attack policy π^{adv} with random parameters θ_{adv}, environment $\mathcal{E}$.

for i = 1 to N_{iter} **do**
  for j = 1 to N_{adv} **do**
    $\{(s_t, a_t^m, a_t^{adv}, R_t^m, R_t^{adv})\} \leftarrow$ rollout($\mathcal{E}$, π^m, π^{adv})
    Update θ_{adv} with $\mathcal{D}_{adv} := \{(s_t, a_t^{adv}, R_t^m, R_t^{adv})\}$
  end for
  for j = 1 to N_m **do**
    $\{(s_t, a_t^m, a_t^{adv}, R_t^m, R_t^{adv})\} \leftarrow$ rollout($\mathcal{E}$, π^m, π^{adv})
    Update θ_m with $\mathcal{D}_m := \{(s_t, a_t^m, R_t^m)\}$
  end for
end for

In this work, we formulate the adversarial training between the motion policy π^m and the attack policy π^{adv} as a non-zero-sum game. The attack policy maximizes the attack reward R^{adv} associated with the robot's safety failure while negatively affecting the motion policy, which can be formally expressed as:

$$J(\theta_{adv}) = \underset{\theta_{adv}}{maximize}\ (V_\pi^{adv}(s) - V_\pi^m(s)), \tag{6}$$

where θ_{adv} is the parameter of the attack policy network, $V_\pi^{adv}(s)$ and $V_\pi^m(s)$ represent the attacker's value function and the motion agent's value function, respectively. A trivial solution to solving Eq. (6) involves applying policy gradient methods. However, this solution does not guarantee monotonic changes in both value functions. Taking advantage of the work (Guo et al. 2021), a new optimization objective that can guarantee the monotonicity of Eq. (6) is obtained by approximating the value function:

$$\begin{aligned} & \underset{\theta_{adv}}{argmax}\, \mathbb{E}_{(a_t^{adv}, s_t)\sim\pi_k^{adv}}[min(clip(\rho_t, 1-\epsilon, 1+\epsilon)A_t^{adv}, \\ & \rho_t A_t^{adv}) - min(clip(\rho_t, 1-\epsilon, 1+\epsilon)A_t^m, \rho_t A_t^m)], \\ & \rho_t = \frac{\pi^{adv}(a_t^{adv}|s_t)}{\pi_k^{adv}(a_t^{adv}|s_t)}, A_t^{adv} = A_{\pi_k^{adv}}^{adv}(a_t^{adv}, s_t), \\ & A_t^m = A_{\pi_k^{adv}}^m(a_t^{adv}, s_t), \end{aligned} \tag{7}$$

where π_k^{adv} denotes the old attack policy and ϵ represents the clipping parameter in the PPO algorithm (Schulman et al. 2017). Subsequently, we employ temporal difference learning with two separate neural networks to approximate the value functions of the attack policy and the motion policy. The policy gradient method is then applied to optimize Eq. (7), learning a powerful attack policy.

We employ an alternating training procedure to optimize the motion policy and the attack policy. Algorithm 1 outlines our approach in detail. The motion policy is initialized from the pre-trained policy, while the attack policy is randomly initialized. In N_{adv} iterations, the motion policy is fixed and embedded into the environment, and the attack policy is optimized to find the intrinsic weaknesses of the motion policy and impose attacks. Then in N_m iterations, the attack

policy is fixed, enabling the motion policy to be further optimized for resilience against adversarial attacks. Through dynamic alternating optimization over N_{iter} iterations, our framework achieves online robust adversarial training.

Experiments

Experimental Setup

We apply the SA2RT to two humanoid robot tasks, Perceptive Locomotion (PL) and Whole-Body Control (WBC), and implementation details are provided in the Appendix. Through comparative experiments and ablation studies, we address the following research questions: (1) Can adversarial attacks induce stronger perturbations than domain randomization? (2) Can SA2RT improve the robustness of motion policies? (3) Does SA2RT contribute to the motion performance of real robots?

Baselines We design four different training settings for comparative analysis, and each motion policy is represented using the following notation:

- **PL/WBC-CE**: The motion policies trained in Clean Environments (CE), without DR and adversarial attacks. Specifically, PL-CE represents the baseline of the perceptive locomotion policy, while WBC-CE represents the baseline of the whole-body control policy.
- **PL/WBC-DR**: Applying DR in training environments.
- **PL/WBC-PAP**: Robust adversarial training with persistent attack policies.
- **PL/WBC-SAP**: Robust adversarial training with selective attack policies.

Metrics To evaluate the performance of motion policies, we use several metrics:

- The mean linear velocity error $E_{vel}(m/s)$ and the mean angular velocity error $E_{ang}(rad/s)$ are used to evaluate the velocity tracking accuracy.
- The gravitational projection component $E_g(m/s^2)$ is used to evaluate the dynamic stability of the robot.
- The Mean Per Keybody Position Error (MPKPE) $E^{upper}_{mpkpe}(m)$ and $E^{lower}_{mpkpe}(m)$ are used to evaluate the keypoint tracking accuracy.
- The Mean Per Joint Position Error (MPJPE) $E^{upper}_{mpjpe}(rad)$ and $E^{lower}_{mpjpe}(rad)$ are used to evaluate the joint position tracking ability.
- The success rate $R_{sr}(\%)$ for robot survival in complex terrain traversal and whole-body trajectory tracking.

Simulation Results

Effectiveness of Adversarial Perturbations To ensure the comparability of experimental results, we maintain consistent perturbation spaces and thresholds for both DR and SAP during training. We then evaluate the task success rates of motion policies under different training settings by separately applying DR and SAP. As shown in Table 1, DR sharply reduces PL/WBC-CE's success rate yet leaves adversarially trained PL/WBC-SAP unaffected. Conversely, the SAP compromises both PL/WBC-CE and PL/WBC-DR. The results show that simple DR cannot ensure the robustness of the motion policy, while SAP effectively exposes vulnerabilities and introduces stronger perturbations.

Envs	Motion Policies $R_{sr}(\%)$		
	PL/WBC-CE	PL/WBC-DR	PL/WBC-SAP
DR	77.3 ± 7.3	94.4 ± 5.7	97.4 ± 2.1
SAP	**0.0 ± 0.0**	**0.0 ± 0.0**	**95.7 ± 6.4**

Table 1: Success Rates in DR and SAP Environments.

Envs	Policies	Metrics		
		$E_{vel}(m/s)$	$E_{ang}(rad/s)$	$E_g(m/s^2)$
CE	PL-DR	0.25 ± 0.03	0.29 ± 0.02	0.34 ± 0.01
	PL-SAP	**0.19 ± 0.02**	**0.22 ± 0.03**	**0.30 ± 0.01**
DR	PL-DR	0.28 ± 0.02	0.34 ± 0.03	0.42 ± 0.01
	PL-SAP	**0.21 ± 0.03**	**0.27 ± 0.02**	**0.36 ± 0.01**

Table 2: Robustness Analysis of PL Policies.

Robustness of Motion Policies We evaluate motion policies trained with different methods in both clean and DR environments. For the perceptive locomotion task, we evaluate velocity tracking and gravitational projection component on flat ground, slopes, stairs, and discrete terrains. As shown in Table 2, PL-SAP outperforms PL-DR, indicating that that the SA2RT can effectively address vulnerabilities in motion policies, enhancing robustness and task performance. For whole-body control, we evaluate trajectory tracking performance, as shown in Fig. 3. WBC-SAP outperforms WBC-DR across all evaluation metrics, particularly in velocity tracking. Experimental results demonstrate that the SA2RT effectively enhances the robot's trajectory imitation performance, enabling it to maintain body stability in perturbed environments without compromising tracking accuracy.

Ablation Study To evaluate the impact of SAP on the performance of motion policies, we compare the performance of motion policies adversarially trained by PAP and SAP under different perturbation levels $L_p \in \{1, 2, 3, 4\}$. For each level, we train PL-PAP-L_p and PL-SAP-L_p policies and record their normalized rewards, as shown in Fig. 4(a). The results demonstrate that adversarial training outperforms DR in robustness. As the perturbation level L_p increases, normalized rewards of PL-SAP-L_p decrease less than those of PL-PAP-L_p, indicating that SAP's selective attacks on vulnerable states prevent the excessive conservatism from invalid stable-state attacks. In addition, we compare the performance of adversarially trained motion policies under different perturbation levels L_P, as shown in Fig. 4(b). PL-SAP-L_p exhibits slower performance degradation compared to PL-PAP-L_p, while the degradation of PL-PAP-L_p exceeds that of PL-DR. The results indicate that while PAP exhibits stronger attack capabilities, its full-time attack induces a significant shift in the state distribution of the motion policy, leading to a marked decline in the policy's performance

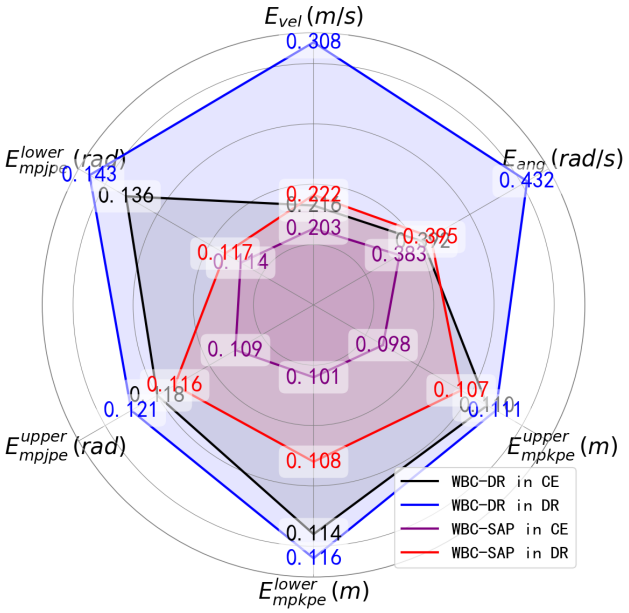
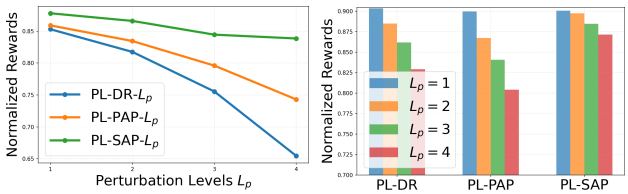


Figure 3: Performance analysis of whole-body control. Trajectory tracking errors of WBC-DR and WBC-SAP are evaluated in clean environments and DR environments. WBC-SAP outperforms WBC-DR across all evaluation metrics, demonstrating that the SA2RT effectively enhances the robustness and tracking performance of WBC policies.



(a) (b)

Figure 4: Impact of different attack policies on motion policy performance. (a) Rewards of motion policies learned via different attack policies under varying perturbation levels L_p. (b) Performance comparison in unperturbed environments for motion policies trained under different L_p.

in low-perturbation environments. In contrast, SAP only attacks the vulnerable states of the motion policy, featuring certain high efficiency and stealthiness, and effectively balances the motion policy's performance and robustness.

Analysis of Selective Attack Policy We evaluate SAP's attack ratio across different motion tasks (Fig. 5), testing perceptive locomotion on flat ground, slopes, stairs, and discrete terrain, while categorizing whole-body control into simple walking and dynamic dancing trajectories. The results demonstrate that the attack ratio increases with task difficulty, confirming SAP's capability to learn and execute targeted attacks. In addition, we analyze SAP's attack behavior during the robot's transition from flat terrain to stairs (Fig. 6). The SAP selectively attacks vulnerable transition

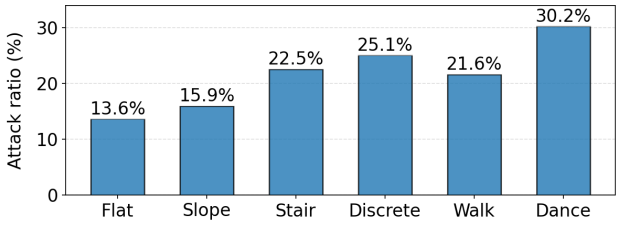
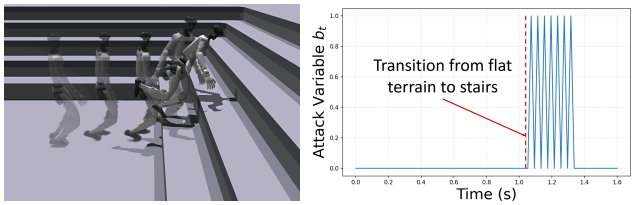


Figure 5: SAP's attack ratio varies significantly across motion tasks. As the difficulty of the task increases, the attack rate gradually increases.



(a) (b)

Figure 6: SAP attack sequences during the robot's transition from flat terrain to stairs. (a) Robot motion states visualization. (b) Attack variable sequence. As the robot readies to traverse stairs, SAP identifies the state's vulnerability, applies a few attacks, and successfully induces a fall.

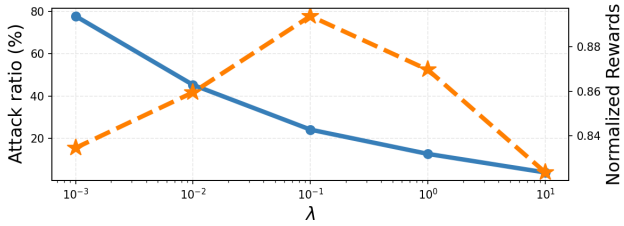


Figure 7: Comparison of WBC-SAP's attack ratio and performance across parameter λ. The blue line represents the attack ratio and the orange line represents the reward. As λ increases, the attack ratio decreases, but the performance shows a trend of first increasing and then decreasing.

phases, thereby inducing falls. This indicates that SAP can clearly distinguish differences in vulnerability across the robot's various motion states and apply targeted perturbations, prompting the motion policy to continuously compensate for vulnerable states during adversarial training and enhancing its robustness.

Analyzing Hyperparameter λ In selective adversarial training, hyperparameter λ regulates the number of adversarial attacks. We analyze WBC-SAP's attack ratio and motion performance under different λ. As illustrated in Fig. 7, as λ increases, the attack ratio drops significantly, while motion performance first rises then falls. When λ increases from 10^{-3} to 10^{-1}, SAP reduces attacks while identifying robot motion vulnerabilities, improving attack efficiency and policy robustness. However, further increasing λ from 10^{-1} to

Policies	Terrains Traversal R_{sr} (%)				
	Slopes	Stairs	Gravel	Sands	Grass
PL-DR	54 ± 6.9	47 ± 7.5	71 ± 4.3	73 ± 4.5	82 ± 4.9
PL-SAP	$\mathbf{90 \pm 3.7}$	$\mathbf{84 \pm 5.4}$	$\mathbf{93 \pm 2.1}$	$\mathbf{90 \pm 2.0}$	$\mathbf{95 \pm 1.6}$

Table 3: Success Rates of Complex Terrain Traversal.

10^1, leads to excessive penalties on the number of attacks, thereby severely reducing the attack ratio. This weakens targeted attacks and hinders improvements in robustness. Thus, λ can adjust the number of attacks, constraining the SAP to learn targeted attacks and avoiding inefficient persistent attacks that compromise task performance.

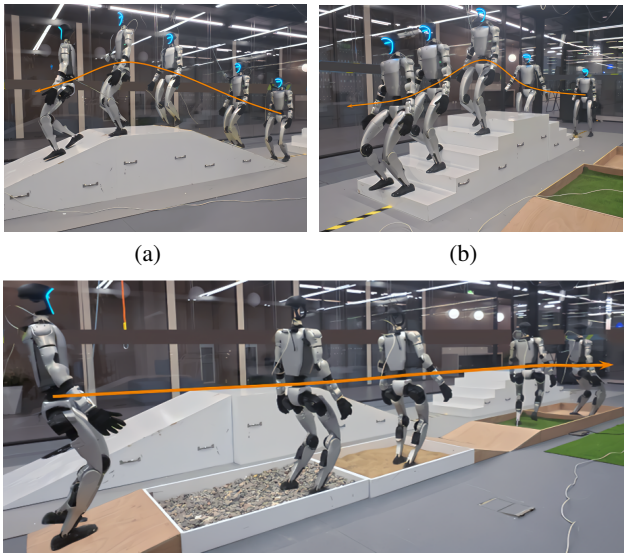
Real-World Experiments

Complex Terrain Traversability To evaluate the terrain traversability of PL-SAP, we construct a standardized test environment with challenging terrains (stairs, slopes, gravel, sand, grass) and compare traversal success rates between PL-DR and PL-SAP, as illustrated in Table 3. Each terrain is tested in 5 groups, with 10 trials per group. Benefiting from effective vulnerability identification and targeted perturbation attacks, the SA2RT improves the traversal success rates of the real robot by 40%. Fig. 8 shows snapshots of complex terrain traversal. PL-SAP exhibit significantly fewer foot trips and slips than PL-DR on stairs and slopes. Additionally, PL-SAP effectively adapts to thin, sparse ridges in terrain connections (hard to perceive via elevation maps), avoiding trips and maintaining stability, whereas PL-DR often falls here. These results demonstrate that the SA2RT significantly enhances the robustness of motion policies.

Whole-Body Trajectory Imitation Performance We further evaluate the performance of the whole-body control on the real robot by tracking 120 seconds of dynamic dance motion. The two joint trajectory tracking indicators E^{lower}_{mpjpe} and E^{upper}_{mpjpe} are tested, as shown in Fig. 9. Due to the high dynamic balancing demands of the dance trajectory, WBC-DR exhibits progressively accumulating tracking errors, ultimately causing instability and premature termination after only 65 seconds of execution. In contrast, WBC-SAP effectively reduces trajectory tracking error by 32% and successfully completes the 120 seconds motion trajectory. Snapshots of the imitation of dance motions are presented in Fig. 1, demonstrating the robot's ability to dynamically regulate the balance of the whole body while performing complex trajectory tracking. These results indicate that the SA2RT significantly enhances the robustness of motion policies and effectively bridges the sim-to-real gap in dynamic motion control.

Conclusions

In this paper, we propose a novel selective adversarial attack for robust training that aims to enhance the robustness of RL-based motion skills and improve long horizon mobility and tracking performance in real robots. By introducing the selective attack policy, which is capable of identifying vulnerable states and imposing targeted perturbations under an attack-budget constraint, the motion policies are strengthened through dynamic non-zero-sum game optimization. Extensive experiments conducted on the Unitree G1 humanoid robot have verified the superiority of the SA2RT. Future work will explore potential perturbation attacks in the interaction between the robot and the environment to enhance the robot's adaptability to dynamic environments. In addition, optimizing the attack-budget constraint to further improve the trade-off between attack effectiveness and strategy agility remains a promising direction.



(a) (b)

(c)

Figure 8: Evaluating the motion performance of PL-SAP in traversing complex terrain. (a) The robot traverses 22° slopes. (b) The robot traverses 16 cm stairs. (c) The robot traverses gravel and sandy terrain. PL-SAP exhibits higher success rates in complex terrain traversal, indicating the SA2RT effectively enhances motion policy robustness.

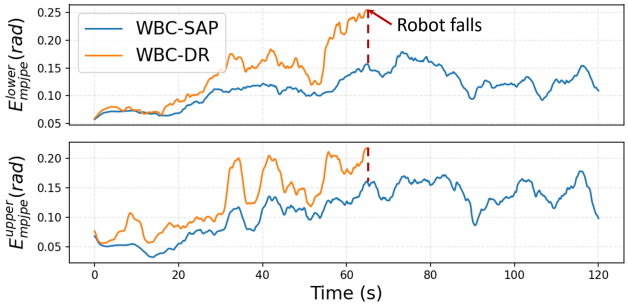


Figure 9: The Mean Per Joint Position Error of each frame in the dance trajectory imitation. Due to the insufficient robustness of PL-DR, the robot loses balance and terminates prematurely at 65 seconds. PL-SAP effectively reduces trajectory tracking errors and enables complete long horizon trajectory tracking.

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

Barbara, N. H.; Wang, R.; and Manchester, I. R. 2024. On Robust Reinforcement Learning with Lipschitz-Bounded Policy Networks. *arXiv preprint arXiv:2405.11432*.

Chen, Z.; He, X.; Wang, Y.-J.; Liao, Q.; Ze, Y.; Li, Z.; Sastry, S. S.; Wu, J.; Sreenath, K.; Gupta, S.; et al. 2024. Learning smooth humanoid locomotion through lipschitz-constrained policies. *arXiv preprint arXiv:2410.11825*.

Cheng, X.; Ji, Y.; Chen, J.; Yang, R.; Yang, G.; and Wang, X. 2024. Expressive whole-body control for humanoid robots. *arXiv preprint arXiv:2402.16796*.

Cho, N.; and Kim, Y. 2024. On the stability of Lipschitz continuous control problems and its application to reinforcement learning. *arXiv preprint arXiv:2404.13316*.

Cui, W.; Li, S.; Huang, H.; Qin, B.; Zhang, T.; Zheng, L.; Tang, Z.; Hu, C.; Yan, N.; Chen, J.; et al. 2024. Adapting humanoid locomotion over challenging terrain via two-phase training. In *8th Annual Conference on Robot Learning*.

Fujimoto, T.; Suetterlein, J.; Chatterjee, S.; and Ganguly, A. 2024. Assessing the Impact of Distribution Shift on Reinforcement Learning Performance. *arXiv preprint arXiv:2402.03590*.

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*.

Gu, X.; Wang, Y.-J.; Zhu, X.; Shi, C.; Guo, Y.; Liu, Y.; and Chen, J. 2024. Advancing humanoid locomotion: Mastering challenging terrains with denoising world model learning. *arXiv preprint arXiv:2408.14472*.

Gu, Z.; Li, J.; Shen, W.; Yu, W.; Xie, Z.; McCrory, S.; Cheng, X.; Shamsah, A.; Griffin, R.; Liu, C. K.; et al. 2025. Humanoid Locomotion and Manipulation: Current Progress and Challenges in Control, Planning, and Learning. *arXiv preprint arXiv:2501.02116*.

Guo, W.; Wu, X.; Huang, S.; and Xing, X. 2021. Adversarial policy learning in two-player competitive games. In *International conference on machine learning*, 3910–3919. PMLR.

He, T.; Gao, J.; Xiao, W.; Zhang, Y.; Wang, Z.; Wang, J.; Luo, Z.; He, G.; Sobanbab, N.; Pan, C.; et al. 2025. ASAP: Aligning Simulation and Real-World Physics for Learning Agile Humanoid Whole-Body Skills. *arXiv preprint arXiv:2502.01143*.

He, T.; Luo, Z.; Xiao, W.; Zhang, C.; Kitani, K.; Liu, C.; and Shi, G. 2024a. Learning human-to-humanoid real-time whole-body teleoperation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 8944–8951. IEEE.

He, T.; Xiao, W.; Lin, T.; Luo, Z.; Xu, Z.; Jiang, Z.; Kautz, J.; Liu, C.; Shi, G.; Wang, X.; et al. 2024b. Hover: Versatile neural whole-body controller for humanoid robots. *arXiv preprint arXiv:2410.21229*.

Huang, S.; Papernot, N.; Goodfellow, I.; Duan, Y.; and Abbeel, P. 2017. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*.

Ji, M.; Peng, X.; Liu, F.; Li, J.; Yang, G.; Cheng, X.; and Wang, X. 2024. Exbody2: Advanced expressive humanoid whole-body control. *arXiv preprint arXiv:2412.13196*.

Kobayashi, T. 2022. L2c2: Locally lipschitz continuous constraint towards stable and smooth reinforcement learning. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4032–4039. IEEE.

Kuang, Y.; Lu, M.; Wang, J.; Zhou, Q.; Li, B.; and Li, H. 2022. Learning robust policy against disturbance in transition dynamics via state-conservative policy optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 7247–7254.

Liang, Y.; Sun, Y.; Zheng, R.; and Huang, F. 2022. Efficient adversarial training without attacking: Worst-case-aware robust reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 22547–22561.

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*.

Malik, A. A.; Masood, T.; and Brem, A. 2023. Intelligent humanoids in manufacturing to address worker shortage and skill gaps: Case of Tesla Optimus. *arXiv preprint arXiv:2304.04949*.

Mende, M.; Scott, M. L.; Van Doorn, J.; Grewal, D.; and Shanks, I. 2019. Service robots rising: How humanoid robots influence service experiences and elicit compensatory consumer responses. *Journal of Marketing Research*, 56(4): 535–556.

Moos, J.; Hansel, K.; Abdulsamad, H.; Stark, S.; Clever, D.; and Peters, J. 2022. Robust reinforcement learning: A review of foundations and recent advances. *Machine Learning and Knowledge Extraction*, 4(1): 276–315.

Mukherjee, S.; Baral, M. M.; Pal, S. K.; Chittipaka, V.; Roy, R.; and Alam, K. 2022. Humanoid robot in healthcare: a systematic review and future research directions. In *2022 International conference on machine learning, big data, cloud and parallel computing (COM-IT-CON)*, volume 1, 822–826. IEEE.

Mysore, S.; Mabsout, B.; Mancuso, R.; and Saenko, K. 2021. Regularizing action policies for smooth control with reinforcement learning. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 1810–1816. IEEE.

Peng, X. B.; Andrychowicz, M.; Zaremba, W.; and Abbeel, P. 2018. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, 3803–3810. IEEE.

Perolat, J.; Scherrer, B.; Piot, B.; and Pietquin, O. 2015. Approximate dynamic programming for two-player zero-sum

Markov games. In *International Conference on Machine Learning*, 1321–1329. PMLR.

Radosavovic, I.; Zhang, B.; Shi, B.; Rajasegaran, J.; Kamat, S.; Darrell, T.; Sreenath, K.; and Malik, J. 2024. Humanoid locomotion as next token prediction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Ren, J.; Huang, T.; Wang, H.; Wang, Z.; Ben, Q.; Pang, J.; and Luo, P. 2025. Vb-com: Learning vision-blind composite humanoid locomotion against deficient perception. *arXiv preprint arXiv:2502.14814*.

Salvato, E.; Fenu, G.; Medvet, E.; and Pellegrino, F. A. 2021. Crossing the reality gap: A survey on sim-to-real transferability of robot controllers in reinforcement learning. *IEEE Access*, 9: 153171–153187.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shi, F.; Zhang, C.; Miki, T.; Lee, J.; Hutter, M.; and Coros, S. 2024. Rethinking Robustness Assessment: Adversarial Attacks on Learning-based Quadrupedal Locomotion Controllers. *arXiv preprint arXiv:2405.12424*.

Shi, J.; Liu, X.; Wang, D.; Lu, O.; Schwertfeger, S.; Sun, F.; Bai, C.; and Li, X. 2025. Adversarial Locomotion and Motion Imitation for Humanoid Policy Learning. *arXiv preprint arXiv:2504.14305*.

Su, Z.; Huang, X.; Ordoñez-Apraez, 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.

Sun, W.; Cao, B.; Chen, L.; Su, Y.; Liu, Y.; Xie, Z.; and Liu, H. 2025. Learning Perceptive Humanoid Locomotion over Challenging Terrain. *arXiv preprint arXiv:2503.00692*.

Tang, Y.; Tan, J.; and Harada, T. 2020. Learning agile locomotion via adversarial training. In *2020 IEEE/RSJ International Conference On Intelligent Robots And Systems (IROS)*, 6098–6105. IEEE.

Tessler, C.; Efroni, Y.; and Mannor, S. 2019. Action robust reinforcement learning and applications in continuous control. In *International Conference on Machine Learning*, 6215–6224. PMLR.

Tong, Y.; Liu, H.; and Zhang, Z. 2024. Advancements in humanoid robots: A comprehensive review and future prospects. *IEEE/CAA Journal of Automatica Sinica*, 11(2): 301–328.

Unitree G1. 2025. https://www.unitree.com/cn/g1. [Online; accessed 25-June-2025].

van Marum, B.; Shrestha, A.; Duan, H.; Dugar, P.; Dao, J.; and Fern, A. 2024. Revisiting Reward Design and Evaluation for Robust Humanoid Standing and Walking. *arXiv preprint arXiv:2404.19173*.

Wang, J.; Liu, Y.; and Li, B. 2020. Reinforcement learning with perturbed rewards. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 6202–6209.

Wei, W.; Wang, Z.; Xie, A.; Wu, J.; Xiong, R.; and Zhu, Q. 2023. Learning Gait-conditioned Bipedal Locomotion with Motor Adaptation. In *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, 1–7. IEEE.

Yan, S.; Zhang, B.; Zhang, Y.; Boedecker, J.; and Burgard, W. 2024. Learning continuous control with geometric regularity from robot intrinsic symmetry. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 49–55. IEEE.

Ze, Y.; Chen, Z.; AraÃšjo, J. P.; Cao, Z.-a.; Peng, X. B.; Wu, J.; and Liu, C. K. 2025. TWIST: Teleoperated Whole-Body Imitation System. *arXiv preprint arXiv:2505.02833*.

Zhang, H.; Chen, H.; Boning, D.; and Hsieh, C.-J. 2021. Robust reinforcement learning on state observations with learned optimal adversary. *arXiv preprint arXiv:2101.08452*.

Zhang, Q.; Cui, P.; Yan, D.; Sun, J.; Duan, Y.; Han, G.; Zhao, W.; Zhang, W.; Guo, Y.; Zhang, A.; et al. 2024. Whole-body humanoid robot locomotion with human reference. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 11225–11231. IEEE.

Zhang, Y.; Nie, B.; and Gao, Y. 2024. Robust Locomotion Policy with Adaptive Lipschitz Constraint for Legged Robots. *IEEE Robotics and Automation Letters*.

Zhuang, Z.; Yao, S.; and Zhao, H. 2024. Humanoid Parkour Learning. *arXiv preprint arXiv:2406.10759*.