CTBC: Contact-Triggered Blind Climbing for Wheeled Bipedal Robots with Instruction Learning and Reinforcement Learning

Rankun Li[*,1,2], Hao Wang[*,2], Qi Li[*,2], Zhuo Han[1,2], Yifei Chu[2], Linqi Ye[†,1], Wende Xie[†,2], and Wenlong Liao[2]

***Abstract*—In recent years, wheeled bipedal robots have garnered significant attention due to their exceptional mobility on flat terrain. However, while stair climbing has been achieved in prior studies, these existing methods often suffer from a severe lack of versatility, making them difficult to adapt to varying hardware specifications or diverse complex terrains. To overcome these limitations, we propose a generalized Contact-Triggered Blind Climbing (CTBC) framework. Upon detecting wheel-obstacle contact, the framework triggers a leg-lifting motion integrated with a strongly-guided feedforward trajectory. This allows the robot to rapidly acquire agile climbing skills, significantly enhancing its capability to traverse unstructured environments. Distinct from previous approaches, CTBC demonstrates superior robustness and adaptability, having been validated across multiple wheeled bipedal platforms with different wheel radii and tire materials. Real-world experiments demonstrate that, relying solely on proprioceptive feedback, the proposed framework enables robots to achieve reliable and continuous climbing over obstacles well beyond their wheel radius. Project page: https://ctbc-for-wheeled-bipedal-robots.github.io/**

***Index Terms*—Contact-Based Control, Instruction Learning, Reinforcement Learning, Wheeled Bipedal Robots.**

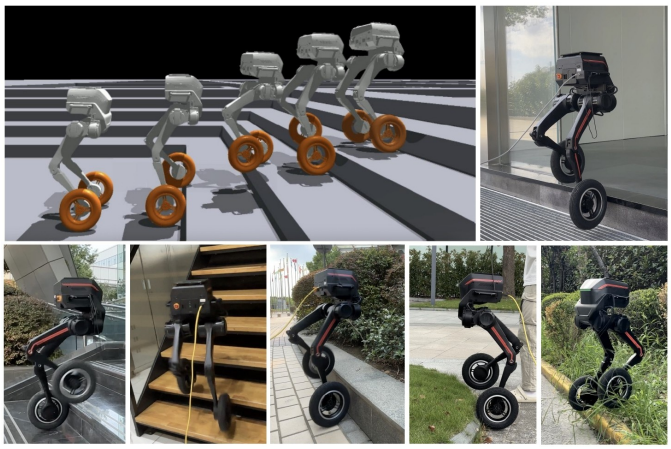

Fig. 1. We have developed a contact-triggered blind climbing control policy that works for wheeled-legged robots of various tire sizes, enabling them to conquer a variety of challenging terrains.

I. Introduction

TRADITIONAL legged robots have demonstrated remarkable agility and adaptability on complex terrain, yet their locomotion efficiency and speed remain comparatively low [1], making it difficult to satisfy the demands of rapid mobility. Wheeled-legged robots, benefiting from their high energy efficiency over long distances and superior travel speed [2], have been extensively studied and deployed across various domains. Nevertheless, when confronted with challenging environments such as staircases or uneven surfaces, the inherent limitations of wheeled robots become evident, as they lack the flexibility required to surmount obstacles effectively.

For wheeled-legged robots, tire dimensions exert a decisive influence on the feasibility of stair-climbing. Larger wheels confer a clear advantage; for instance, Simon Chamorro et al. [3] demonstrated that the Ascento robot can ascend a 15 cm stair with wheels of 25 cm radius, but only with deflated tires and limited to climbing a single step.

To overcome the pain points of wheeled-legged robots in complex terrain, we draw inspiration from the contact-triggered reflexes observed in human gait and propose a contact-triggered control framework that synergizes feedforward instruction learning with reinforcement learning. By leveraging privileged information within an asymmetric actor–critic architecture, we develop a biomimetic gait reflex that enables wheeled-biped robots to ascend stairs with ease. In contrast, our method is suitable for wheeled-biped robots with arbitrary wheel diameters and tire types, and has been successfully deployed on robots with 11 cm rubber solid tires and 12.7 cm pneumatic tires. The key contributions of this work are summarized as follows:

1) **Contact-Triggered RL Task Framework.** We propose a contact-triggered reinforcement learning task framework applicable to wheeled-legged robots of various tire sizes, including the small-tire robot *LimX Dynamics Tron1*, to achieve stair-climbing strategies.
2) **Feedforward Trajectory Instruction Learning.** By combining instruction learning with reinforcement learning, feedforward trajectories are utilized to teach the robot when to lift its legs appropriately. This approach avoids unnecessary exploration and significantly improves learning efficiency.
3) **A Universal Framework for Various Wheel Sizes.** This method enables the *LimX Dynamics Tron1* robot to continuously climb 20 cm stairs (Fig. 1), far exceeding its tire radius. It also enables the *Cowarobot R0* heavy-duty wheeled-biped robot to continuously climb 7.5 cm stairs. The strategy supports both fast movement on flat ground and motion on complex terrain, enhancing the practical adaptability of the robot.

* Indicates Equal Contribution.
† Indicates Corresponding Author.
[1] The School of Future Technology, Shanghai University, 200444 Shanghai, China. rankunli@shu.edu.cn, yelinqi@shu.edu.cn
[2] COWAROBOT Co. Ltd., China.

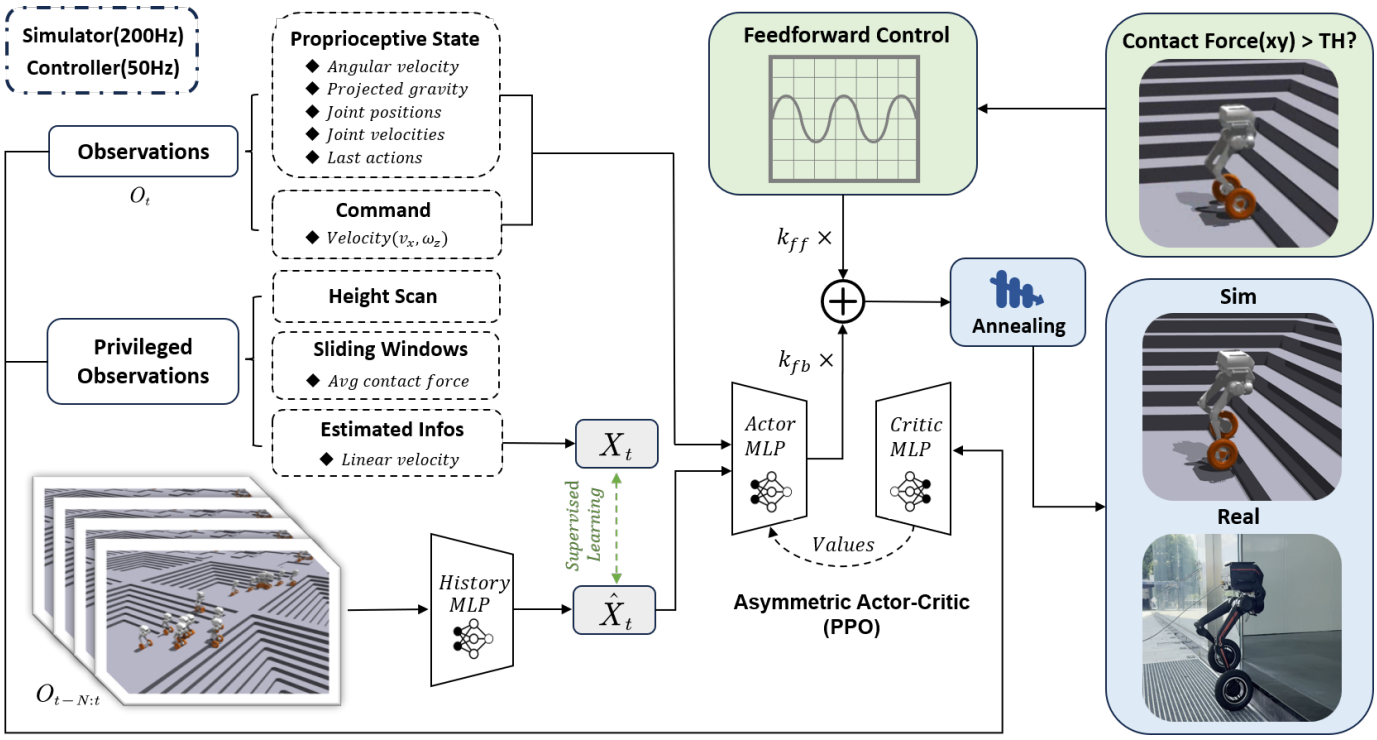

Fig. 2. Overview of our universal contact-triggered blind stair-climbing framework. The overall framework is mainly composed of a state estimator and an asymmetric actor-critic network. For elastic tires, we introduce a contact force sliding window to simulate more realistic contact. When the contact force in the xy direction of the foot exceeds a threshold, it triggers the designed feedforward reference trajectory to guide the robot to lift its leg. After annealing the feedforward trajectory, the method can be zero-shot transferred to the physical robot.

II. RELATED WORK

A. RL-Based Legged Locomotion

In the realm of legged locomotion, model-based optimization techniques such as Model Predictive Control (MPC) and trajectory optimization have long been widely adopted [4]. Nevertheless, the reliance of these methods on accurate and intricate dynamic models often constrains their robustness and generalization. Recently, model-free reinforcement learning has risen to prominence, offering an end-to-end learning paradigm that simultaneously strengthens robustness and markedly enhances generalization across diverse motion-control tasks, positioning itself as a powerful alternative to traditional approaches.

In 2019, Hwangbo et al. [5] introduced reinforcement learning to legged robot control by proposing a policy conditioned on desired velocity that outputs joint-position targets, augmented with an actuator network to enhance motor modeling accuracy. Building directly on this seminal framework, Lee et al. [6] developed a teacher–student architecture that enables robust traversal of hills, steps, and other challenging terrains using only on-board proprioception. Siekmann et al. [7] further demonstrated that an RL policy trained solely on proprioception can drive the bipedal robot Cassie to blindly ascend real-world stairs. Leveraging massively GPU-parallel simulation in Isaac Gym, Rudin et al. [8] trained a locomotion policy with Proximal Policy Optimization (PPO) in just 20 minutes that transfers zero-shot to real hardware. Extending this promising work, Rudin et al. [9] introduced a task formulation based on positional goals: within a strict time limit, the robot must reach a target location, autonomously planning both path and motion to overcome obstacles and complete navigation without additional motion priors.

B. Wheeled-Legged Locomotion on Rough Terrain

In complex-terrain locomotion, conventional model-based methods often hinge on either simple heuristics that dictate when to walk or when to drive [10], or on fixed, pre-defined gait sequences [11]. Most policies for legged robots still embed hand-engineered gait patterns [12], [13] or biologically inspired motion primitives [6], [14].

Bjelonic et al. [15] introduced an online gait generator driven by leg availability: when the availability score of any leg drops below a threshold, the leg is automatically switched to swing while the others continue to drive or support; a single MPC parameter set suffices for all gaits, eliminating manual cost-weight tuning. Klemm et al. [16] leveraged non-smooth trajectory optimization to co-solve global motion planning and contact switching for stairs, steps and jumps in one pass, creating a closed perception–control loop and demonstrating continuous stair climbing on the Ascento wheeled-leg platform. Lee et al. [17] trained a quadrupedal wheeled robot with RL to switch on-the-fly between high-speed wheel driving and legged obstacle clearance in response to commands and terrain, enabling robust obstacle traversal. Lee et al. [18] further proposed a fully-integrated end-to-end framework that fuses model-free RL, privileged learning and hierarchical control, allowing seamless transitions between walking and driving for

tasks such as table jumping and stair climbing. Chamorro et al. [3] demonstrated that a blind RL policy, operating without vision or localization, allows the Ascento robot to climb 15 cm stairs by relying solely on positional objectives, binary terrain flags, and an asymmetric actor-critic architecture. However, to the best of our knowledge, while the Ascento robot possesses a wheel diameter of 25 cm, both its simulation and hardware experiments only demonstrate the traversal of a single-step stair and lack the capability to ascend stairs with narrow treads.

Currently, a universal obstacle-traversal framework for bipedal wheeled robots with arbitrary wheel sizes remains elusive. In particular, when a robot is required to surmount high steps exceeding its wheel radius without any additional exteroceptive sensing, existing methods often struggle with performance and training convergence.

III. METHODOLOGY

As illustrated in Fig. 2, our universal contact-triggered blind climbing framework is depicted. The following sections systematically detail the training environment, reinforcement-learning task formulation, training pipeline, design of the contact-triggered mechanism, integration of feedforward instruction learning, and the sim-to-real transfer strategy with concrete deployment specifics.

A. Learning Environment

1) Simulator: We select Isaac Gym [19] as our training platform because it is specifically designed for reinforcement-learning applications and is equipped with a GPU-accelerated architecture that dramatically increases agent-training speed thanks to its high degree of parallelism. In addition, Isaac Gym supports domain randomization techniques [20], which improve the robustness of reinforcement-learning agents by introducing environmental variations during training and thus facilitate the transfer of policies to the real world. To validate our sim-to-real pipeline, we also leverage MuJoCo [21] for cross-validation of the trained policies [22]. MuJoCo is renowned for its high-fidelity simulation and is widely used to verify reinforcement-learning policies; a policy that successfully deploys in MuJoCo is generally expected to transfer seamlessly to the physical environment.

2) Learning Algorithm: We adopt PPO with an asymmetric actor–critic architecture [23]. This variant of the standard actor–critic framework employs separate networks for the actor and the critic, permitting independent updates [24], [25]. Building upon this, we further introduce a Multi-Layer Perceptron (MLP) encoder for state estimation [26]. All training and experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU with 24 GB of VRAM.

3) Terrains: Our environment is structured as an 8 m × 8 m terrain divided into 10 columns: one column of smooth slope, one of rough slope, six columns of stairs, and two columns of discrete obstacles, as shown in Fig. 3. To progressively increase the curriculum difficulty, the terrain is further split into 10 rows. Generally, the greater the diversity of obstacles encountered during training, the more robust the resulting policy becomes.

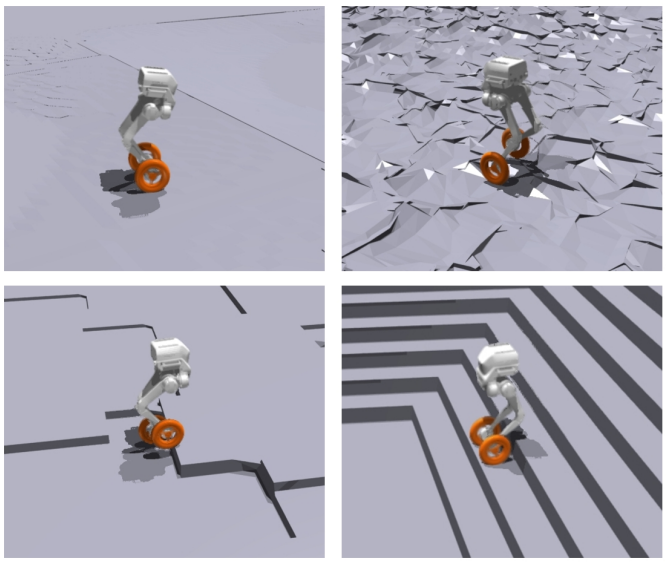

Fig. 3. Terrain type. From left to right and top to bottom: smooth slope, rough slope, discrete obstacles, and stairs.

TABLE I
OBSERVATION & PRIVILEGED INFORMATION

Symbol	Description	Units	Coeff.	Size	Noise (%)
Observation (Actor / Critic)					
$\vec{\theta}$	Angular velocity	rad/s	1.0	3	±20
$\vec{\gamma}$	Projected gravity	–	1.0	3	±5
$\vec{q}$	Joint positions	rad	1.0	6	±1
$\vec{\dot{q}}$	Joint velocities	rad/s	0.05	8	±50
$a_{\vec{last}}$	Last actions	rad & rad/s	1.0	8	0
Privileged Information (Critic only)					
v_x	Linear velocity	m/s	2.0	3	–
μ_{contact}	Avg contact forces	N	1.0	6	–
h_{height}	Height scan	m	5.0	77	–

B. Task Formulation

1) State: We adopt an asymmetric actor–critic architecture, so we partition the state into two parts: (i) Observations, which are accessible to both the Actor and the Critic, and (ii) Privileged Information, which is revealed to the Critic only during training. As summarized in Table I, we explicitly list the observations available to the actor during both training and deployment, as well as the privileged information reserved for the critic at training time. In particular, the term last actions denotes a composite action vector obtained by a weighted fusion of the raw actions directly output by the network and the actions derived from feedforward trajectory.

2) Actions: In this study, the robot's action space has a dimension of 8. These actions correspond to the robot's individual joints, including both leg and wheel joints. For the leg joints, the action commands are directly used as target positions for low-level proportional–derivative (PD) controllers, i.e., the controllers strive to drive the joints to these preset positions. For the wheel joints, the action vector represents target angular velocities; in other words, the joint motors operate in velocity-control mode, aiming to reach the specified

TABLE II
REWARD TERMS AND CLASSIFICATIONS

Reward	Formula	Coeff.
Task Rewards		
Lin. vel tracking x	$\exp\left(-20\left(v_{\text{cmd},x}-v_{\text{base},x}\right)^2\right)$	1.2
Lin. vel tracking y	$\exp\left(-20\left(v_{\text{cmd},y}-v_{\text{base},y}\right)^2\right)$	1.0
Lin. vel tracking x pb	$\frac{\Delta\phi_x}{\Delta t}$	1.0
Lin. vel tracking y pb	$\frac{\Delta\phi_y}{\Delta t}$	0.8
Ang. vel tracking	$\exp\left(-20\left\|\omega_{\text{cmd}}-\omega_{\text{base}}\right\|\right)$	1.0
Ang. vel tracking pb	$\frac{\Delta\phi_\omega}{\Delta t}$	0.5
Tracking target pos	$\exp\left(-2\\|q-q_{\text{target}}\\|\right)-0.2\\|q-q_{\text{target}}\\|$	0.8
Feet air time	$\sum_i \min(t_{\text{air},i},0.5)\,\mathbb{I}_{\text{first contact},i}$	2.0
Feet contact number	$\sum_i\left[\mathbb{I}_{\text{contact}_i=\text{stance}_i}-1.3\,\mathbb{I}_{\text{contact}_i\neq\text{stance}_i}\right]$	2.0
Feet clearance	$\sum_i \mathbf{1}_{\text{swing},i}\cdot\mathbf{1}_{h_{\min}<h_i<h_{\max}}$	2.0
Style Rewards		
Nominal foot position	$\frac{1}{N}\sum_i\exp\left[-\left(\frac{(z_i-z_{\text{nom}})^2}{\sigma_z^2}+\frac{\\|\mathbf{v}_{\text{cmd}}\\|^2}{\sigma_v^2}\right)\right]$	1.0
Default pose	$\sum_j\left\|q_j-q_{j,\text{default}}\right\|$	-1.0
Feet distance	$\max(0,d_{\min}-d)+\max(0,d-d_{\max})$	-10.0
Wheel zero velocity	$\exp\left(-\sum_{j\in\{3,7\}}\mathbf{1}_{\text{swing},j}\,\dot{\theta}_j^2\right)$	0.5
Same foot x position	$\left\|x_0-x_1\right\|$	-2.0
Base height	$\left\|h_{\text{base}}-h_{\text{target}}\right\|$	-20.0
Orientation	$\tilde{g}_x^2+\tilde{g}_y^2$	-12.0
Wheel spin	$\sum_j\max\left(0,0.8\left\|r\dot{\theta}_j\right\|-\\|\mathbf{v}_{\text{foot},j}\\|-0.1\right)$	-5.0
Opposite base vel	$\max\left(0,-\text{sgn}(v_{\text{cmd}})\,v_x\right)$	-40.0
Opposite wheel vel	$\sum_{j\in\{L,R\}}\max\left(0,-\text{sgn}(v_{\text{cmd}})\,\dot{\theta}_j\right)$	-2.0
Regularization Rewards		
Lin vel z	v_z^2	-0.3
Ang vel xy	$\omega_x^2+\omega_y^2$	-0.01
Torques	$\sum_j\tau_j^2$	-1×10^{-5}
Dof acc	$\sum_j\ddot{q}_j^2$	-2.5×10^{-7}
Dof vel	$\sum_j\dot{q}_j^2$	-1×10^{-5}
Action rate	$\sum_j(a_j-a_j^{\text{prev}})^2$	-0.01
Action smooth	$\sum_j(a_j-2a_j^{\text{prev}}+a_j^{\text{prev2}})^2$	-0.005
Collision	$\sum_{i\in\mathscr{I}_{\text{penalised}}}\mathbf{1}_{\\|\mathbf{F}_i\\|>1}$	-50.0
Feet contact forces	$\max\left(0,\bar{F}_z-F_{\max}\right)$	-5.0
Dof pos limits	$-\sum_j\max\left(0,\left\|q_j-q_j^{\text{limit}}\right\|\right)$	-2.0

angular velocities.

3) Rewards: Our task-specific rewards are summarized in Table II. The reward function is composed of three main components:

1) **Task rewards**: including velocity-tracking and foot-lifting terms, which ensure the robot moves at the desired speed and follows the prescribed gait pattern.
2) **Style rewards**: comprising foot-pose and body-pose terms, which encourage the robot to maintain a natural and stable gait.
3) **Regularization rewards**: used to optimize motion smoothness and prevent superfluous joint movements.

Conditioned Reward Formulation: The rewards associated with leg-lifting maneuvers, namely Target Position Tracking, Feet Air Time, Feet Contact Number, and Feet Clearance, are formulated as conditional terms activated solely upon wheel-obstacle contact. This design effectively decouples the locomotion tasks. Specifically, it allows the robot to maintain stable and efficient high-speed wheeled cruising on flat surfaces. Meanwhile, the policy immediately triggers agile leg-lifting behaviors to negotiate obstacles once contact is detected. Consequently, the framework achieves robust obstacle traversal while fully preserving the inherent mobility of the wheeled-bipedal platform.

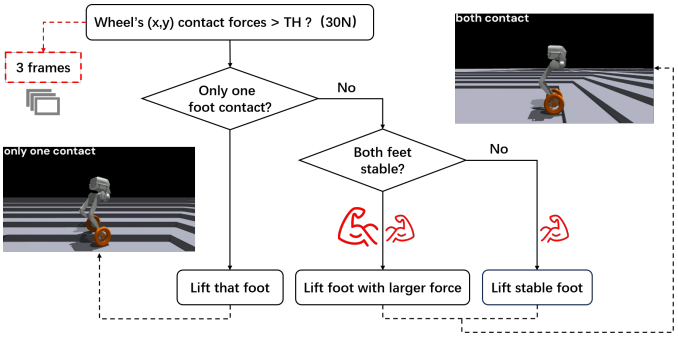

Fig. 4. Logical architecture of the contact-triggered mechanism.

C. Contact-Triggered Mechanism

The methodology is primarily inspired by [17], with the core objective of updating the robot's state and determining the gait phase (stance or swing) based on the horizontal (xy-plane) contact forces measured at the feet (Fig. 4). We extend this framework with three key enhancements:

1) **Threshold-based Triggering:** When the contact force on either wheel exceeds a predefined threshold, the feedforward trajectory is immediately activated. This triggers an initial lift of the contacting leg, followed by a synchronized response from the contralateral leg, resulting in a coordinated alternating ascent.
2) **Sliding-Window Filtering:** To address the contact flickering inherent in rigid-body simulators like Isaac Gym, we adopt a three-frame sliding window to aggregate historical contact states, following the approach in [3]. To formally define the stable contact condition, we employ an indicator-based sliding window mechanism:

$$C_t=\prod_{i=0}^{2}\mathbb{1}(F_{t-i}>\tau) \tag{1}$$

where $C_t\in\{0,1\}$ denotes the binary stable contact state at time t. The term F_{t-i} represents the measured contact force magnitude in the xy-plane at the i-th preceding frame, τ is the predefined force threshold (e.g., 30N), and $\mathbb{1}(\cdot)$ is the indicator function which equals 1 if the condition is satisfied and 0 otherwise. This filtering mechanism effectively suppresses high-frequency noise while ensuring the reliability of genuine contact triggers.
3) **Wheel-Leg Synergetic Integration:** The triggering mechanism is tightly coupled with the rolling wheel dynamics, facilitating seamless transitions between continuous rolling and discrete stepping. This integration optimizes the trade-off between energy efficiency and high-performance obstacle traversal.

The triggering mechanism determines the lifting sequence by continuously monitoring the contact forces on both feet in real time. For each foot, the system stores the latest three

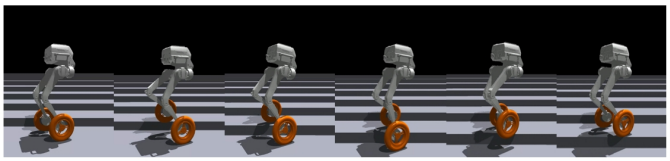

Fig. 5. When either wheel makes contact, only the contacting leg lifts.

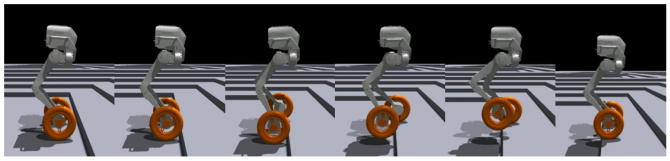

Fig. 6. If both wheels are in contact, the leg with stable contact or the larger contact force is chosen to lift.

frames of force data and designates the contact as stable contact only if all three frames exceed the threshold.

The prioritized leg-lifting decision logic is structured as follows:

1) **Asymmetric Contact:** If contact is detected on a single foot only, the system initiates a unilateral lift of that specific leg while the other maintains support or driving (Fig. 5).
2) **Bilateral Contact:** When both wheels encounter obstacles simultaneously (Fig. 6), the sequence is determined by the reliability and magnitude of the contact signals:
   i) **Stability-First:** If only one foot satisfies the stable contact criteria (as defined by the sliding window), that foot is prioritized for lifting to ensure the maneuver is based on a genuine physical obstacle.
   ii) **Force-Dominant:** If both feet exhibit stable contact, the system selects the leg with the greater instantaneous contact force to lift first. This minimizes the resistance torque against the base and facilitates a more agile ascent.

The mechanism determines the lifting sequence by monitoring bilateral contact forces in real-time. Specifically, the system maintains a buffer of the three most recent force frames for each foot; a "stable contact" is confirmed only if all frames in the buffer consistently exceed the threshold. This logic allows the robot to adapt its locomotion strategy dynamically to ground-truth contact conditions, fostering more natural and stable maneuvers over unstructured terrain.

D. Feedforward Instruction Learning

The concept of feedforward instruction learning, primarily inspired by [27], utilizes a baseline gait motion as a heuristic signal to guide policy exploration. In this work, we adapt this mechanism to bipedal wheeled robots by injecting feedforward trajectories exclusively into the hip-pitch and knee-pitch joints. The composite desired joint action a_t is formulated as:

$$a_t = k_{\text{fb}} a_\pi(t) + k_{\text{ff}}(n) a_{\text{ff}}(t), \tag{2}$$

$$a_{\text{ff}}(t) = A\left(1 - \cos\left(\frac{2\pi}{T}t\right)\right), \quad T = 0.6 \text{ s}, \tag{3}$$

where $a_\pi(t)$ is the neural-network policy output and $a_{\text{ff}}(t)$ is the reference trajectory.

The trajectory amplitude A and the 1:2 ratio between the hip and knee joints are chosen to initiate a basic lifting motion. While the lifting clearance is generally proportional to the trajectory amplitude, we purposefully avoid setting an excessively large value. The core objective of the feedforward signal is to serve as a behavioral guide that "seeds" the lifting primitive rather than dictating a precise geometric path. Consequently, once the robot learns the basic intent of lifting its legs, the specific clearance height is autonomously refined and optimized by the RL policy through environmental interaction.

To ensure that the final policy can autonomously execute maneuvers without external guidance during deployment, we implement a linear difference annealing schedule for the feedforward weight $k_{\text{ff}}(n)$:

$$k_{\text{ff}}(n) = \max\left(0, k_0 - n \cdot \frac{k_0}{N_{\text{ann}}}\right), \tag{4}$$

where n is the current training iteration and N_{ann} denotes the annealing horizon. As training progresses and the policy converges, k_{ff} gradually decreases to zero. At this stage, the feedforward "scaffolding" is completely removed, and the robot's motion is governed entirely by the trained network policy, ensuring a seamless transition from simulation to real-world deployment.

E. Domain Randomization

To achieve zero-shot sim-to-real transfer, we introduce a broad set of randomization factors in simulation to model real-world uncertainties and enhance the policy's generalization ability, as detailed in Table III.

In particular, we apply wide-range randomization for **friction** and **restitution** to compensate for the simplified contact physics in simulation. A broad friction range enables the blind policy to secure reliable traction across diverse real-world surfaces, while varied restitution coefficients effectively model the tires' passive damping and unpredictable energy dissipation during high-impact collisions with obstacle edges.

TABLE III
DOMAIN RANDOMIZATION

Parameter	Range	Unit
Payload mass	$[-0.5,\ 2]$	kg
Center of mass shift	$[-3,\ 3] \times [-2,\ 2] \times [-3,\ 3]$	cm
Kp Factor	$[0.8,\ 1.2]$	N/rad
Ka Factor	$[0.8,\ 1.2]$	N·s/rad
Friction	$[0.2,\ 1.6]$	–
Restitution	$[0.0,\ 1.0]$	–
Inertia	$[0.8,\ 1.2]$	–
Motor torque	$[0.8,\ 1.2]$	N
IMU offset	$[-1.2,\ 1.2]$	–
Default dof pos	$[-0.05,\ 0.05]$	rad
Step delay	$[0,\ 20]$	ms
Push interval	8	s
Push vel (xy)	1.0	m/s

IV. EXPERIMENTS

A. Simulation Experiments

To quantify the contribution of each component in the proposed CTBC method, we conducted controlled ablation experiments across four distinct configurations. To ensure statistical reliability and eliminate the influence of stochasticity, each method was trained across three independent runs using random seeds ($s = 1, 2, 3$). All policies were evaluated after 80,000 iterations under identical simulation environments and hyperparameter setups. The configurations compared are:

- **CTBC (Proposed)**: Integrates the contact-triggered leg-lifting mechanism with feedforward instructions to achieve coordinated climbing.
- **CTBC w/o Feedforward**: Retains the contact-triggered logic but removes the explicit feedforward trajectory, relying solely on indirect lifting rewards for motion generation.
- **CTBC w/o Contact-Trigger**: Maintains the feedforward trajectory but lacks the contact-dependent state transition logic, resulting in repetitive or non-adaptive lifting behaviors.
- **CTBC w/o Both**: Strips away both enhancements to serve as a baseline reinforcement learning policy for performance comparison.

For both the CTBC and CTBC w/o feedforward variants, we observed that an initial longitudinal (fore-aft) leg motion serves as a behavioral heuristic that facilitates the learning of the lifting primitive. Consequently, we adopted a two-stage training curriculum:

1) **Stage I (Exploration):** The policy is trained without constraints on lateral or longitudinal foot placement to encourage the discovery of successful climbing maneuvers.
2) **Stage II (Refinement):** Building upon the pre-trained weights from Stage I, we introduce a "same foot x-position" reward to regularize the fore-aft motion, thereby ensuring a more graceful and stable robot posture.

The comparative performance of the proposed method and its variants is illustrated in Fig. 7. As shown in Fig. 7, the complete CTBC method consistently outperforms all other variants, reaching the maximum terrain level with significantly higher sample efficiency and stability across different random seeds.

- **Impact of Feedforward Trajectory:** Without the feedforward component (CTBC w/o feedforward), the agent lacks effective heuristic guidance during exploration, resulting in a slower progression of terrain levels and a restricted performance ceiling. This underscores that the feedforward signal not only significantly accelerates the policy convergence but is also essential for maximizing the robot's obstacle-clearing capacity.
- **Impact of Contact-Triggered Mechanism:** The variant lacking the contact-triggered mechanism (CTBC w/o contact-trigger) exhibits poor scalability. Failing to dynamically switch between rolling and stepping modes based on real-time contact states, the robot merely executes repetitive lifting motions. This rigid behavior leads to excessive energy consumption and severely limits its adaptability to varied terrains.
- **Baseline Performance:** The baseline policy (CTBC w/o both) fails to surmount even the most fundamental obstacles. This outcome demonstrates that neither component alone is sufficient for mastering high-dimensional stair-climbing tasks, further highlighting the necessity of the synergy between feedforward guidance and contact-triggered state transitions.

Fig. 7. Progressive terrain-level training curves for all ablation variants. The solid lines represent the mean performance across three independent random seeds ($s = 1, 2, 3$), with the shaded areas indicating the standard deviation. Step height increases from 8 cm to 20 cm while width decreases from 50 cm to 28 cm.

TABLE IV
SUCCESS RATE (%) ON STAIRS OF INCREASING HEIGHT

Ablation Experiments	Step height (cm)						
	8	10	12	15	18	20	22
CTBC (Proposed)	**100**	**100**	**100**	**98**	**96**	**86**	**70**
CTBC w/o feedforward	96	96	96	92	80	58	38
CTBC w/o contact-trigger	62	60	56	46	18	2	0
CTBC w/o both	46	34	28	8	4	0	0

The success rates presented in Table IV further quantify these observations. To ensure the robustness of the evaluation, each success rate was statistically computed across 100 parallel environments with domain randomization fully enabled, accounting for uncertainties in friction, mass, and motor characteristics. While all methods perform reasonably well on 8 cm steps, the advantage of CTBC becomes increasingly pronounced as the difficulty escalates. Notably, our method maintains a high success rate of 86% at the 20 cm training limit and remains viable (70%) even at an extreme 22 cm height. In contrast, removing the contact-trigger causes the success rate to plummet to 2% at 20 cm. This failure occurs because the robot cannot dynamically synchronize its lifting action with actual ground contact, leading to catastrophic instability on higher obstacles.

Fig. 8. Snapshots of real-world experiments on the *LimX Dynamics Tron1* robot. Top: Escaping a 10 cm deep hole through rapid contact-triggered leg lifting. Middle and Bottom: Successfully ascending 16 cm and 20 cm high stairs, respectively, demonstrating robust blind locomotion over varied obstacle heights.

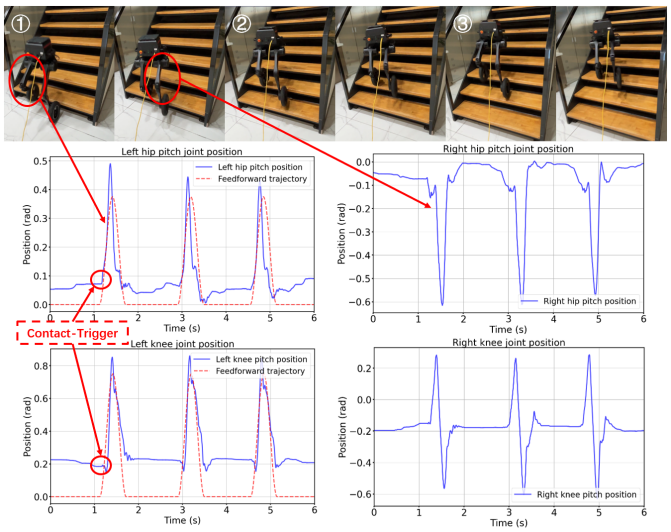

Fig. 9. Continuous blind ascent on 20 cm high open-gap stairs. The results demonstrate the policy's precision in handling extreme surface sparsity. The real-time joint data plots confirm the effectiveness of the guidance: the executed trajectories (solid blue) effectively track the prescribed feedforward instructions (dashed red) upon being activated. Notably, the "Contact-Trigger" points mark the precise moments where the agent transitions from rolling to stepping based on impact, enabling stable clearance of consecutive hollow treads.

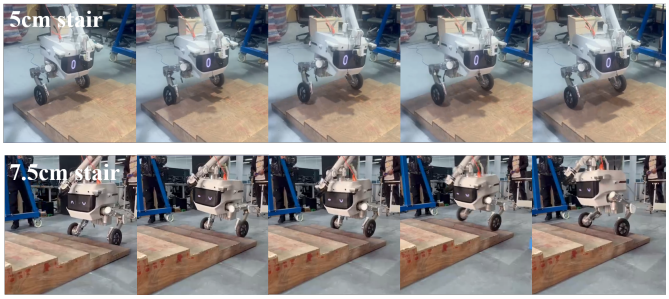

Fig. 10. *Cowarobot R0* ascending 5 cm (top) and 7.5 cm (bottom) steps. Despite the 65 kg mass and low joint velocity (< 2 rad/s), the robot achieves continuous ascent without fine-tuning, demonstrating the zero-shot transferability of CTBC across platforms with distinct scale and dynamics.

B. Real-World Experiments

The CTBC policy, operating at a control frequency of 50 Hz, was deployed on the 8-DoF wheeled-legged robot, *LimX Dynamics Tron1*, without any exteroceptive sensing (e.g., LiDAR or cameras). To evaluate the robustness and sim-to-real transferability of the policy, we conducted experiments in two challenging scenarios: hole escape and stair climbing (Fig. 8).

- **Hole Escape:** Upon a wheel dropping into a 10 cm deep void, the resultant contact force immediately exceeds the trigger threshold. Despite the discretized 50 Hz control loop, the policy responds by executing a rapid lifting motion, enabling the robot to extricate itself smoothly without losing balance.
- **Stair Climbing:** When encountering 16 cm and 20 cm steps, the policy dynamically determines the lead leg based on the transient contact forces at the impact moment. This facilitates a rapid ascent while maintaining postural stability under blind conditions.

As illustrated in Fig. 9, the robot successfully ascends a series of 20 cm continuous open-gap stairs. This scenario represents a heightened level of difficulty due to the extreme discontinuity and the hollow structure of the support surfaces. Even in the absence of a solid riser, our method effectively covers these complex terrains: the contact-triggered mechanism provides the necessary precision to initiate lifting only upon impact with the narrow tread, preventing the wheels from falling into the gaps. This robust performance demonstrates that the learned policy can reliably handle repetitive, high-impact locomotion on sparse terrain.

Furthermore, to verify the cross-platform versatility of our framework, we transferred the CTBC policy to the *Cowarobot R0*, a heavy-duty bipedal wheeled-legged robot integrated with a robotic arm and 8-DoF legs. The *Cowarobot R0* presents a stark contrast to *LimX Dynamics Tron1* (which weighs less than 20 kg) with a total mass of 65 kg and significantly different hardware characteristics. Specifically, it utilizes small-diameter (11 cm) solid rubber tires instead of the pneumatic ones used on *LimX Dynamics Tron1*, and its maximum knee joint velocity is constrained to within 2 rad/s. Despite these substantial discrepancies in physical scale, contact dynamics, and actuator bandwidth, the robot successfully achieved continuous ascent of 5 cm and 7.5 cm steps without any fine-tuning (as shown in Fig. 10). Such successful deployment on a high-inertia platform, characterized by distinct contact dynamics and stringent kinematic constraints, strongly underscores the exceptional robustness and generalization capabilities of the proposed method.

Remarkably, the policy exhibits exceptional generalization: even if the annealing schedule is bypassed and the feedforward signal is abruptly removed, the robot maintains its ability to surmount 20 cm steps. This confirms that the neural network has successfully internalized the lifting maneuver, transitioning from guided exploration to autonomous execution. Furthermore, although the feedforward trajectory was originally designed for a 10 cm lift height, the learned policy successfully

scales this behavior adaptively to clear 20 cm obstacles. For even more challenging terrains, the framework remains highly extensible; by increasing the feedforward amplitude and the reward-constrained lifting range, the robot can acquire even more powerful climbing capabilities through re-training.

V. CONCLUSIONS

In conclusion, this paper presents a contact-triggered, blind locomotion framework for bipedal wheeled-legged robots, successfully bridging the gap between prescribed feedforward guidance and adaptive reinforcement learning. By internalizing motion primitives through a two-stage training scheme, the robot achieves robust stair-climbing and hole-traversal without any exteroceptive sensing. Experimental validations on both the 20 kg *LimX Dynamics Tron1* and the 65 kg *Cowarobot R0* demonstrate the framework's exceptional cross-hardware versatility. Despite significant discrepancies in physical scale, wheel diameters, and tire materials (pneumatic vs. solid rubber), the policy generalizes effectively to obstacles multiple times the height of the initial guidance.

Nevertheless, certain limitations remain: the extended gait in the first training phase restricts the contact frequency of the rear legs, leading to a persistent kinematic bias where the policy favors leading with the front legs. Furthermore, the current system remains purely blind, limiting its efficiency in complex, unstructured environments. Future work will explore symmetry-breaking rewards to eliminate gait bias and integrate this robust blind policy as a low-level reactive controller within a hierarchical perceptive architecture. This synergy between "blind-reflex" and "vision-based planning" will ultimately empower bipedal wheeled-legged robots with fully autonomous navigation and traversal capabilities in unknown and hazardous terrains.

REFERENCES

[1] P. Biswal and P. K. Mohanty, "Development of quadruped walking robots: A review," *Ain Shams Engineering Journal*, vol. 12, no. 2, pp. 2017–2031, 2021.

[2] M. Bjelonic, V. Klemm, J. Lee, and M. Hutter, "A survey of wheeled-legged robots," in *Climbing and walking robots conference*. Springer, 2022, pp. 83–94.

[3] S. Chamorro, V. Klemm, M. d. L. I. Valls, C. Pal, and R. Siegwart, "Reinforcement learning for blind stair climbing with legged and wheeled-legged robots," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 8081–8087.

[4] P. M. Wensing, M. Posa, Y. Hu, A. Escande, N. Mansard, and A. Del Prete, "Optimization-based control for dynamic legged robots," *IEEE Transactions on Robotics*, vol. 40, pp. 43–63, 2023.

[5] J. Hwangbo, J. Lee, A. Dosovitskiy, D. Bellicoso, V. Tsounis, V. Koltun, and M. Hutter, "Learning agile and dynamic motor skills for legged robots," *Science Robotics*, vol. 4, no. 26, p. eaau5872, 2019.

[6] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Science robotics*, vol. 5, no. 47, p. eabc5986, 2020.

[7] J. Siekmann, K. Green, J. Warila, A. Fern, and J. Hurst, "Blind bipedal stair traversal via sim-to-real reinforcement learning," *arXiv preprint arXiv:2105.08328*, 2021.

[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*. PMLR, 2022, pp. 91–100.

[9] N. Rudin, D. Hoeller, M. Bjelonic, and M. Hutter, "Advanced skills by learning locomotion and local navigation end-to-end," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 2497–2503.

[10] M. Bjelonic, P. K. Sankar, C. D. Bellicoso, H. Vallery, and M. Hutter, "Rolling in the deep–hybrid locomotion for wheeled-legged robots using online trajectory optimization," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3626–3633, 2020.

[11] M. Hosseini, D. Rodriguez, and S. Behnke, "State estimation for hybrid locomotion of driving-stepping quadrupeds," in *2022 Sixth IEEE International Conference on Robotic Computing (IRC)*. IEEE, 2022, pp. 103–110.

[12] C. D. Bellicoso, F. Jenelten, C. Gehring, and M. Hutter, "Dynamic locomotion through online nonlinear motion optimization for quadrupedal robots," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2261–2268, 2018.

[13] F. Jenelten, J. Hwangbo, F. Tresoldi, C. D. Bellicoso, and M. Hutter, "Dynamic locomotion on slippery ground," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4170–4176, 2019.

[14] T. Miki, J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning robust perceptive locomotion for quadrupedal robots in the wild," *Science robotics*, vol. 7, no. 62, p. eabk2822, 2022.

[15] M. Bjelonic, R. Grandia, O. Harley, C. Galliard, S. Zimmermann, and M. Hutter, "Whole-body mpc and online gait sequence generation for wheeled-legged robots," in *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2021, pp. 8388–8395.

[16] V. Klemm, Y. de Viragh, D. Rohr, R. Siegwart, and M. Tognon, "Non-smooth trajectory optimization for wheeled balancing robots with contact switches and impacts," *IEEE Transactions on Robotics*, 2023.

[17] J. Lee, M. Bjelonic, and M. Hutter, "Control of wheeled-legged quadrupeds using deep reinforcement learning," in *Climbing and Walking Robots Conference*. Springer, 2022, pp. 119–127.

[18] J. Lee, M. Bjelonic, A. Reske, L. Wellhausen, T. Miki, and M. Hutter, "Learning robust autonomous navigation and locomotion for wheeled-legged robots," *Science Robotics*, vol. 9, no. 89, p. eadi9641, 2024.

[19] 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.

[20] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2017, pp. 23–30.

[21] 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.

[22] 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.

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

[24] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International conference on machine learning*. PmLR, 2016, pp. 1928–1937.

[25] I. Grondman, L. Busoniu, G. A. Lopes, and R. Babuska, "A survey of actor-critic reinforcement learning: Standard and natural policy gradients," *IEEE Transactions on Systems, Man, and Cybernetics, part C (applications and reviews)*, vol. 42, no. 6, pp. 1291–1307, 2012.

[26] G. Ji, J. Mun, H. Kim, and J. Hwangbo, "Concurrent training of a control policy and a state estimator for dynamic and robust legged locomotion," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4630–4637, 2022.

[27] L. Ye, J. Li, Y. Cheng, X. Wang, B. Liang, and Y. Peng, "From knowing to doing: learning diverse motor skills through instruction learning," *arXiv preprint arXiv:2309.09167*, 2023.