

Humanoid Whole-Body Locomotion on Narrow Terrain via Dynamic Balance and Reinforcement Learning

Weiji Xie^{1,2}, Chenjia Bai^{2*}, Jiyuan Shi², Junkai Yang^{1,2}, Yunfei Ge², Weinan Zhang¹, Xuelong Li²

Abstract—Humans possess delicate dynamic balance mechanisms that enable them to maintain stability across diverse terrains and under extreme conditions. However, despite significant advances recently, existing locomotion algorithms for humanoid robots still struggle to traverse extreme environments, especially in cases that lack external perception (e.g., vision or LiDAR). This is because current methods often rely on gait-based or perception-condition rewards, lacking effective mechanisms to handle unobservable obstacles and sudden balance loss. To address this challenge, we propose a novel whole-body locomotion algorithm based on dynamic balance and Reinforcement Learning (RL) that enables humanoid robots to traverse extreme terrains, particularly narrow pathways and unexpected obstacles, using only proprioception. Specifically, we introduce a dynamic balance mechanism by leveraging an extended measure of Zero-Moment Point (ZMP)-driven rewards and task-driven rewards in a whole-body actor-critic framework, aiming to achieve coordinated actions of the upper and lower limbs for robust locomotion. Experiments conducted on a full-sized Unitree H1-2 robot verify the ability of our method to maintain balance on extremely narrow terrains and under external disturbances, demonstrating its effectiveness in enhancing the robot’s adaptability to complex environments. The videos are given at <https://whole-body-loco.github.io>.

I. INTRODUCTION

Recent advances in humanoid locomotion control have achieved significant progress, benefited from large-scale interaction and policy learning [1] in a Reinforcement Learning (RL) framework. These methods mainly include phase-based gait learning [2], [3], motor skill control [4], [5], and motion imitation [6], [7]. By leveraging large-scale parallelized simulation [8] and policy optimization techniques [9], current humanoid locomotion methods demonstrate well adaptation capabilities in conventional terrains. Despite these achievements, the locomotion ability of humanoid robots still lags far behind that of humans in terms of dynamic balance under extreme conditions. Especially, humans can quickly adjust their foot placements and centroids when faced with situations such as slipping or stepping off the edge, thus regaining stability. In contrast, current RL-based controllers lack such abilities primarily due to their reliance on periodic gait [10], [3] or motion primitives [1], [6], which cannot achieve fast and diverse gait adjustment at critical moments of instability. We argue that a robust control policy should fully leverage the information on contact forces, the support polygon, and the centroid of the robot, which describes the fine-grained relationship between the robot and the support surface and is crucial for dynamic equilibrium.

Alternatively, classical biped locomotion research addresses this problem by considering two types of foot-ground contact during a walk cycle [11], [12]. Specifically, there is a double-support phase when the robot is supported on both feet, and a single-support phase when only one foot of the robot is in contact with the ground while the other is transitioning from the rear to the front position. In both cases, it is crucial to determine whether the contact can be maintained between the robot and the ground at a specific moment. Consequently, the concept of the Zero Moment Point (ZMP) is introduced to measure the influence of all forces acting on the robot that can be represented by a single force [13]. Specifically, the ZMP is defined as the point where the inertial and gravitational forces have no component along the horizontal axes. It has been demonstrated that if the ZMP lies within the support polygon of the foot and ground, the entire system is in dynamic balance. Inspired by this, we intergrate ZMP into learning-based humanoid whole body control, demonstrating significant improvements in the dynamic stability of humanoid robots when navigating complex terrains and resisting external disturbances.

In this paper, we propose a novel RL framework for whole-body locomotion in extreme scenarios, named *Dynamic Balanced Humanoid Locomotion (DBHL)*. To enhance the locomotion policy’s ability to traverse complex terrains, particularly narrow pathways and sudden obstacles, we extend the concept of ZMP to non-planar surfaces, thereby forming a line of ZMPs. We then design a reward function for the RL policy that encourages the ZMP coordinates to be close to the center of the humanoid’s support polygon. This reward function is calculated using privileged information obtained from simulations, while the policy is learned solely based on proprioception via an asymmetric actor-critic framework. This design allows the policy to be deployed in real-world scenarios without relying on external perception. Within the RL framework, we train a whole-body control policy that leverages upper-body swings to assist dynamic balance. We introduce angular-momentum regularization and action-space noise to constrain undesired body rotation and action range. Finally, we integrate the ZMP-based reward with command-following and regularization rewards using a reward vectorization technique, where each reward term and value function are estimated independently, avoiding inaccurate estimation of small items in an accumulated reward.

In experiments, we evaluate DBHL under various challenging conditions, including walking on surfaces with unknown disturbances, stepping on stairs with different widths and heights, and traversing narrow slopes with varying

¹Shanghai Jiao Tong University, China

²Institute of Artificial Intelligence (TeleAI), China Telecom

*Correspondence to: Chenjia Bai (baicj@chinatelecom.cn)

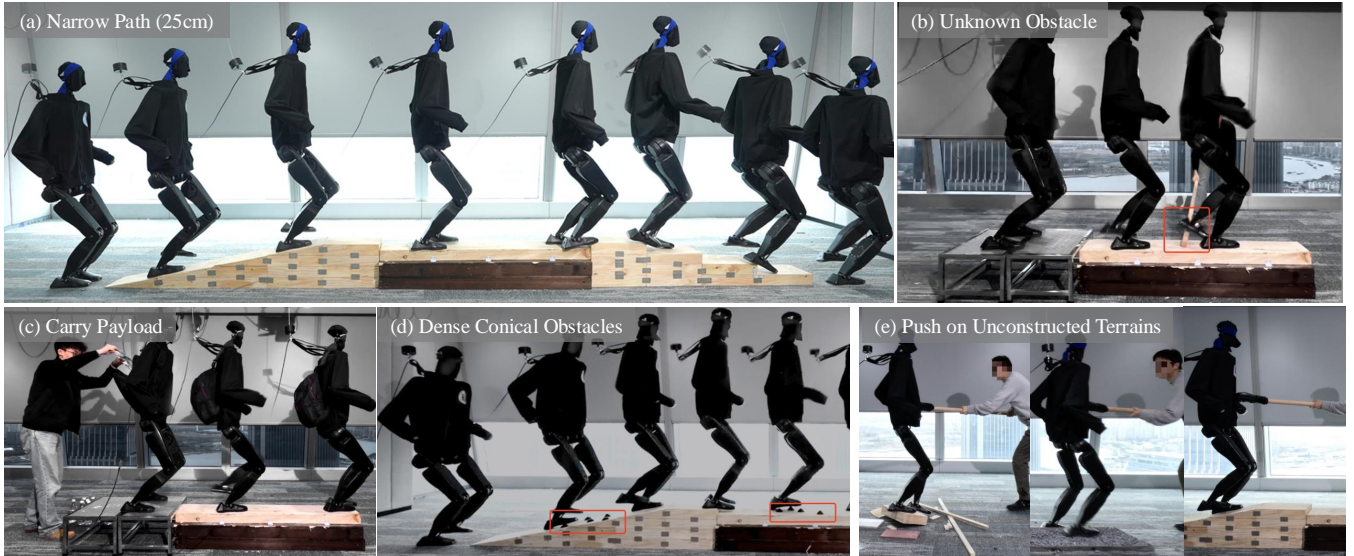


Fig. 1: **The locomotion capabilities of full-sized Humanoid without vision or LiDAR sensors.** (a) *Narrow Path (25cm)*: The humanoid traverses a narrow pathway, including slopes and stairs, demonstrating precise foot placement and dynamic balance. (b) *Unknown Obstacle*: The humanoid robot showcases its dynamic balance control by swiftly adapting to the moving stick’s attempts to trip it, maintaining stability even in this challenging scenario. (c) *Carry Payload*: Our method can maintain stability while carrying loads, highlighting its robust control. (d) *Dense Conical Obstacles*: The humanoid steps over a series of closely spaced cones, exhibiting agility and coordination. (e) *External Pushes*: The system responds to external forces applied during locomotion over uneven terrain, proving its resilience against disturbances. Each scenario underscores the DBHL’s versatility and effectiveness in handling complex conditions.

widths and degrees. The results demonstrate that DBHL exhibits significantly better stability compared to other mainstream methods. We also analyze the role of the ZMP-reward through comparison, highlighting its crucial role in maintaining dynamic balance. Additionally, we conducted a detailed ablation study on other design elements. Our real-world experiments using the full-sized Unitree H1-2 robot illustrate the robot’s capability to move forward and backward in narrow terrain, handle disturbances such as pushes and trips, and showcase enhanced stability and adaptability in various extreme scenarios, as shown in Fig. 1.

The key contributions are summarized as follows.

- We integrate ZMP into the RL-based humanoid control framework as a novel reward function, realizing dynamic balance in complex terrains.
- We construct a whole-body control framework with newly introduced techniques including reward vectorization, angular-momentum regularization and action-space noise.
- We evaluate the proposed method via extensive experiments in both the simulation and the real-world using full-sized humanoid robots.

II. RELATED WORK

A. Humanoid Locomotion

Research on humanoid locomotion can be traced back to the 1970s [14]. The fundamental idea for a locomotion controller is to decompose it into planning and tracking modules, where the planning module is responsible for generating desired trajectories, while the tracking module ensures

that the robot follows these trajectories accurately [15], [16]. Methods such as Whole-Body Control (WBC) and Model Predictive Control (MPC) have achieved significant success in this domain [17], [18], [19]. However, these methods typically require precise modeling of the dynamics [20], [21], which poses substantial challenges for complex robot structures. In recent years, learning-based algorithms have emerged as a promising alternative to legged locomotion [1], [22] with efficient parallel simulation [8], significantly reducing the cost of interaction between the robot and the environment. Policies trained extensively in simulation environments can then be transferred to real robots [23]. In the field of quadrupedal locomotion, RL algorithms have demonstrated excellent performance in complex tasks such as complex-terrain walking [24], [25], gait control [26], [27], and even parkour [28], [29].

For RL-based humanoid locomotion, things become more difficult due to their limited support areas and high gravity. Meanwhile, classical control algorithms are also limited by the inaccurate modeling of the complex dynamical system. Recent approaches have proposed using RL algorithms for phase-based gait learning [2], [3], motor skill control [4], [5], and motion imitation [6], [7]. However, these methods still lag far behind that of humans in terms of dynamic balance under complex terrain (e.g., narrow paths) and extreme conditions (e.g., sudden disturbance). In our work, we address this problem by measuring the relationship between the ZMP and the humanoid support polygon in an RL framework. We note that a concurrent work uses foothold rewards to pass through narrow areas, while it relies on a

LiDAR-based elevation map for real-world deployment [30]. In contrast, our method can traverse complex terrains only using proprioception without vision or LiDAR perception.

B. Zero-moment Point

ZMP became a crucial tool in classical humanoid locomotion a few decades ago [11], which provides a framework for ensuring dynamic stability in bipedal robots. Formally, ZMP is defined as the point on the ground at which the net moment of the inertial forces and the gravity forces has no component along the horizontal axes [12]. In subsequent research, the ZMP concept has been instrumental in gait synthesis and has been integrated with advanced sensors to facilitate real-time balance adjustments [31]. In addition, it has inspired the exploration of innovative materials and foot designs to enhance the interaction of robots [32]. In our work, we extend the ZMP as a reward function to measure the relationship between the line of ZMPs and the support polygon, which enables the humanoid to maintain dynamic balance in complex terrains without relying on external perception.

III. METHOD

In this section, we present our method for training an end-to-end RL policy that enables humanoid robots to traverse extreme terrains using only proprioceptive information. We formulate our problem as a goal-conditioned RL task, where the policy π is trained to follow the target velocity command. The action \mathbf{a}_t represents the target joint positions, which are fed into the PD controller to actuate the robot's degrees of freedom. The agent's observation \mathbf{o}_t comprises velocity command \mathbf{c}_t and the history of proprioception information $\mathbf{s}_t^{\text{prop}}$. We employ the Proximal Policy Optimization (PPO) algorithm [33] with asymmetric actor-critic networks [34] to maximize the cumulative discounted reward of the policy. The whole architecture is shown in Fig. 2.

A. ZMP-based Dynamic Balance

To achieve dynamic balance, we design a ZMP-based reward and integrate it into the RL framework. With the standard definition, ZMP is the point on the ground where the horizontal component of the moment of the ground reaction forces becomes zero, assuming the ground is a horizontal plane. A necessary condition for the locomotion mechanism to be in dynamic equilibrium is that the candidate ZMP is inside the support polygon [31]. Traditional control methods achieve stable locomotion by planning the ZMP trajectory, and ensuring it stays within the support polygon. Furthermore, to handle non-planar surfaces, the concept of the Zero Moment Line (ZML) is introduced, which extends the ZMP concept to account for complex contact surfaces [35]. In DBHL, the ZMP-based reward is calculated by the distance between the ZML and the center of the support polygon formed by the robot's feet, as shown in Fig. 3.

Then we give the calculation process as follows. Assuming that the ground reaction force is the only external force applied to the robot system, the moment of the ground

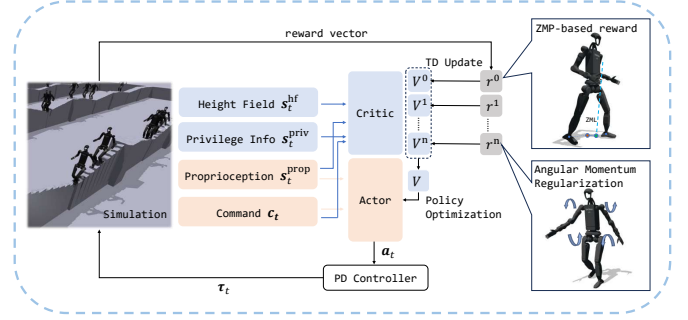


Fig. 2: The overall training process of the proposed method.

reaction force about the origin of the world frame is given by

$$\boldsymbol{\tau} = \mathbf{p}_{\text{zmp}} \times \mathbf{f} + \boldsymbol{\tau}_p, \quad (1)$$

where \mathbf{p}_{zmp} is the position of ZMP, $\boldsymbol{\tau}_p$ is the moment about the ZMP, and \mathbf{f} represents the ground reaction force. By Newton-Euler equations, we have the following relationship

$$\begin{cases} \dot{\mathbf{P}} = M\mathbf{g} + \mathbf{f} \\ \dot{\mathbf{L}} = \mathbf{p}_{\text{CoM}} \times M\mathbf{g} + \boldsymbol{\tau}, \end{cases} \quad (2)$$

where \mathbf{P} , \mathbf{L} , M , \mathbf{p}_{CoM} represents the linear momentum, the angular momentum, the total mass, the center of mass (CoM) position of the system, respectively, and $\mathbf{g} = [0, 0, -g]^\top$. By the definition of ZMP, we have

$$\tau_{p,x} = \tau_{p,y} = 0, \quad (3)$$

Solving the above equations with respect to \mathbf{p}_{zmp} , we have

$$p_{\text{zmp},x} = \frac{Mgp_{\text{CoM},x} + p_{\text{zmp},z}\dot{P}_x - \dot{L}_y}{Mg + \dot{P}_z} \quad (4)$$

$$p_{\text{zmp},y} = \frac{Mgp_{\text{CoM},y} + p_{\text{zmp},z}\dot{P}_y + \dot{L}_x}{Mg + \dot{P}_z}. \quad (5)$$

According to Eq. (4) and Eq. (5), ZMP at different heights $p_{\text{CoM},z}$ lies at different locations, whose trajectory forms a line of ZMPs, called ZML [35]. The condition for stable locomotion is that the ZML must cross the support polygon.

Then, we need to estimate the support polygon of the humanoid robot to determine the relationship between ZML and the support. In general cases, the support polygon is determined by computing the convex hull of all contact points for scenarios with multiple contact points. However, the calculation becomes intricate when dealing with irregular contact surfaces, such as slopes or uneven terrain, where the contact points must be projected onto a horizontal plane, and the resulting convex hull is adjusted to account for the tilt [36]. Also, in movements, the support polygon should be updated to reflect the changing contact points. In DBHL, we can simplify the standard calculation because, in a RL framework, we only need to design a reward function to encourage the ZML to be close to the center of the support polygon, which can be achieved without strictly estimating the support polygon. Specifically, the geometric center of

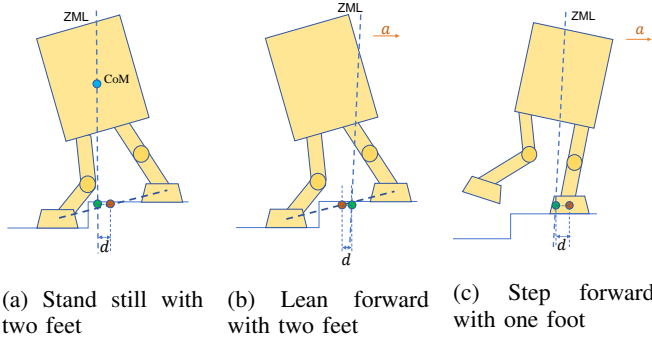


Fig. 3: The illustration of our ZMP-based reward in different locomotion conditions. The brown dot represents the approximated center of the support polygon, and the green dot on the ZML is at the same horizontal height as the brown dot.

support polygon (denoted as \mathbf{p}_{csp}) is approximated by the center of supporting feet, as

$$\mathbf{p}_{\text{csp}} = \frac{\mathbf{p}_{\text{left-foot}} \cdot (c_{\text{left-foot}} + \epsilon) + \mathbf{p}_{\text{right-foot}} \cdot (c_{\text{right-foot}} + \epsilon)}{c_{\text{left-foot}} + c_{\text{right-foot}} + 2\epsilon}, \quad (6)$$

where $\mathbf{p}_{\text{left-foot}}$ and $\mathbf{p}_{\text{right-foot}}$ are the center of the left and right foot, respectively; $c_{\text{left-foot}} = \mathbb{1}[\|\mathbf{f}_{\text{left-foot}}\|_2 > 0]$ and $c_{\text{right-foot}} = \mathbb{1}[\|\mathbf{f}_{\text{right-foot}}\|_2 > 0]$ are indicator functions to determine whether the foot is in contact with the ground. According to Eq. (6), if a humanoid robot is supported by two feet, \mathbf{p}_{csp} is approximately equal to the center of two feet. And if a humanoid robot is supported by one foot, \mathbf{p}_{csp} is approximately equal to either $\mathbf{p}_{\text{left-foot}}$ or $\mathbf{p}_{\text{right-foot}}$.

Based on the simplified \mathbf{p}_{csp} , we design a computationally tractable reward function, denoted as r_{zmp} , with the horizontal distance between \mathbf{p}_{csp} and the ZML, as

$$r_{\text{zmp}} = \exp(-\|\mathbf{p}_{\text{csp}} - \text{Proj}_{\text{ZML}}(\mathbf{p}_{\text{csp}})\|/0.05), \quad (7)$$

where $\text{Proj}_{\text{ZML}}(\mathbf{p}_{\text{csp}})$ is a horizontal projection function that projects \mathbf{p}_{csp} to the ZML. Intuitively, dynamic stability is guaranteed when \mathbf{p}_{csp} is close to the ZML, and a smaller distance indicates better stability, as illustrated by Fig. 3.

B. Whole-Body Locomotion

To achieve coordinated whole-body motion and enhance dynamic stability, we propose a whole-body locomotion framework that leverages upper-body swings to assist dynamic balance. To achieve this, we incorporate two key techniques: angular momentum regularization and multiplicative action noise injection.

1) *Angular Momentum Regularization*: It is introduced to minimize the undesired rotational motion during locomotion, thus improving coordination and improving resistance to external disturbances. As extensively discussed in prior work [37], [38], the motion of swinging legs generates significant angular momentum, which can disrupt whole-body motion. This effect can be mitigated through the proper use of upper limbs to counterbalance the momentum. We introduce a regularization reward r_{AM} based on the L_2 norm of the

total angular momentum about the robot base position $\mathcal{L}_{\text{base}}$, which is defined as

$$r_{\text{AM}} = \exp(-\|\mathcal{L}_{\text{base}}\|/5), \quad (8)$$

where

$$\mathcal{L}_{\text{base}} = \sum_{i=1}^n \mathbf{p}_i \times (m_i \mathbf{v}_i) + \mathbf{I}_i \boldsymbol{\omega}_i, \quad (9)$$

where \mathbf{p}_i , m_i , \mathbf{I}_i , \mathbf{v}_i , $\boldsymbol{\omega}_i$ represent the position vector of the CoM, mass, inertia tensor, linear velocity, and angular velocity of the i -th link, respectively.

2) *Multiplicative Action Noise Injection*: It is employed to constrain the range of motion of the upper body joints to enhance the robustness of the policy. If the policy is not restricted, the locomotion policy can often result in unstable and large-angle upper limb movements. Such a technique is implemented by modifying the input to the PD controller, where the nominal action \mathbf{a}_t is perturbed by a multiplicative noise term, as

$$\mathbf{a}'_t = \mathbf{a}_t(1 + \sigma_{\text{AN}} \boldsymbol{\epsilon}_t), \quad (10)$$

where $\boldsymbol{\epsilon}_t \sim N(0, \mathbf{I})$. By applying large perturbations to actions with greater magnitudes, the policy is encouraged to favor small actions for the upper body joints, promoting cautious behavior when significant movements are required.

C. Policy Learning Details

1) *Task and Command*: We developed three types of narrow terrain in policy training, including flat, stairs, and slopes. Meanwhile, a terrain-curriculum mechanism introduced in [8] is used in training with 20 difficulty levels, where the path width gradually decreases from 1.0 to 0.2 m, the slope gradient increases from 0 to 0.3, and the step height increases from 0 to 0.12 m. The robot is trained in a mixing of these narrow terrains, each containing 30% with an additional 10% planar terrain. For each episode, the linear velocity command is uniformly sampled in the range $\hat{v}_x \in [-0.5\text{m/s}, 1.0\text{m/s}]$ and $\hat{v}_y \in [-0.2\text{m/s}, 0.2\text{m/s}]$, while the yaw velocity command is determined by $\hat{\omega}_{\text{yaw},t} = \text{clip}(0.5 * \Delta\theta_{\text{yaw}}, -1 \text{ rad/s}, 1 \text{ rad/s})$ where $\Delta\theta_{\text{yaw}}$ is the horizontal angle between positive X-axis direction and the orientation of the robot.

2) *Asymmetric Actor-Critic Framework*: The observation for the actor network $\mathbf{o}_t \in \mathbb{R}^{279}$ comprises velocity command $\mathbf{c}_t = [\hat{v}_{x,t}, \hat{v}_{y,t}, \hat{\omega}_{\text{yaw},t}]$ and the robot's proprioception $\mathbf{s}_t^{\text{prop}} = [\mathbf{q}_{t-3:t}, \dot{\mathbf{q}}_{t-3:t}, \boldsymbol{\omega}_{t-3:t}, \mathbf{g}_{t-3:t}, \mathbf{a}_{t-4:t-1}]$ with 4-step history of joint position $\mathbf{q}_t \in \mathbb{R}^{21}$, joint velocity $\dot{\mathbf{q}}_t \in \mathbb{R}^{21}$, base angular velocity $\boldsymbol{\omega}_t \in \mathbb{R}^3$, base projected gravity $\mathbf{g}_t \in \mathbb{R}^3$ and last action $\mathbf{a}_{t-1} \in \mathbb{R}^{21}$. The observation for the critic network \mathbf{s}_t includes actor observation \mathbf{o}_t , privileged information $\mathbf{s}_t^{\text{priv}} \in \mathbb{R}^{70}$ and surrounding height field $\mathbf{s}_t^{\text{hf}} \in \mathbb{R}^{187}$. The privileged observation $\mathbf{s}_t^{\text{priv}}$ contain the linear velocity, the base height, feet contact indicator, randomized PD parameter and randomized link mass. The height field \mathbf{s}_t^{hf} is sampled from a $1.6\text{m} \times 1.0\text{m}$ area around the robot, with a point spacing of 0.1m. This framework leverages privileged information to enhance value function estimation for policy guidance, while restricting the actor to

a local state ensures the policy’s transferability to real-world environments.

3) *Reward and Value Vectorization*: To facilitate the learning of the value function with different reward functions, we introduce the vectorization of the reward and value function. Instead of aggregating all reward terms into a single scalar and learning a single value function, we combine the different rewards as a vector and learn the corresponding value functions via Temporal-Difference (TD) learning independently. Then, we obtain a set of value functions, each associated with a specific TD target. To achieve this, the value function is implemented by a neural network with multiple output heads. Then, all value functions are aggregated in computing the action advantage function. This method addresses a key limitation in traditional approaches, where summing all rewards makes it difficult for the value function to capture changes of reward terms with relatively small magnitudes. Specifically, we have

$$V^{\text{total}}(s_t) = \sum_{i=1}^{\text{\#Reward}} V^i(s_t), \quad (11)$$

and the loss function to train a value function is given by

$$\mathcal{L}_{\text{value}} = \sum_{i=1}^{\text{\#Reward}} \mathbb{E} \left[\left\| r_t^i + \gamma V^i(s_{t+1}) - V^i(s_t) \right\|^2 \right]. \quad (12)$$

The reward functions of DBHL are given in Table. I.

TABLE I: Reward functions of DBHL. The feet contact reward encourages single contact between the feet and the ground. The feet edge distance reward encourages the feet to stay far from the ground edge. The action closeness reward encourages the action to be close to the current DOF position.

Group	Reward Function
Task	Linear Velocity Tracking, Angular Velocity Tracking, Low Speed Penalty
Gait	ZMP, Feet Air Time, Feet Contact, Feet Separation, Feet Slippage, Feet Height, Base Height, Feet Edge Distance
Regularization	Angular Momentum, Orientation, Base Acceleration, Action Smoothness, Action Closeness, Torque, DOF Velocity, DOF Position Limit, Collision

4) *Symmetry Regularization*: We introduce a symmetry loss term, inspired by [39], to enhance sample efficiency and promote more harmonious gaits. This loss leverages the symmetry of the robot’s motion with respect to the x - z plane, defined as

$$\mathcal{L}_{\text{sym}} = \mathbb{E} \left[\left\| \mathbf{V}(G(s_t)) - \mathbf{V}(s_t) \right\|^2 + \left\| \boldsymbol{\pi}(G(o_t)) - G(\boldsymbol{\pi}(o_t)) \right\|^2 \right], \quad (13)$$

where G denotes the reflection operator across the x - z plane.

5) *Domain Randomization*: To enable zero-shot sim-to-real transfer, we randomize the physical parameters of the simulated environment and humanoids, listed in Table II.

IV. EXPERIMENT

In this section, we present extensive experiments to evaluate the proposed method. Our experiments aim to address the following key questions:

TABLE II: Domain randomization settings.

Term	Value
External Push	$\Delta T \sim \text{Exp}(6)\text{s}$, $\Delta \mathbf{v}_{xy} \sim \mathcal{U}(-0.6, 0.6)\text{m/s}$, $\Delta \boldsymbol{\omega} \sim \mathcal{U}(-0.8, 0.8)\text{rad/s}$
Action Delay	$\mathcal{U}(4, 20)\text{ms}$
P Gain	$\mathcal{U}(0.8, 1.2) \times \text{default}$
D Gain	$\mathcal{U}(0.8, 1.2) \times \text{default}$
Friction	$\mathcal{U}(0.1, 2)$
Link Mass	$\mathcal{U}(0.8, 1.2) \times \text{default}$
Load Mass	$\mathcal{U}(-1, 3)\text{kg}$
Base CoM Offset	$\mathcal{U}(-0.1, 0.1)\text{m}$
Torque RFI	$\Delta \boldsymbol{\tau} \sim \mathcal{U}(-0.1, 0.1) \times \tau_{\text{rfi}} \times \text{torque limit N}\cdot\text{m}$, $\tau_{\text{rfi}} \sim \mathcal{U}(0.5, 1.5)$
Action Noise	$\sigma_{\text{AN}} = 0.03$, see Sec. III-B.2

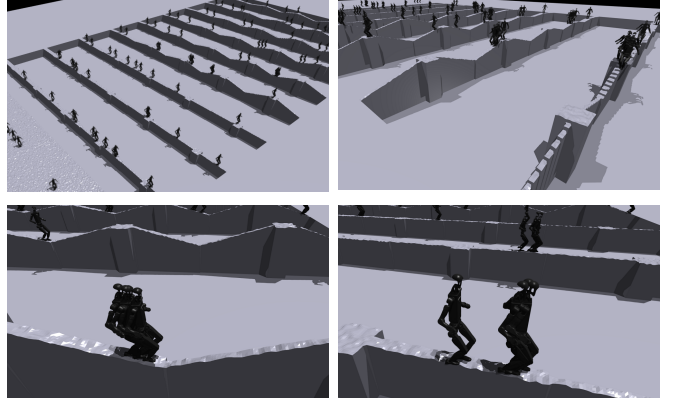


Fig. 4: Visualization of the various training terrains of our method in Isaac Gym.

- Q1: Can DBHL outperform other methods in various extreme terrains?
- Q2: How does ZMP reward help with dynamic balance?
- Q3: How does design choices (i.e., reward vectorization and action noise) influence training performance?
- Q4: Can DBHL transfer to real-world hardware?

Experiments Setup. We conduct experiments on Unitree H1-2, which is a full-sized humanoid robot with 27 DoF. The policy controls 21 DoF, excluding the 3 DoF in each wrist of the hands. We train the RL policy in Isaac Gym [40] and use MuJoCo [41] as a *sim-to-sim* verification platform. As shown in Fig. 4, we visualize the different training terrains in the Isaac Gym simulation. In our experiments, we evaluate the performance across the three types of terrains (i.e., narrow flat, narrow slope, and narrow stairs), where each terrain has varying road widths and difficulties (i.e., maximum linear velocity of random push in flat, slope gradient in slope, and step height in stairs). For evaluation, we sample 10^4 episodes with 20s duration and 0.5m/s heading velocity in Isaac Gym, and 100 episodes with 25s duration and 0.3m/s heading velocity in MuJoCo. The discrepancy in evaluation episode is due to MuJoCo is renowned for its high-fidelity physics modeling, enabling reliable results with fewer episodes. Meanwhile, the policy is trained on Isaac Gym and transferred to MuJoCo, which often leads to decreased performance. Thus, we adopt a longer duration and more conservative command for MuJoCo evaluation. We report the success rate (where success means the move distance

TABLE III: Comparison of our method to baselines in various terrains under one of the difficulty settings (i.e., flat push 0.6m/s, slope gradient 0.2, stair height 8cm). The second row indicates the road widths in meters. The best results are highlighted in boldface. Our method significantly outperforms baselines.

Metric	Method	Flat			Slope			Stairs		
		0.25	0.3	0.35	0.25	0.3	0.35	0.25	0.3	0.35
Success Rate	DBHL	0.90 \pm 0.02	0.94 \pm 0.02	0.99 \pm 0.01	0.92 \pm 0.02	0.96 \pm 0.02	1.00 \pm 0.00	0.89 \pm 0.02	0.93 \pm 0.02	0.98 \pm 0.01
	DBHL w/o Upper	0.00 \pm 0.00	0.04 \pm 0.03	0.18 \pm 0.03	0.01 \pm 0.01	0.13 \pm 0.05	0.24 \pm 0.04	0.00 \pm 0.00	0.12 \pm 0.04	0.25 \pm 0.04
MXD	URG	0.01 \pm 0.01	0.14 \pm 0.12	0.12 \pm 0.04	0.03 \pm 0.02	0.18 \pm 0.12	0.34 \pm 0.21	0.02 \pm 0.02	0.12 \pm 0.08	0.22 \pm 0.14
	DBHL	6.58 \pm 0.17	6.78 \pm 0.21	7.02 \pm 0.20	5.83 \pm 0.11	6.20 \pm 0.15	6.52 \pm 0.20	5.62 \pm 0.20	5.79 \pm 0.20	6.03 \pm 0.23
	DBHL w/o Upper	0.85 \pm 0.04	1.16 \pm 0.17	1.92 \pm 0.25	1.07 \pm 0.08	1.47 \pm 0.15	1.81 \pm 0.17	1.07 \pm 0.06	1.55 \pm 0.16	2.10 \pm 0.16
	URG	1.33 \pm 0.11	1.93 \pm 0.73	2.02 \pm 0.37	1.61 \pm 0.24	2.24 \pm 0.59	3.09 \pm 1.04	1.48 \pm 0.15	1.91 \pm 0.43	2.42 \pm 0.67

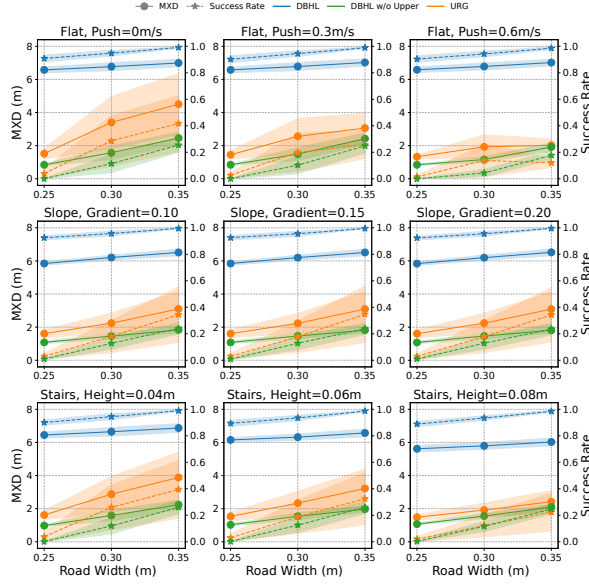


Fig. 5: The result comparison across various terrains and difficulties. The result shows whole-body control is essential for DBHL, and the dynamic balance mechanism is more effective than phase-based gait in challenging conditions. Each setting is evaluated over 3 random seeds.

$\geq 4m$) and the Mean X-Displacement (MXD) as metrics.

A. Result Comparison in Narrow Terrains

To address Q1, we compare our methods with the following baselines in Isaac Gym:

- DBHL w/o Upper: This variant of our method fixes the upper body, focusing solely on the lower-body control.
- Unitree RL Gym (URG): The official RL framework provided by Unitree, which basically follows [1] by employing asymmetric actor-critic for lower-body control and relying on phase-based rewards to learn periodic gaits. We also incorporate the height field as part of the observation space to ensure a fair comparison.

According to Fig. 5 and Table III, our method consistently outperforms the baselines in terms of both success rate and MXD. The results highlight the critical role of whole-body control in our approach. The performance drops significantly when only the lower body is controlled, indicating that narrow-terrain locomotion requires the coordinated effort of the entire robot. Meanwhile, our method demonstrates superior robustness compared to the phase-based control

method across various terrains, achieving higher success rates and better MXD at all difficulty levels. This underscores the importance of phase-free locomotion and whole-body control in complex environments.

B. Analysis of ZMP-based Reward

To address Q2, we compare the performance of policy trained with and without the ZMP-based reward, focusing on dynamic balance and task performance.

1) *Quantitative Comparison:* Fig. 6 presents a quantitative comparison of success rate and MXD between variants with and without ZMP-based reward in MuJoCo. The reason to compare in MuJoCo is to leverage its high-fidelity simulation for sensitive detection of subtle policy differences. The results clearly show that incorporating ZMP-based rewards allows the robot to traverse narrow and challenging terrains more effectively than the non-ZMP variant, resulting in a significantly better success rate and MXD in most settings. An special case for non-ZMP variant arises when the slope gradient is 0.2, where the performance surpasses that observed on slopes of 0.1 and 0.15. We hypothesize that the non-ZMP variant may overfit to this condition, resulting in better performance compared to less challenging settings.

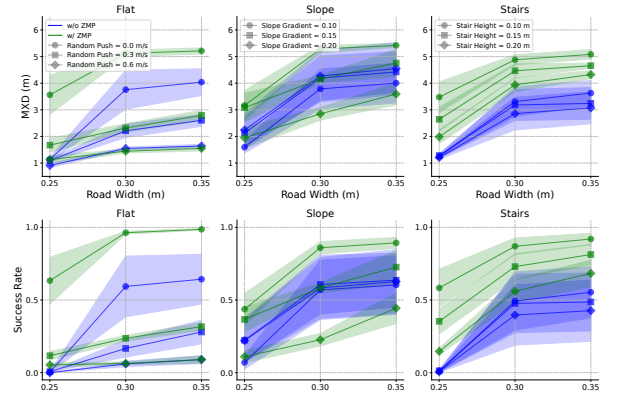
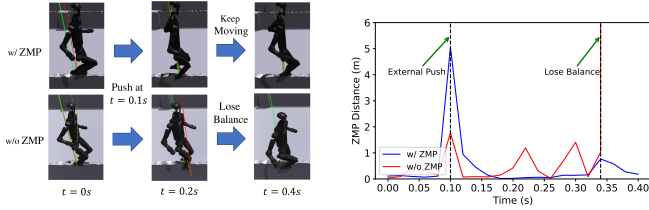


Fig. 6: Quantitative comparison for ZMP-based rewards. The inclusion of the ZMP reward improves the robot’s dynamic stability, leading to a better success rate and MXD on narrow terrains. Each setting is evaluated over 3 random seeds. The shaded region around each curve represents $\pm 0.5\sigma$ area, indicating the variability of the results.

2) *Qualitative Comparison:* In Fig. 7(a), we provide two example trajectories for a qualitative comparison. At $t = 0.1s$, an external push ($\Delta v_y = 0.6m/s$) is exerted



(a) Snapshots of trajectories w/ and w/o. ZMP-based reward.

(b) ZMP-distance over time

Fig. 7: Comparison of two example trajectories under external pushes. Without the ZMP-based reward, the robot loses balance and falls after pushing, which can be observed by the diverging ZMP distance. In contrast, DBHL with ZMP rewards maintains dynamic balance. In snapshots, the yellow lines represent ZML while the green and red ones represent vertical lines crossing p_{csp} and $Proj_{ZML}(p_{csp})$, respectively.

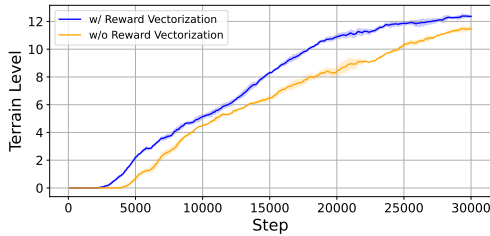


Fig. 8: Comparison of training curves with and without reward vectorization. The terrain level represents the average difficulty level determined by the terrain-curriculum mechanism, with a total of 20 predefined levels. Each setting is evaluated over 3 random seeds. The training curves are sampled at 100-step intervals and smoothed using an exponential moving average with a smoothing factor of 0.1.

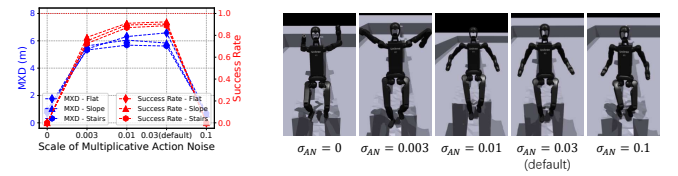
on the robot. The trajectory without ZMP-based reward results in the robot losing balance, where the corresponding ZMP-distance, defined as $\|p_{csp} - Proj_{ZML}(p_{csp})\|$ in Eq. 7, becomes particularly large and eventually diverges, as shown in Fig. 7(b). In contrast, the policy incorporating ZMP-based reward successfully maintains the robot's stability throughout the disturbance since the distance is constrained by maximizing the ZMP-reward in Eq. (7).

C. Ablation Study

To address Q3, we perform ablation studies in Isaac Gym to investigate the effect of two key design choices: reward vectorization and multiplicative action noise, which aims to evaluate how each of these factors influences the performance of our method.

1) *Effect of Reward Vectorization*: Fig. 8 shows the training curves for the policy with and without reward vectorization. The use of reward vectorization significantly accelerates the learning process, facilitating faster policy convergence. This improvement is attributed to the reward and value vectorization framework, which enables DBHL to learn each value term associated with each reward term independently, thus increasing overall learning efficiency.

2) *Effect of Multiplicative Action Noise*: Fig. 9 presents the results for varying scales of action noise. Our experiments



(a) Success rate and MXD for varying scales of the action noise.

(b) Snapshots of the policy behaviors for varying scales of the action noise.

Fig. 9: Ablation study for action noise. A proper scale of action noise can constrain the range of motion of the upper body joints and improve task performance. We conduct experiments on narrow terrains with a width of 25cm.

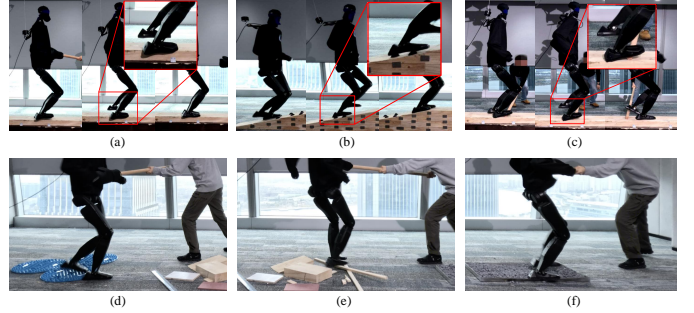


Fig. 10: The real-world experiments, including (a) narrow path under perturbations, (b) sloped terrain with conical obstacles, (c) moving stick trips, (d) acupuncture plates, (e) wooden block obstacles, and (f) stone roads, showcasing DBHL's efficacy in addressing real-world challenges.

indicate that an appropriately calibrated level of action noise effectively confines the motion range of the upper body joints. Conversely, both excessively low and high levels of action noise lead to instabilities in accomplishing the task.

D. Real-World Experiment

To address Q4, we deploy DBHL on the Unitree H1-2 robot and evaluate its performance in both narrow and planar terrains through a series of real-world experiments. The evaluation platform consists of a narrow wooden track measuring 25 cm in width. This track includes three sections: a sloped ramp with a gradient of 0.2 and a length of 1.6 m, a bridge section spanning 1.6 m, and a set of stairs with a step width of 40 cm and a step height of 8 cm.

1) *External Disturbances in Narrow Terrains*: As shown in Fig. 1 and Fig. 10 (a-c), DBHL can traverse the narrow track successfully. The controller effectively handles a range of challenging disturbance conditions, including carrying a 5 kg payload, passing through dense conical obstacles, withstanding human pushes, and walking on a levered stick.

2) *Irregular Obstacles in Uneven Ground*: Since the previous experiments focus solely on narrow terrains, we train an additional policy on irregular planar terrain. Then we evaluate the policy's capabilities by incorporating a variety of irregular obstacles on flat terrain in the real world, as depicted in Figures 10 (d-f). This environment features acupuncture plates, wooden blocks, planks, and stone-paved paths. In addition, the robot must contend with applied

external forces while traversing these challenging conditions. The results shows our method can overcome these complex terrain challenges.

V. CONCLUSIONS

This work presents DBHL, a novel reinforcement learning framework that enables humanoid robots to traverse extreme terrains by introducing a ZMP-based reward function and a whole-body control framework. Through extensive simulations and real-world experiments, DBHL demonstrates superior performance in narrow and uneven terrains, highlighting its robustness and adaptability. The proposed methodology opens new possibilities for real-world applications requiring extreme mobility and balance.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (Grant No.62306242).

REFERENCES

- [1] X. Gu, Y.-J. Wang, and J. Chen, "Humanoid-gym: Reinforcement learning for humanoid robot with zero-shot sim2real transfer," *arXiv preprint arXiv:2404.05695*, 2024.
- [2] X. Gu, Y.-J. Wang, X. Zhu, C. Shi, Y. Guo, Y. Liu, and J. Chen, "Advancing humanoid locomotion: Mastering challenging terrains with denoising world model learning," in *RSS*, 2024.
- [3] Y. Xue, W. Dong, M. Liu, W. Zhang, and J. Pang, "A unified and general humanoid whole-body controller for fine-grained locomotion," *arXiv preprint arXiv:2502.03206*, 2025.
- [4] Y. Guo, Z. Jiang, Y.-J. Wang, J. Gao, and J. Chen, "Decentralized motor skill learning for complex robotic systems," *IEEE RA-L*, 2023.
- [5] Q. Zhang, C. Weng, G. Li, F. He, and Y. Cai, "Hilo: Learning whole-body human-like locomotion with motion tracking controller," *arXiv preprint arXiv:2502.01122*, 2025.
- [6] T. Peng, L. Bao, J. Humphreys, A. M. Delfaki, D. Kanoulas, and C. Zhou, "Learning bipedal walking on a quadruped robot via adversarial motion priors," in *Annual Conference Towards Autonomous Robotic Systems*. Springer, 2024, pp. 118–129.
- [7] Z. Zhuang and H. Zhao, "Embrace collisions: Humanoid shadowing for deployable contact-agnostics motions," *arXiv preprint arXiv:2502.01465*, 2025.
- [8] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," in *Conference on Robot Learning*, 2022, pp. 91–100.
- [9] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, F. Janoos, L. Rudolph, and A. Madry, "Implementation matters in deep rl: A case study on ppo and trpo," in *ICLR*, 2019.
- [10] G. B. Margolis and P. Agrawal, "Walk these ways: Tuning robot control for generalization with multiplicity of behavior," in *CoRL*, 2023, pp. 22–31.
- [11] M. Vukobratovic and J. Stepanenko, "On the stability of anthropomorphic systems," *Mathematical Biosciences*, vol. 15, pp. 1–37, 1972.
- [12] M. Vukobratović, B. Borovac, D. Surla, and D. Stokić, *Biped Locomotion: Dynamics, Stability, Control and Application*. Berlin: Springer-Verlag, 1990.
- [13] I. Yamaguchi, A. Takanishi, and I. Kato, "Development of a biped walking robot compensation for three — axis moment by trunk motion," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robot and Systems*, Yokohama, Japan, 1993.
- [14] I. Kato, S. Ohteru, H. Kobayashi, K. Shirai, and A. Uchiyama, "Information-power machine with senses and limbs (wabot 1)," in *First CISM-IFTOMM Symposium on Theory and Practice of Robots and Manipulators*, vol. 1. Vienna: Springer-Verlag, 1974, pp. 11–24.
- [15] R. Grandia, F. Jenelten, S. Yang, F. Farshidian, and M. Hutter, "Perceptive locomotion through nonlinear model-predictive control," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3402–3421, 2023.
- [16] A. Meduri, P. Shah, J. Viereck, M. Khadiv, I. Havoutis, and L. Righetti, "Biconmp: A nonlinear model predictive control framework for whole body motion planning," *IEEE Transactions on Robotics*, vol. 39, no. 2, pp. 905–922, 2023.
- [17] J. Li and Q. Nguyen, "Multi-contact mpc for dynamic locomotion on humanoid robots," in *American Control Conference (ACC)*. IEEE, 2023, pp. 1215–1220.
- [18] L. Sentis and O. Khatib, "A whole-body control framework for humanoids operating in human environments," in *ICRA*. Orlando, FL, USA: IEEE, 2006, pp. 2641–2648.
- [19] J.-P. Sleiman, F. Farshidian, M. V. Minniti, and M. Hutter, "A unified mpc framework for whole-body dynamic locomotion and manipulation," *IEEE RA-L*, vol. 6, no. 3, pp. 4688–4695, 2021.
- [20] J. Koenemann, A. D. Prete, Y. Tassa, E. Todorov, O. Stasse, M. Bennewitz, and N. Mansard, "Whole-body model-predictive control applied to the HRP-2 humanoid," in *IROS*. Hamburg, Germany: IEEE, 2015, pp. 3346–3351.
- [21] G. Schultz and K. Mombaur, "Modeling and optimal control of human-like running," *IEEE/ASME Transactions on Mechatronics*, vol. 15, no. 5, pp. 783–792, 2009.
- [22] S. Ha, J. Lee, M. van de Panne, Z. Xie, W. Yu, and M. Khadiv, "Learning-based legged locomotion: State of the art and future perspectives," *The International Journal of Robotics Research*, 2024.
- [23] H. He, P. Wu, C. Bai, H. Lai, L. Wang, L. Pan, X. Hu, and W. Zhang, "Bridging the sim-to-real gap from the information bottleneck perspective," in *CoRL*, 2024.
- [24] J. Levy, T. Westenbroek, and D. Fridovich-Keil, "Learning to walk from three minutes of real-world data with semi-structured dynamics models," in *CoRL*, 2024.
- [25] J. Shi, C. Bai, H. He, L. Han, D. Wang, B. Zhao, M. Zhao, X. Li, and X. Li, "Robust quadrupedal locomotion via risk-averse policy learning," in *ICRA*. IEEE, 2024, pp. 11 459–11 466.
- [26] L. Han, Q. Zhu, Sheng, *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.
- [27] X. Huang, Y. Chi, R. Wang, Z. Li, X. B. Peng, S. Shao, B. Nikolic, and K. Sreenath, "Diffuselo: Real-time legged locomotion control with diffusion from offline datasets," in *CoRL*, 2024.
- [28] Z. Zhuang, Z. Fu, J. Wang, C. G. Atkeson, S. Schwertfeger, C. Finn, and H. Zhao, "Robot parkour learning," in *CoRL*, 2023.
- [29] X. Cheng, K. Shi, A. Agarwal, and D. Pathak, "Extreme parkour with legged robots," in *ICRA*. IEEE, 2024, pp. 11 443–11 450.
- [30] H. Wang, Z. Wang, J. Ren, Q. Ben, J. Pang, T. Huang, and W. Zhang, "Beamdojo: Learning agile humanoid locomotion on sparse footholds," 2024.
- [31] M. Vukobratović and B. Borovac, "Zero-moment point — thirty five years of its life," *International Journal of Humanoid Robotics*, vol. 1, no. 1, pp. 157–173, 2004.
- [32] S. Kajita, H. Hirukawa, K. Harada, and K. Yokoi, *Introduction to Humanoid Robotics*, 1st ed. Springer Publishing Company, 2016.
- [33] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
- [34] L. Pinto, M. Andrychowicz, P. Welinder, W. Zaremba, and P. Abbeel, "Asymmetric actor critic for image-based robot learning," in *RSS*. MIT Press Journals, 2018.
- [35] T. Breceelj and T. Petrič, "Zero moment line—universal stability parameter for multi-contact systems in three dimensions," *Sensors*, vol. 22, no. 15, p. 5656, 2022.
- [36] S. Caron, Q.-C. Pham, and Y. Nakamura, "Zmp support areas for multicontact mobility under frictional constraints," *IEEE Transactions on Robotics*, vol. 33, no. 1, pp. 67–80, 2017.
- [37] M. Popovic, A. Hofmann, and H. Herr, "Angular momentum regulation during human walking: biomechanics and control," in *IEEE International Conference on Robotics and Automation*, 2004. *Proceedings. ICRA '04*. 2004, vol. 3, 2004, pp. 2405–2411 Vol.3.
- [38] X. Zhang, X. Wang, L. Zhang, G. Guo, X. Shen, and W. Zhang, "Achieving stable high-speed locomotion for humanoid robots with deep reinforcement learning," 2024.
- [39] Z. Su, X. Huang, D. Ordonez-Apaez, *et al.*, "Leveraging symmetry in rl-based legged locomotion control," *IROS*, pp. 6899–6906, 2024.
- [40] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.
- [41] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 5026–5033.