



## Research Article

## Learning-based locomotion control fusing multimodal perception for a bipedal humanoid robot

Chao Ji<sup>a,b</sup>, Diyuan Liu<sup>b</sup>, Wei Gao<sup>a,\*</sup>, Shiwu Zhang<sup>a,\*</sup><sup>a</sup> School of Engineering Science, University of Science and Technology of China, Hefei 230026, China<sup>b</sup> iFLYTEK Co., Ltd., Hefei 230088, China

## ARTICLE INFO

## Article history:

Received 31 August 2024

Revised 2 January 2025

Accepted 7 January 2025

Available online 18 January 2025

## Keywords:

Bipedal humanoid robot

Deep reinforcement learning

Multimodal perception

## ABSTRACT

The ability of bipedal humanoid robots to walk adaptively on varied terrain is a critical challenge for practical applications, drawing substantial attention from academic and industrial research communities in recent years. Traditional model-based locomotion control methods have high modeling complexity, especially in complex terrain environments, making locomotion stability difficult to ensure. Reinforcement learning offers an end-to-end solution for locomotion control in humanoid robots. This approach typically relies solely on proprioceptive sensing to generate control policies, often resulting in increased robot body collisions during practical applications. Excessive collisions can damage the biped robot hardware, and more critically, the absence of multimodal input, such as vision, limits the robot's ability to perceive environmental context and adjust its gait trajectory promptly. This lack of multimodal perception also hampers stability and robustness during tasks. In this paper, visual information is added to the locomotion control problem of humanoid robot, and a three-stage multi-objective constraint policy distillation optimization algorithm is innovatively proposed. The expert policies of different terrains to meet the requirements of gait aesthetics are trained through reinforcement learning, and these expert policies are distilled into student through policy distillation. Experimental results demonstrate a significant reduction in collision rates when utilizing a control policy that integrates multimodal perception, especially in challenging terrains like stairs, thresholds, and mixed surfaces. This advancement supports the practical deployment of bipedal humanoid robots.

© 2025 The Author(s). Published by Elsevier B.V. on behalf of Shandong University. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Bipedal humanoid robots have a broad spectrum of potential applications, including logistics distribution, household services, and eldercare [1]. Realizing the full potential of these robots necessitates developing an effective control scheme. This study primarily investigates the terrain-adaptive walking capabilities of bipedal humanoid robots, aiming to minimize collision rates through sim-to-real reinforcement learning—a crucial capability for such applications. Currently, the prevailing locomotion control techniques for biped robots are optimization-based methods, primarily involving hierarchical control based on simplified models to optimize trajectories and employing feedback linearization to monitor them. Over the past decades, researchers have advanced various model-based locomotion control methodologies for biped robots. For instance, M. Shafiee-Ashtiani et al. introduced a comprehensive, resilient bipedal walking control policy that integrates model predictive control with the motion

divergence component [2].

Similarly, Ayonga Hereid et al. employed a hybrid zero dynamics model to quickly and reliably generate efficient walking gaits for multi-contact robots [3]. Reduced-order models, such as the linear inverted pendulum model [4] and the spring-loaded inverted pendulum model, have also become prominent in bipedal locomotion research. Notably, Agility Robotics' biped robot Cassie employs a locomotion control model based on the SLIP model [5–7]. Whole-body dynamic models for bipedal robots are typically high-dimensional and nonlinear, complicating optimization efforts [8–11].

Recent advancements in deep reinforcement learning have introduced learning-based methods to develop locomotion control policies [12–15] that aim to achieve dynamic locomotion behaviors. Unlike traditional optimization or heuristic control methods, reinforcement learning generates suitable policies by allowing the agent to interact with its environment, using sensory inputs and historical data while rewarding or penalizing behavior during training [16–18]. Most reinforcement learning studies on locomotion control for bipedal humanoid robots primarily rely on proprioceptive sensory data alone to formulate control policies. This approach requires the robot to perceive terrain information

\* Corresponding authors.

E-mail addresses: [weigao@ustc.edu.cn](mailto:weigao@ustc.edu.cn) (W. Gao), [swzhang@ustc.edu.cn](mailto:swzhang@ustc.edu.cn) (S. Zhang).

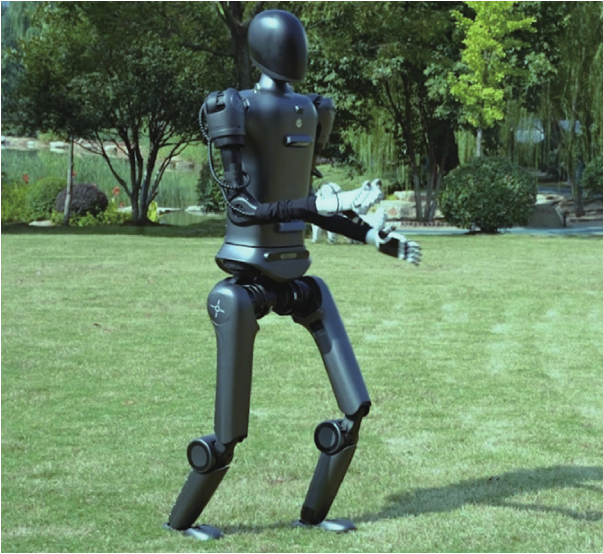


Fig. 1. Description of the overall hardware system.

via foot-ground collisions. As a result, in complex environments, bipedal humanoid robots may incur unpredictable damage from frequent terrain collisions. Moreover, this limited perception reduces the stability and adaptability of bipedal humanoid robots in real-world tasks. Without addressing this issue, the practical application of these robots remains restricted. Thus, minimizing collision rates between robots and terrains during locomotion is essential.

In recent years, researchers have explored enhancing training effectiveness by incorporating multimodal information, such as visual data, into the process [19–21]. R. Yang et al. achieved end-to-end visual navigation for a quadruped robot using depth images, demonstrating vision-based zero-sample transfer on flat, sloping, grassy, and sandy terrain [22]. H. Duan et al. proposed a vision-based learning approach for enabling biped robots to walk on complex terrain [23]. Their approach uses elevation maps for motion control based on external perception, though it only addresses foot-raising when the robot encounters elevated terrain. In contrast, our study focuses on more diverse policies, determining both foot-raising and stepping actions, which are integrated into a unified model. In the paper, our study integrates deep reinforcement learning with the proprioceptive sensing capabilities of a novel full-scale bipedal humanoid robot, Xiao-Man. A depth camera is incorporated to capture visual data from an RGBD sensor, offering insights into external environmental conditions. Xiao-Man is 1.5 meters tall and weighs 45 kg, with a total of 19 revolute degrees of freedom (DOFs). To address complex terrain challenges such as stairs and thresholds, the multi-expert policy distillation method is employed for Xiao-Man’s locomotion control.

The following sections begin with an introduction to the hardware design and policy training framework (Section 2). Section 3 details the methodology for policy training and experimental validation. Finally, the conclusions are presented, followed by a discussion of potential future research directions (Section 5).

## 2. System overview

### 2.1. Hardware system design

The comprehensive hardware system of Xiao-Man is shown in Fig. 1. The robot’s mechanical architecture consists of a trunk,

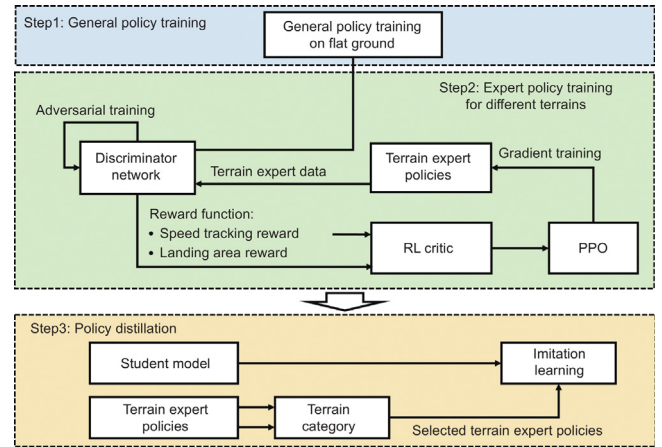


Fig. 2. Description of the policy training framework.

two arms, and two legs. Each leg contains five joints: three at the hip (roll, pitch, and yaw), one at the knee, and one at the ankle. Each arm contains four joints: three at the shoulder and one at the elbow. The trunk includes a single joint for yaw rotation. Each robot joint is equipped with an encoder for real-time position capture, and joint torque is estimated via actuation current sensing. The trunk integrates an Inertial Measurement Unit (IMU) to collect real-time body motion data relative to the center of mass (COM). Additionally, the trunk has a depth camera angled at 70° from the horizontal for terrain observation.

### 2.2. Policy training framework design

Training of the robot’s external perceptual locomotion model occurs in three stages, shown in Fig. 2. In Stage 1, a general policy model is trained to develop an aesthetically pleasing gait on flat ground, using shape-shaping rewards [24,25]. In Stage 2, expert policies are trained for common indoor terrains like stairs and thresholds, employing the Generative Adversarial Imitation Learning (GAIL) algorithm to provide a reward function that emulates the desired gait policy within each expert policy’s constraints. Landing area constraints guide the robot to target areas with larger support surfaces, and sparse speed tracking rewards ensure adherence to speed commands. In Stage 3, external perception noise in real environments is analyzed to construct a noise model for training, using imitation learning to distill multiple expert skills into a unified policy.

## 3. Policy training

### 3.1. General policy training

The general policy model primarily employs an actor-critic reinforcement learning algorithm, specifically Proximal Policy Optimization (PPO), to optimize the robot’s proprioceptive data. The gait optimization problem is initially formulated as a Markov Decision Process (MDP), defining the state space, action space, and reward function. During this training stage, both the actor and critic networks use a multilayer perceptron (MLP) architecture. Practically, the specified command velocity is transmitted to the robot via remote control.

The state space encompasses the robot’s proprioceptive data, as shown in Table 1. This data, received in various vectors, specifically includes the positions of the 19 revolute joints on the bipedal humanoid robot  $q \in \mathbb{R}^{19}$ ; the joint velocities  $\dot{q} \in \mathbb{R}^{19}$ ; the gravity vectors of the robot  $g \in \mathbb{R}^3$ ; the angular velocities of roll,

**Table 1**  
State space of the actor-critic network.

Input	Vector dimension	
	Actor	Critic
Joint Position	19	19
Joint Velocity	19	19
Gravity Vector	3	3
Angular Velocity	3	3
Linear Velocity	3	3
Linear Velocity Commands	3	3
Past Actions	19	19

pitch and yaw  $\omega \in \mathbb{R}^3$ ; linear velocities  $v \in \mathbb{R}^3$ ; linear velocity commands  $v_{cmd} \in \mathbb{R}^3$ ; and previous actions  $a_{t-1} \in \mathbb{R}^{19}$ . The action space delineates the relative variation between the desired joint position and the reference joint position.  $q_{ref}$  represents the default joint position when the bipedal humanoid robot is still standing; therefore, the relationship between action  $a$  and the desired joint position  $q_d$  is  $q_d = a + q_{ref}$ .

Reward functions are broadly divided into two categories: task rewards and auxiliary rewards. The general policy model primarily relies on the following task reward functions:

$$r_v = 0.5 \times e^{-\|v_{xy}^{base} - v_{xy}^{com}\|^2} + e^{-|v_{xy}^{base} - v_{xy}^{com}|} \quad (1)$$

$$r_\omega = 0.5 \times e^{-\|\omega^{base} - \omega^{com}\|^2} + e^{-|\omega^{base} - \omega^{com}|} \quad (2)$$

where  $v_{xy}^{base}$  and  $v_{xy}^{com}$  denote the linear velocity and linear velocity commands, respectively, along the  $x$ -axis and the  $y$ -axis under the coordinate system of the base of the robot.  $\omega^{base}$  and  $\omega^{com}$  denote the angular velocity and the angular velocity command, respectively.

Auxiliary rewards support the development of smooth, fluent gaits, addressing gait optimization and enabling effective transfer from the simulated environment to the physical robot. Limiting joint torque within a reasonable range enhances hardware protection. The torque penalty reward is formulated as:

$$r_{torque} = -\|\tau\|^2 \quad (3)$$

$$r_{lime} = -(\min(\tau_{lim}^{low} - \tau, 0) + \max(\tau - \tau_{lim}^{high}, 0)) \quad (4)$$

where  $\tau$  represents the torque command sent to the joint motor.

Gait regularization was applied to constrain conditions such as adduction, abduction, inner eight, and outer eight in both legs, with the corresponding reward function defined as follows:

$$r_{reg} = e^{-\|q_0\|^2} + e^{-\|q_5\|^2} + e^{-\|q_1 - q_6\|^2} \quad (5)$$

where,  $q_0$  and  $q_5$  respectively represent the joint position of the hip joint of the robot's left and right legs rotating around the  $Z$ -axis.  $q_1$  and  $q_6$  represent the joint position of the hip joint of the robot's left and right legs rotated about the  $X$ -axis, respectively.

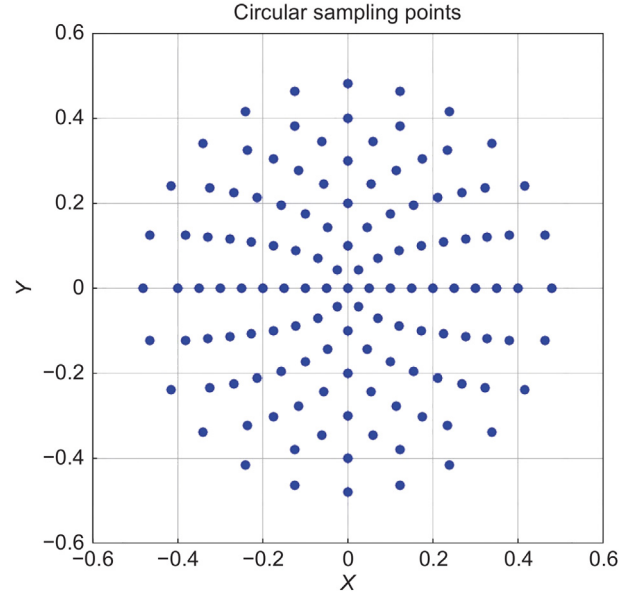
Due to actuator bandwidth limitations in the real environment, control commands with significant jitter can damage the hardware. Therefore, smooth control commands are generated through a motion smoothing constraint. The corresponding reward function is defined as:

$$r_{smooth} = -\|a_t - a_{t-1}\|^2 - \|a_t - 2a_{t-1} + a_{t-2}\|^2 \quad (6)$$

Therefore, the total reward  $r$  of the model was the weighted sum of the above values:

$$r = k_1 \cdot r_v + \dots + k_6 \cdot r_{smooth} \quad (7)$$

All hyperparameters are crucial to the performance of the policy. The highest weights are assigned to task rewards, such as speed tracking rewards  $r_v$ . The initial hyperparameter configuration is first trained on flat terrain, followed by fine-tuning



**Fig. 3.** The visualization of the sampling area of the robot's sole.

**Table 2**  
Parameters of the sampling area.

Radius	Number of sampling point
0	1
0.05	6
0.1	8
0.15	10
0.2	12
0.25	14
0.3	16
0.35	18
0.4	20
0.48	24

based on the robot's locomotion performance in complex terrains. The hyperparameter configuration varies slightly across different terrains.

### 3.2. Expert policy training for different terrains

In addition to proprioceptive information, training the expert policy requires new perceptual data, including the contact state and force direction of the robot's sole.

A series of sampling circles with gradient radii are placed at the center of both soles, with uniformly distributed sampling points, as shown in Table 2. The visualization of the sampling points for one of the robot's feet is shown in Fig. 3, with the sole's coordinate defined as (0,0).

In uneven terrain, it is impractical to require bipedal humanoid robots to traverse at a constant speed, as this limits their ability to navigate such terrains. To enhance the robot's ability to navigate complex indoor terrains, this paper introduces a novel sparse reward to ensure compliance with speed commands. Although sparse rewards increase training difficulty, they enable the robot to acquire unexpected and valuable skills. Many effective gait skills are challenging to achieve through human-designed intensive rewards. Furthermore, to address the challenge of low-efficiency sparse reward training, a combination of intensive and sparse rewards is employed to ensure compliance with speed commands. As training progresses, the weight of the intensive reward gradually decreases, while the weight of

the sparse reward increases. The specific expression is as follows:

$$r_v = (1 - \rho) * e^{-\|v_t - v_d\|^2} + \rho * e^{-\frac{1}{T} \sum_{t=0}^T (v_t - v_d)^2} * I_{\text{episode}} \quad (8)$$

where  $\rho \in (0, 1)$  increases with the number of training iterations,  $v_t$  represents the velocity component of the robot's center of gravity at time  $t$  in the fuselage coordinate system of the X-axis and Y-axis,  $T$  represents the time step of the current episode,  $I_{\text{episode}}$  is the marker of the end of the current episode, and 1 represents the end of the episode.

Due to the differences between the simulation and real environments, it cannot be guaranteed that the robot's methods for ascending and descending stairs in simulation will be successful in reality. For instance, if the contact area between the robot's feet and the steps is small in the simulation, or if collisions with significant force do not affect the success rate of stair traversal, such situations are not reliable in the real environment. Therefore, simulating gait for ascending and descending stairs is crucial to obtain a more robust gait with a higher success rate. This paper proposes a reward function design to constrain the landing area as follows:

$$r_{\text{landing}} = \sum_{i=0}^2 \left( \sum_{k=0}^K |h_i^k - \hat{h}_i^k| \right) * I_{\text{foot}} (F_i > 0) \quad (9)$$

where  $h_i^k$  represents the height in the world coordinate system corresponding to the  $k$ th point on the foot plane of the  $i$ th leg, and  $\hat{h}_i^k$  represents the height of the terrain in the same X-axis and Y-axis coordinates as the  $k$ th point selected in the plane-plane of the foot of the  $i$ th leg.

In addressing the gait optimization challenge for bipedal humanoid robots, relying solely on incrementally increasing rewards can require extensive time to adjust and optimize each reward and its associated weights. Additionally, this tedious optimization may need to be redone if the training task or algorithm hyperparameters change. This paper employs the GAIL algorithm to train a discriminator network that identifies whether a given state-action pair originates from an expert policy. The objective function for optimizing the discriminator network is as follows:

$$L_{\text{disc}}(\vartheta) = -E_{a \sim \pi} [\ln D_{\vartheta}(s, a)] - E_{a \sim \pi^{\text{gait}}} [\ln(1 - D_{\vartheta}(s, a))] \quad (10)$$

where  $\pi^{\text{gait}}$  represents the aesthetic gait trained in Stage 1.

The robot's gait exhibits non-periodic changes in uneven terrains, while the aesthetic gait strategy from Stage 1 is optimized solely for flat surfaces. Therefore, the gait imitation rewards are only applied to flat terrains. The reward function is defined as follows:

$$r_{\text{gait}} = -\ln(1 - D(s, a)) * I_{\text{plane}} \quad (11)$$

where  $I_{\text{plane}}$  represents a flag bit used to determine whether the terrain is flat land within a radius of 0.48 m with the sole of the robot as the center. The primary models trained include the discriminator, actor, and critic networks, with individual policy models developed for stair and threshold terrains.

During training, the weights of these networks are first initialized, and data of a fixed batch size is collected through interactions between the actor network and the simulator. The discriminator network then calculates the gait imitation reward, while the critic network's weights are updated via gradient descent. The objective function  $L_{\text{disc}}(\vartheta)$  is also optimized using gradient descent, with iterative updates to the discriminator network's weights. Training concludes once the specified number of iterations is reached.

Notably, training a single gait model suitable for all terrains remains challenging. For instance, the robot should ideally cross rather than step on thresholds, while on stairs, its foot should contact the horizontal rather than the vertical plane. However,

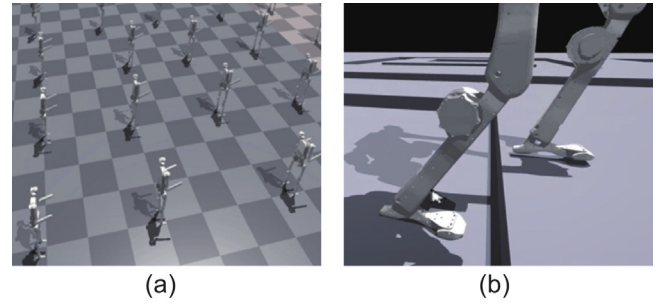


Fig. 4. Description of the policy trained in the simulator.

because threshold and single stair terrains are quite similar, training them jointly can make it difficult for the robot to differentiate between the two, thus hindering the generation of an appropriate gait.

### 3.3. Policy distillation

In this stage, the expert policies for stairway and threshold terrains trained in Stage 2 are distilled into student models through Teacher-Student training via imitation learning. In addition to proprioceptive data, the student model's input includes external environmental data, divided into terrain elevation information and a depth image captured by the trunk-mounted depth camera. The terrain elevation input is a 20 x 21 grid of discrete elevation points, uniformly sampled within a rectangular area: 0.75 m and 0.25 m before and after the fuselage coordinate system origin, 0.5 m from each side, with a 0.05 m interval. Each elevation point has a corresponding validity marker, with 0 indicating an invalid point and 1 indicating a valid point. The topographic elevation and marker point data are concatenated in additional dimensions, resulting in a three-dimensional array with dimensions 21 x 21 x 2. The depth image is obtained directly from the depth camera, with the resolution adjusted through downsampling and cropping, and only single-frame data from the depth image is used.

## 4. Experiments and results

### 4.1. Training process

First, an initial simulation model is constructed within the simulator, as shown in Fig. 4(a), and the robots are trained simultaneously on flat ground to rapidly develop the corresponding general policy. To address complex stair and threshold terrains, a new reward function is introduced to optimize the bipedal humanoid robot's gait in these terrains, with expert policies trained within the simulator, as depicted in Fig. 4(b). A batch size of 4096 x 128 is used. The total number of iterations is 18,400, comprising 15,000 iterations for training and fine-tuning the expert policy and 3,400 iterations for policy distillation.

In the policy training process, we conduct comparative experiments to investigate the impact of fusing terrain elevation and depth image information on model training outcomes. Results show that policies incorporating external environmental information generally achieve enhanced locomotion robustness, often measured by landing constraint reward, throughout the model iteration process, as illustrated in Fig. 5. Although velocity reward serves as the primary reward for the task, emphasis is placed on the landing constraint reward's role in enabling the robot to identify safe landing points.

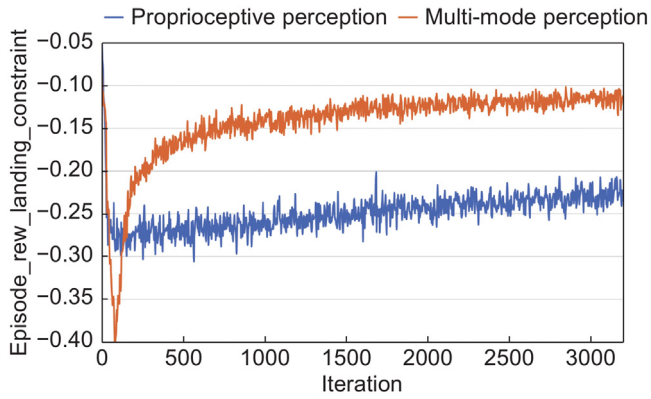


Fig. 5. Result of the comparison of the reward of landing constraint.

#### 4.2. Sim-to-real transfer

The external environmental information utilized in this study comprises terrain elevation data and depth images. To estimate terrain elevation, the elevation data are combined with the pose information of the depth camera and the robot, while the depth image is directly obtained from the depth camera. Deploying this locomotion control algorithm in real environments requires addressing issues of incomplete field of view due to depth camera occlusions, ensuring the external perception of the bipedal humanoid robot is not compromised by the loss of local terrain and elevation data due to the depth camera's field of view angle, measurement range, and occlusions. To address this, terrain elevation information is randomly masked during training. Masking levels are categorized into 11 increments from 0 to 10, where 0 indicates no masking and 10 represents full masking of terrain elevation data. Each level increment raises the masking ratio by 10%.

The entirety of each episode remained unchanged. Additionally, this method enables masking of height information with significant noise or delay through post-processing, based on prior researcher knowledge during real-machine deployment. This approach allows locomotion control with external perception to be simplified to rely solely on proprioceptive information, reducing the influence of noise and delay in the terrain elevation map. However, multimodal perception remains essential. While the simplification to proprioception may eliminate the robot's uncontrollable and potentially hazardous behaviors resulting from external perception noise, it can lead to increased collisions and balance issues, particularly on complex terrains like stairs. More critically, the robot may exhibit destructive behaviors rather than safe navigation. For development, and to maintain alignment between the policy model and PD controller frequencies, the input state is configured consistently with the simulation environment, facilitating the model's adaptation to the real robot system.

#### 4.3. Experiment of balanced standing on flat ground

The robot stands idle when the input velocity command is set to zero. To evaluate its self-balancing performance, a 10 N thrust is applied along the  $x$ -axis. This experiment records the total joint torque curve, illustrating the transition from initial balanced standing to sustained self-balance under applied thrust, as shown in Fig. 6.

In the suspension phase ( $0 \leq t \leq 30$  s), the total joint torque is minimal. In the partial suspension phase ( $30 \leq t \leq 60$  s), the robot is gradually lowered to the ground. The total joint torque shows an increasing trend and gradually stabilizes by 60 s, mainly due

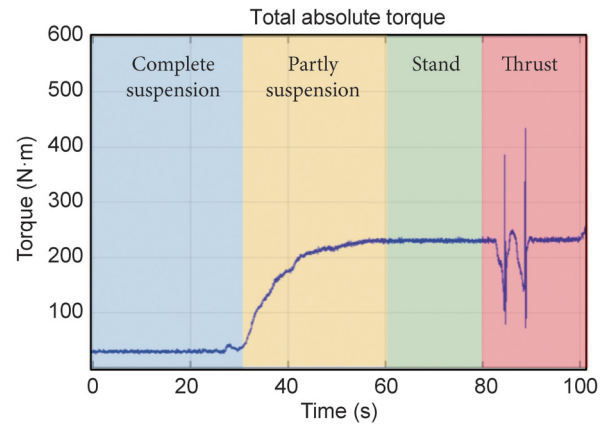


Fig. 6. Description of the robot balanced standing.

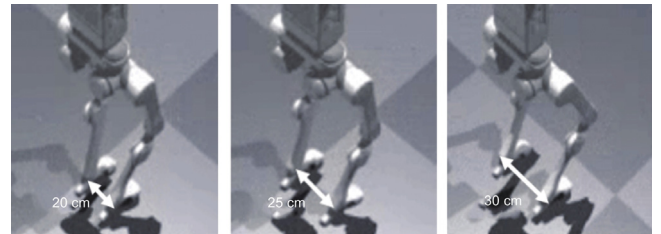


Fig. 7. Results of the walking experiments with different foot spacing.

to the gravitational influence on the leg joints. In the full standing phase ( $60 \leq t \leq 80$  s), the robot stands freely without suspension, maintaining a static state, and the total joint torque remains nearly constant. In the self-balancing test phase ( $80 \leq t \leq 100$  s), an instantaneous thrust is applied artificially. Following several disturbances, the total joint torque returns to the value observed during static standing. These disturbances occur because external force disrupts the robot's initial equilibrium. The robot must swiftly adjust its landing position to ensure the center of mass (COM) projection consistently falls within the support polygon formed by both feet, leading to an instantaneous increase in joint torque. The torque curve demonstrates that the robot equipped with the trained control policy possesses notable self-balancing capabilities.

#### 4.4. Experiment of forward walking on flat ground

In anticipation of deploying the vision sensor on the fuselage in subsequent research, ensuring the fuselage's stability during walking is crucial. A gradient experiment was designed in simulation to examine the influence of foot spacing on walking stability, as shown in Fig. 7.

For walking policy training, the speed was set to 0.4 m/s along the  $x$ -axis, and the trunk's roll-direction offset was recorded for comparative analysis. Results indicate that a foot spacing of 25 cm provides optimal walking stability. A smaller foot spacing increases the likelihood of foot collision during walking, impacting stability, while a larger spacing compromises gait aesthetics.

The IMU records the robot's COM deviation from the  $z$ -axis during walking, and stability is quantitatively assessed using the X offset curve in quaternions. The X offset curve in quaternions was tested at different walking speeds, as shown in Fig. 8. Results reveal that the X offset reaches a minimum at 0.4 m/s, corresponding to a trunk offset of approximately  $1.5^\circ$ , indicating high stability.

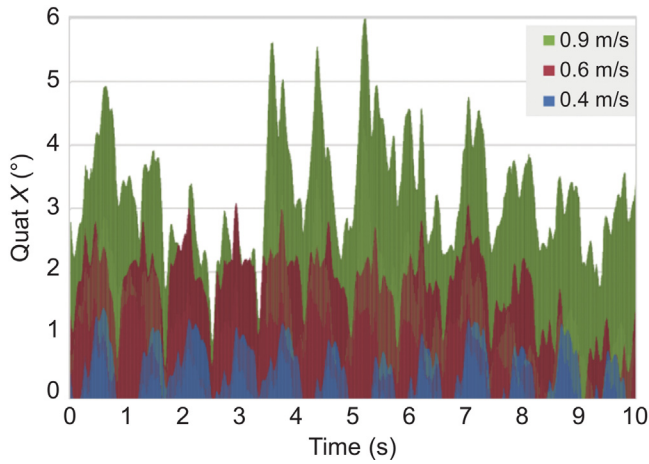


Fig. 8. Description of the trunk offset during walking.

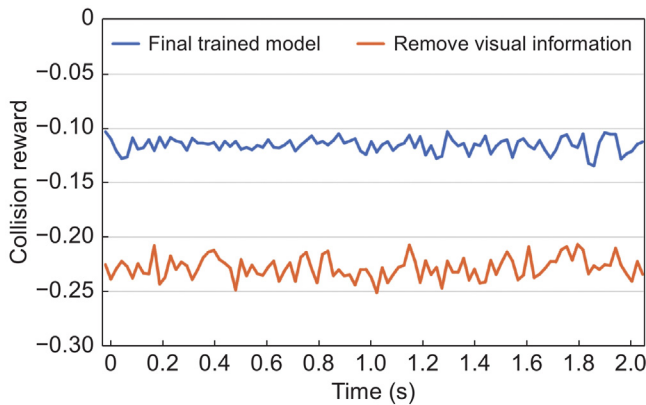


Fig. 9. Result of the ablation experiment for multimodal perception.

#### 4.5. Ablation experiment for multimodal perception

In order to verify the effectiveness of the final trained model, we removed the visual information from the input, so that the robot could not obtain the terrain information, and compared it with the final trained model by setting the stairs terrain in the Issac Sim, as shown in Fig. 9. We chose 100 sets of continuous data for comparison between the two models in the training convergence stage. Each set of data represents the reward score of 20 ms in the dimension of internal collision rate. The lower the score is, the higher the collision rate is. It can be seen that after removing the visual information, the pass-ability of the robot on the stairs terrain is significantly reduced, which is mainly reflected in the improvement of the collision rate. The negative ordinate indicates that the collision reward is negative, with a value no greater than zero.

#### 4.6. Experiment of pass-ability in the terrain of stairs

It is highly valuable for bipedal humanoid robots to identify step-like terrain using external perception data, allowing for anticipatory lifting and crossing. Reducing the collision rate markedly enhances locomotion stability in complex environments. In the simulator, the robot's travel speed was set to 0.4 m/s, with a foot spacing of 25 cm, a single walking duration of 20 s, and continuous single-step terrain traversal. Collisions occur when an object positioned within 2 cm ahead of the toe

Table 3  
Parameters of the stairs.

Parameter	Value
Height	10 cm
Width	50 cm
Interval	1 m

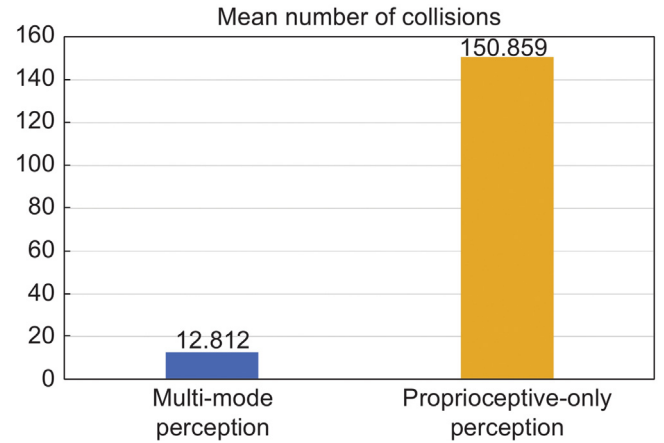


Fig. 10. Result of pass-ability on the terrain of stairs in the simulator.



Fig. 11. Result of pass-ability on the terrain of stairs.

exceeds foot height. Details of the terrain parameter settings are presented in Table 3.

This experiment was repeated 50,000 times, revealing that integrating the external perception policy model reduced the robot's collision rate by 91.51%, as shown in Fig. 10. This result underscores the efficacy of incorporating external environmental perception in motion control.

The policy model was subsequently deployed and validated on a physical robot, with results demonstrating the effectiveness of combining external perception data for enhanced bipedal control.

The bipedal humanoid robot demonstrated the capability to perceive stair terrain information and execute adaptive control autonomously, as depicted in Fig. 11.

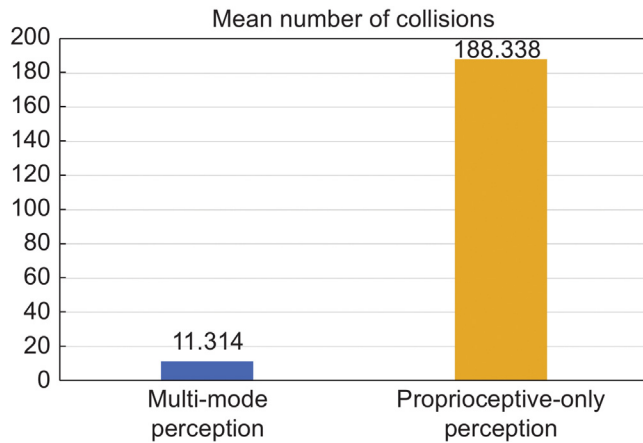
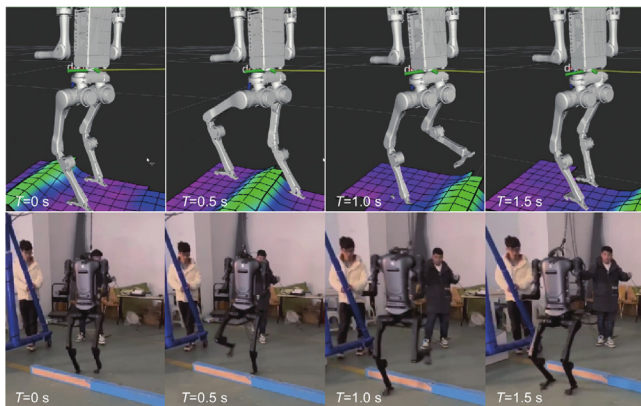
#### 4.7. Experiment of pass-ability in the terrain of thresholds

Thresholds are a key terrain that biped humanoid robots must navigate in domestic environments. In the simulator, the robot's travel speed was set to 0.4 m/s, foot spacing to 25 cm, and a walking duration of 20 s for continuous traversal over discrete threshold terrains. An elevation map centered on the robot was created, sampling height points within a 2 cm x 2 cm area ahead of the robot's toes. A collision was recorded if any sampled height point exceeded the height of the toes. Terrain parameter settings are detailed in Table 4.

This experiment was repeated 50,000 times, demonstrating that the external perception policy model integration reduced the robot's collision rate by 93.99%, as shown in Fig. 12.

**Table 4**  
Parameters of the thresholds.

Parameter	Value
Height	10 cm
Width	10 cm
Interval	1 m

**Fig. 12.** Result of pass-ability on the terrain of thresholds in the simulator.**Fig. 13.** Result of pass-ability on the terrain of thresholds.

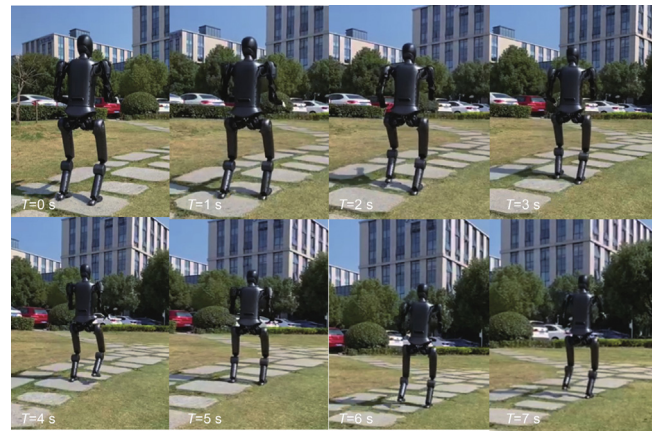
The proposed policy model was effectively implemented and validated on a physical robot. Results emphasized the effectiveness of combining the policy model with external perceptual data for bipedal humanoid robots. As shown in Fig. 13, the robot demonstrated the ability to identify threshold terrain information and perform adaptive control autonomously.

#### 4.8. Experiment of pass-ability in the terrain of mixed terrains

To assess the generalization performance of the policy model in real-world scenarios, we conducted a mixed terrain experiment, where the biped humanoid robot was tasked with traversing diverse terrains. Results indicate that the policy model maintains high pass-ability across varied conditions, demonstrating robust generalization capabilities, as shown in Fig. 14.

## 5. Conclusion and future research

This paper introduces a novel learning-based locomotion control method that integrates multimodal perception for bipedal

**Fig. 14.** Result of pass-ability on the terrain of the mixed terrains.

humanoid robots. By incorporating visual information into the locomotion control solution, we propose a three-stage, multi-objective constrained policy distillation optimization algorithm. Through reinforcement learning, expert policies can be trained on various terrains to ensure gait adaptability to varied terrains. These expert policies are distilled into a student model using policy distillation. Experimental results demonstrate that the collision rate of a control policy that incorporates multimodal perception information is significantly reduced in complex terrains such as stairs and thresholds. Concurrently, it also demonstrates generalization ability in navigating mixed terrains.

Future work will explore integrating large language models (LLMs) with reinforcement learning-based locomotion control for bipedal humanoid robots. Additionally, incorporating embodied intelligence technology will enhance autonomous navigation and operation of robots in complex environments.

#### CRedit authorship contribution statement

**Chao Ji:** Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation. **Diyuan Liu:** Software. **Wei Gao:** Methodology. **Shiwu Zhang:** Writing – review & editing, Methodology, Funding acquisition.

#### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

#### Acknowledgments

This work is supported by the National Natural Science Foundation of China (U21A20119, 62103395, and 51975550).

#### Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.birob.2025.100213>.

#### References

- [1] H. Chai, Y. Li, R. Song, et al., A survey of the development of quadruped robots: Joint configuration, dynamic locomotion control method and mobile manipulation approach, *Biomim. Intell. Robot.* 2 (1) (2022) 100029, <http://dx.doi.org/10.1016/j.birob.2021.100029>.

- [2] M. Shafiee-Ashtiani, A. Yousefi-Koma, M. Shariat-Panahi, Robust bipedal locomotion control based on model predictive control and divergent component of motion, in: Proc. IEEE Int. Conf. Robot. Autom., ICRA, Singapore, 2017, pp. 3505–3510.
- [3] A. Hereid, E.A. Cousineau, C.M. Hubicki, et al., 3D dynamic walking with underactuated humanoid robots: A direct collocation framework for optimizing hybrid zero dynamics, in: Proc. IEEE Int. Conf. Robot. Autom., ICRA, Stockholm, Sweden, 2016, pp. 1447–1454.
- [4] T. Appgar, P. Clary, K. Green, et al., Fast online trajectory optimization for the bipedal robot Cassie, in: Robotics: Sci. and Syst., 2018, p. 14.
- [5] W.C. Martin, A. Wu, H. Geyer, Experimental evaluation of deadbeat running on the ATRIAS biped, IEEE Robot. Autom. Lett. 2 (2) (2017) 1085–1092, <http://dx.doi.org/10.1109/LRA.2017.2658020>.
- [6] K. Green, Y. Godse, J. Dao, et al., Learning spring mass locomotion: Guiding policies with a reduced-order model, IEEE Robot. Autom. Lett. 6 (2) (2021) 3926–3932, <http://dx.doi.org/10.1109/LRA.2021.3066833>.
- [7] J. Reher, W.L. Ma, A.D. Ames, Dynamic walking with compliance on a Cassie bipedal robot, in: Proc. Eur. Control Conf., ECC, Naples, Italy, 2019, pp. 2589–2595.
- [8] K. Sreenath, H.W. Park, I. Poulakakis, et al., Embedding active force control within the compliant hybrid zero dynamics to achieve stable, fast running on MABEL, Int. J. Robot. Res. 32 (3) (2013) 324–345, <http://dx.doi.org/10.1177/0278364912473344>.
- [9] Y. Gong, R. Hartley, X. Da, et al., Feedback control of a Cassie bipedal robot: Walking, standing, and riding a segway, in: Proc. Amer. Control Conf., ACC, Philadelphia, PA, USA, 2019, pp. 4559–4566.
- [10] Z. Li, C. Zhou, N. Tsagarakis, et al., Compliance control for stabilizing the humanoid on the changing slope based on terrain inclination estimation, Auton. Robots 40 (6) (2016) 955–971, <http://dx.doi.org/10.1007/s10514-015-9504-6>.
- [11] J. Chen, K. Xu, X. Ding, Adaptive gait planning for quadruped robot based on center of inertia over rough terrain, Biomim. Intell. Robot. 2 (1) (2022) 100031, <http://dx.doi.org/10.1016/j.birob.2021.100031>.
- [12] X. Da, R. Hartley, J.W. Grizzle, Supervised learning for stabilizing underactuated bipedal robot locomotion, with outdoor experiments on the wave field, in: Proc. IEEE Int. Conf. Robot. Autom., ICRA, Singapore, 2017, pp. 3476–3483.
- [13] G.A. Castillo, B. Weng, A. Hereid, et al., Reinforcement learning meets hybrid zero dynamics: A case study for RABBIT, in: Proc. IEEE Int. Conf. Robot. Autom., ICRA, Montreal, QC, Canada, 2019, pp. 284–290.
- [14] G.A. Castillo, B. Weng, W. Zhang, et al., Hybrid zero dynamics inspired feedback control policy design for 3D bipedal locomotion using reinforcement learning, in: Proc. IEEE Int. Conf. Robot. Autom., ICRA, Paris, France, 2020, pp. 8746–8752.
- [15] G.A. Castillo, B. Weng, W. Zhang, et al., Robust feedback motion policy design using reinforcement learning on a 3D digit bipedal robot, in: Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst., IROS, Prague, Czech Republic, 2021, pp. 5136–5143.
- [16] J. Siekmann, Y. Godse, A. Fern, J. Hurst, Sim-to-real learning of all common bipedal gaits via periodic reward composition, in: Proc. IEEE Int. Conf. Robot. Autom., ICRA, Xi'an, China, 2021, pp. 7309–7315.
- [17] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, 2017, pp. 1–15, <http://dx.doi.org/10.48550/arXiv.1707.06347>, arXiv e-prints, arXiv:1707.06347.
- [18] X.B. Peng, M. Andrychowicz, W. Zaremba, et al., Sim-to-real transfer of robotic control with dynamics randomization, in: Proc. IEEE Int. Conf. Robot. Autom., ICRA, Brisbane, QLD, Australia, 2018, pp. 3803–3810.
- [19] A. Agrawal, S. Chen, A. Rai, K. Sreenath, Vision-aided dynamic quadrupedal locomotion on discrete terrain using motion libraries, in: Proc. IEEE Int. Conf. Robot. Autom., ICRA, 2022, pp. 4708–4714.
- [20] S. Gangapurwala, M. Geisert, R. Orsolino, M. Fallon, I. Havoutis, RLOC: Terrain-aware legged locomotion using reinforcement learning and optimal control, IEEE Trans. Robot. 38 (5) (2022) 2908–2927, [Online]. Available: <http://arxiv.org/abs/2012.03094>.
- [21] W. Yu, D. Jain, A. Escontrela, A. Iscen, Visual-locomotion: Learning to walk on complex terrains with vision, in: 5th Conf. Robot Learn., CoRL, 2021, pp. 1–12.
- [22] R. Yang, M. Zhang, N. Hansen, H. Xu, X. Wang, Learning vision-guided quadrupedal locomotion end-to-end with cross-modal transformers, 2021, CoRR, abs/2107.03996.
- [23] H. Duan, et al., Learning vision-based bipedal locomotion for challenging terrain, in: Proc. IEEE Int. Conf. Robot. Autom., ICRA, 2024, pp. 56–62.
- [24] C. Ji, D. Liu, W. Gao, S. Zhang, Blind walking balance control and disturbance rejection of the bipedal humanoid robot Xiao-Man via reinforcement learning, in: Proc. IEEE Int. Conf. Robot. Biomimetics, ROBIO, Koh Samui, Thailand, 2023, pp. 1–7, <http://dx.doi.org/10.1109/ROBIO58561.2023.10354629>.
- [25] C. Yang, K. Yuan, S. Heng, et al., Learning natural locomotion behaviors for humanoid robots using human bias, 2020, <http://dx.doi.org/10.1109/LRA.2020.2972879>.