

Article

Reinforcement Learning-Based Landing Impact Mitigation and Stabilization Control for Lunar Quadruped Robots Under Complex Operating Conditions

Jianfei Li ^{1,2}, Yeqing Yuan ², Zhiyong Liu ¹ and Shengxin Sun ^{3,*} 

- ¹ State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; lijianfei2020@ia.ac.cn (J.L.); zhiyong.liu@ia.ac.cn (Z.L.)
² Beijing Key Laboratory of Intelligent Space Robotic System Technology and Applications, Beijing Institute of Spacecraft System Engineering, Beijing 100094, China; yqyuan@stu.hit.edu.cn
³ School of Astronautics, Harbin Institute of Technology, Harbin 150001, China
* Correspondence: 20210218@hit.edu.cn

Abstract

Lunar quadruped robots face landing challenges including weak gravity, large mass variations, uncertain sloped terrain, and strict payload acceleration limits, requiring effective impact mitigation and rapid post-landing stabilization. This paper presents a novel end-to-end reinforcement learning-based landing controller with three key novelties: (i) a phase-structured yet implicitly encoded formulation that distinguishes contact preparation, energy dissipation, and stabilization without explicit phase switching; (ii) a terrain-agnostic state and control representation using equivalent support direction construction and contact-gated modulation to decouple normal-tangential dynamics; and (iii) an extremum oriented learning strategy that directly captures peak impact suppression and buffering sufficiency, addressing limitations of cumulative rewards in hybrid, peak-dominated tasks. A hybrid control model for lunar quadruped landing dynamics is established, incorporating variable mass, low impact, and full stroke as key constraints during training. Simulation and full-scale experimental prototypes are built to validate the controller. Simulation results demonstrate robust landing buffering and stability control under varying mass, landing velocity, and slope conditions, with favorable robustness against parameter variations. Experimental verification is conducted under diverse conditions including different masses (200 kg, 250 kg), vertical/horizontal landing velocities (0.8 m/s, 0.2 m/s), and slopes (0°, 8°). The deviation between simulation and experimental results does not exceed 30%, confirming the effectiveness and transferability of the proposed approach.

Keywords: quadruped robots; landing buffering; reinforcement learning; friction modeling; intelligent control



Academic Editor: Giovanni B. Palmerini

Received: 3 March 2026

Revised: 30 March 2026

Accepted: 6 April 2026

Published: 9 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

In low-gravity environments such as the lunar surface, quadruped robots typically achieve deployment and obstacle negotiation with the assistance of jumping/propulsion. Their ground contact phase is characterized by multi-point contact switching, significant uncertainties in contact stiffness and friction, and the coupling difficulty of normal cushioning and tangential anti-sliding is further exacerbated on sloped terrain and soft/deformable ground [1–3]. The landing impact dynamics exhibit strong nonlinearity and transience,

where hazards are often dominated by short-term peak accelerations/loads, thereby inducing the risks of structural damage and attitude instability. Under the conditions of uncertain terrain and soft ground, stability prediction and verification have also highlighted the importance of such peak-dominated failure mechanisms [4–6]. Meanwhile, in low gravity, the energy dissipation mechanisms (cushion stroke utilization, contact energy consumption) are more prone to forming an “energy dissipation–stability” contradiction with the demands for attitude stabilization and anti-sliding, making control strategies that rely solely on empirical parameters or simplified contact assumptions unable to cover all working conditions [7,8].

For impact mitigation and stability recovery, research approaches have evolved from model-based soft landing concepts of energy loss minimization and impedance/cushioning [9], cushioning design of landing mechanisms [10], and multi-physics coupled modeling and analysis of the landing process [8], to dynamic motion generation and tracking combined with optimization and predictive control [11,12]. Against this backdrop, reinforcement learning (RL) has been systematically summarized and widely adopted due to its adaptability to continuous state-action spaces, unknown dynamics, and complex contact interactions [13,14]. On the one hand, neuro-inspired/central pattern generator (CPG) methods and hierarchical learning have improved the energy efficiency of quadruped locomotion and adaptability to complex environments [15,16] and been validated in rough-terrain walking and robust locomotion skill learning [3,17]. On the other hand, deep RL has been applied to the control and gait transition of typical quadruped platforms [18,19], the stability enhancement and adaptability of hexapods on complex terrain [20], and extended to scenarios such as low-gravity jumping/landing, redirection-landing control, and stable jumping in planetary missions [21,22]. In addition, image/meta-RL-based guidance, navigation, and control (GNC) for autonomous landing with safe site selection, as well as mission decision-making methods such as multi-criteria group decision-making criterion reduction, have also driven the learning-oriented trend of the “perception–decision–control” closed loop [23–25].

Despite the progress made in learning paradigms, existing landing-related studies still have three common limitations in their general approaches [26,27]. First, the assumptions about the environment and system are overly idealized: many studies focus on single-leg jumping balance or horizontal ground contact scenarios, or verify strategies under fixed parameters and nominal models, making it difficult to cope with real mission disturbances such as sloped terrain, friction drift, and large-range mass variations [2,19,28,29]. Even though research on trajectory optimization of lunar jumpers, image/meta-learning-based autonomous landing and safe site selection has considered task-level uncertainties, the coupling problems of mass/contact randomness and control in the landing phase are often weakly addressed, resulting in insufficient cross-condition consistency [30]. Second, the semantic modeling of physical coupling and constraints is inadequate: the suppression of normal impact and tangential anti-sliding stability are strongly coupled during sloped landing, while the normal terrain properties and friction are often not directly measurable [31–33]. Although relevant studies have revealed the coupling mechanisms from the perspectives of stability analysis on uncertain terrain, predictive verification on soft ground, bionic lander design, and discrete element method–finite element method–multi-body dynamics (DEM–FEM–MBD) coupled simulation [4–6,8], there is a lack of a unified learnable constraint representation with end-to-end policy learning, leading to a common reliance on friction priors or decoupled modeling [34,35]. Furthermore, studies on spring–damper devices and fuzzy RL for space robot capture, attitude–orbit coupled sliding mode and event-triggered transmission for spacecraft formation, and RL-based robust control for underactuated multi-agent coordination have provided experience in

handling “coupled systems + constraints”, but have not directly solved the optimizable expression of extreme-type safety metrics such as foot-end contact peaks and ground clearance during the landing transient [36–38]. Finally, rewards and objectives fail to reflect the phase characteristics of the landing mission: in the field of locomotion control, the advantages of phase/hierarchical decomposition have been demonstrated through CPG hierarchical control, hexapod hierarchical stability, and multi-agent hierarchical energy management [16,20]. However, in most end-to-end RL methods for landing, the three key phases of contact preparation, energy absorption, and stability recovery are still compressed into a single reward function. This leads to the dilution of critical transient peaks by step-wise cumulative rewards, unclear learning guidance, and limited training efficiency and policy performance [39]. Even though transfer learning and integrated objectives have been introduced in planetary redirection landing and integrated attitude-landing control [21,22,40], more explicit phase semantics and extreme safety characterization are still required to support engineering deployability [41,42].

Based on the above limitations, this paper proposes a learning-based control framework for the landing of lunar quadruped robots on complex sloped terrain with large-range mass uncertainty. Compared with existing reinforcement learning landing methods that often assume flat terrain, fixed mass, or rely on explicit terrain priors, the proposed approach offers three distinctive contributions: (1) a phase-structured yet implicitly encoded control formulation that differentiates contact preparation, energy dissipation, and stabilization without requiring explicit phase switching; (2) a terrain-agnostic state and control representation based on equivalent support direction and contact-gated modulation, which decouples normal–tangential dynamics and reduces reliance on terrain geometry; and (3) an extremum-oriented learning objective that directly captures peak impact suppression and buffering sufficiency, addressing the limitations of standard cumulative rewards in hybrid, peak-dominated tasks. These elements collectively enable robust impact mitigation and stability recovery under uncertain mass, varying impact velocities, and sloped terrain conditions.

The remainder of this paper is organized as follows. The lunar sloped-terrain landing problem is formulated in Section 2 as a hybrid control task, and the limitations of naive end-to-end reinforcement learning are discussed. The phase-structured landing framework and the implicit phase encoding strategy are presented in Section 3. The terrain-aware state and control representation, including equivalent support direction construction and contact-gated modulation, is introduced in Section 4. The learning-based control design, reward function, and success criteria are detailed in Section 5. The training configuration and implementation details are described in Section 6. Simulation results and robustness analysis are reported in Section 7. Hardware experimental validation under multiple test conditions is presented in Section 8. The paper is concluded in Section 9.

2. Problem Formulation and Modeling

2.1. Dynamics Model of a Lunar Quadruped Robot

To address the challenges of mobile exploration across the complex terrain of the lunar surface, this paper proposes and designs a lunar quadruped robot. Adopting a bio-inspired leg configuration, the platform exhibits high-dynamic locomotion, terrain adaptability, and efficient obstacle-negotiation capabilities, making it suitable for long-distance, multi-task scientific exploration under the low-gravity, soft-soil, and rugged topographical conditions of the Moon. The robot employs a symmetric quadruped layout, with each leg featuring three active degrees of freedom (DoFs) that respectively achieve hip abduction/adduction, hip pitch, and knee pitch motion. The overall structure is compact and highly integrated; while ensuring sufficient structural stiffness and load-bearing ca-

capacity, lightweight design and high-performance actuation enable efficient mobility. This section focuses on the design of the robot's leg mechanism, including configuration selection, kinematic layout, materials and lightweighting strategies, and the design of the joint drive units.

As the core components responsible for both locomotion and support, the leg configurations directly influence the robot's locomotion performance, terrain adaptability, and energy efficiency. The legs of the proposed quadruped robot adopt a hybrid parallel–serial mechanism, with the main structure based on a Parallel Five-Bar Mechanism (PFBM) for knee actuation, combined with a serial hip abduction/adduction DoF. This design ensures a large workspace and high load-bearing capacity while enabling a concentrated mass distribution of the drive units, significantly reducing the leg's moment of inertia and benefiting joint dynamic response and walking efficiency.

As illustrated in Figure 1, the robot's thigh and shank are connected via a set of parallel five-bar linkages, where the actual knee motion is jointly controlled by two active actuators located at the hip. Specifically, the thigh link is connected to the shank link through two parallel rocker arms, forming a closed kinematic chain. This layout relocates all knee-driving motors upward near the hip joint, substantially reducing the mass of the shank and foot-end, thereby lowering the inertial load of the swing leg and improving gait frequency and energy efficiency. In addition, all electrical components are positioned above the hip, avoiding direct exposure of motors, encoders, and wiring to harsh lunar conditions such as dust and extreme temperatures, thus enhancing environmental tolerance and system reliability.

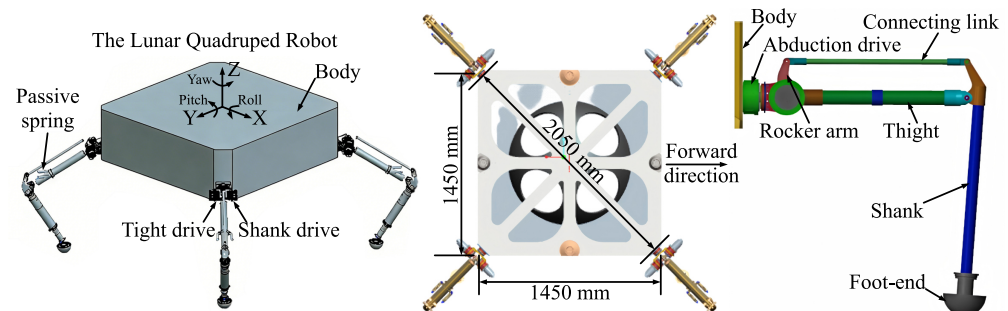


Figure 1. Mechanical Design of a Lunar Quadruped Robot.

Each leg possesses three active DoFs:

- Hip abduction/adduction: Enables leg spreading and retraction in the lateral plane, used to adjust the support polygon, achieve turning, and maintain lateral balance.
- Hip pitch: Cooperates with the knee to generate forward and backward leg swinging, serving as the primary source of propulsion.
- Knee pitch: Implemented via the PFBM, works in conjunction with hip pitch to accomplish foot-end trajectory tracking, ground contact force control, and terrain adaptation.

The motion range of each joint is optimized to balance workspace requirements with structural interference avoidance: the hip abduction/adduction range is $\pm 90^\circ$, the hip pitch range is -90° to 30° , and the equivalent knee pitch range is -180° to 70° . This configuration allows the robot to perform various locomotion modes on the lunar surface, including large-stride walking, in-place turning, and leg lifting for obstacle negotiation.

Leg geometric parameters directly affect kinematic performance and force transmission efficiency. Considering the lunar low-gravity environment and soil mechanical properties, this paper selects equal lengths of 0.5 m for both the thigh and shank links. This equal-length design maximizes the workspace for a given total leg length and ensures that the hip and knee joints share a similar torque distribution during the stance phase, which

is beneficial for actuator load balancing and thermal management. Moreover, equal-length links enhance kinematic symmetry, simplifying gait generation and control algorithms.

In terms of structural materials, the primary load-bearing components of the legs—such as the thigh, shank, and connecting rods—are fabricated from carbon fiber composite materials into hollow thin-walled tubular structures, achieving extreme lightweighting while maintaining adequate stiffness and strength. Specifically, the thigh link weighs 2.1 kg, the shank 1.2 kg, and the foot pad 0.6 kg. Carbon fiber materials offer high specific strength, a low coefficient of thermal expansion, and good vibration damping characteristics, making them well-suited for the large diurnal temperature variations and mechanical vibration environment of the Moon.

To improve energy efficiency during walking and enhance adaptability to ground impacts, a passive linear tension spring (with a stiffness of 6000 N/m and a free length of 0.21 m) is integrated at the knee joint. This spring stores part of the kinetic energy during the stance phase and releases it during the swing phase, helping to reduce joint motor power consumption and achieve tendon-like elastic energy cycling. During continuous walking, the passive spring can smooth joint torque fluctuations, reduce peak actuator loads, and improve overall system efficiency and endurance.

Each active joint is driven by a custom-developed Integrated Drive Unit (IDU), which is compact, lightweight, and capable of high torque output. The IDU comprises an outer-rotor servo motor, a harmonic drive reducer, a torque sensor, an encoder, and an integrated housing, with a total mass of only 3.0 kg and a maximum output torque of 140 N·m. The motor features a high torque density design, with a rated torque of 1.8 N·m and a peak torque of 2.4 N·m; when combined with a harmonic reducer with a transmission ratio of 100, it meets the high-torque and dynamic response requirements of lunar locomotion. The torque sensor provides real-time measurement of joint output torque for force control and ground contact detection, while the high-resolution encoder offers precise joint position feedback to support high-accuracy trajectory tracking.

The robot's body adopts a semi-monocoque structure combining an aluminum alloy frame with carbon fiber skin, measuring 1.45 m (length) $\times$ 1.45 m (width) $\times$ 0.4 m (height). The interior integrates the control system, communication module, energy storage battery, and interfaces for scientific payloads. The total mass of the robot is adjustable within the range of 200–360 kg to accommodate different mission configurations. The four legs are symmetrically arranged at the four corners of the body, forming a stable support base. All hip joint actuators are embedded within the body, further lowering the center of mass and enhancing both static and dynamic stability. In summary, through the synergistic design of parallel actuation layout, equal-length link optimization, carbon fiber lightweighting, passive elasticity integration, and high-performance drive units, the robot's leg mechanism achieves a favorable balance among structural stiffness, motion agility, environmental adaptability, and energy efficiency, laying a solid mechanical foundation for robust mobility across the complex terrain of the lunar surface.

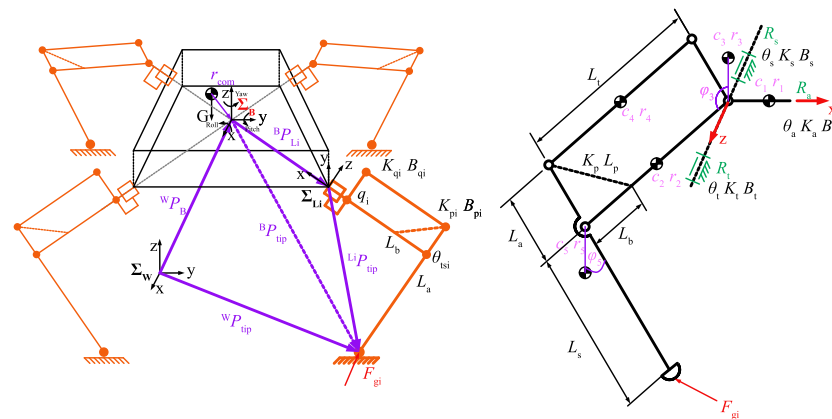
The specific parameters of the lunar quadruped robot are shown in Table 1.

As shown in Figure 2, the body frame Σ_B is defined with its origin at the lander's body center. The x -axis of Σ_B points forward (in the direction of travel) and the z -axis is perpendicular to the upper plane of the body (pointing roughly upward). Each leg i (for $i = 1, \dots, 4$) has an associated leg frame Σ_{L_i} with origin at the leg's hip joint (the intersection point of the axes of the three leg joints). In frame R_{L_i} , the x -axis points from the leg hip toward the body center, and the y -axis is chosen parallel to the z -axis of Σ_B (so all frames are oriented consistently in the vertical direction). The world frame Σ_W is a fixed reference frame (e.g., lunar surface frame) that coincides with Σ_B at the initial time. All coordinate systems are right-handed Cartesian frames.

Table 1. Design of Key Parameters for a Lunar Quadraped Robot.

Name	Unit	Number
Mass	kg	200–360
Thigh link length	m	0.5
Shank link length	m	0.5
Body length	m	1.45
Body width	m	1.45
Body height	m	0.4
Maximum joint torque	N·m	140
Roll joint motion range	°	−90–90
Hip pitch joint motion range	°	−90–30
Knee pitch joint motion range	°	−180–70
Tension spring stiffness	N/m	6000
Tension spring free length	m	0.21
Joint mass	kg	3
Thigh link mass	kg	2.1
Shank link mass	kg	1.2
Foot pad mass	kg	0.6

For each leg, we define the vector of generalized coordinates (active joint angles) as $\mathbf{q}_i = [\theta_{a_i}, \theta_{t_i}, \theta_{s_i}]^T$, where θ_{a_i} is the abduction angle, θ_{t_i} is the thigh angle, and θ_{s_i} is the rocker arm angle (related to the knee joint). In the following, since all legs have identical structure, we derive the leg dynamics for a generic leg and omit the leg index i for simplicity.

**Figure 2.** Dynamic model of the Lunar Quadraped Robot.

The Lunar Quadraped Robot is modeled as a central rigid body supported by four legs, as shown in Figure 2. Let m_b be the total mass of the body, and I_b its inertia tensor about the body frame origin O_b . We denote by $\mathbf{r}_{com}$ the vector from O_b to the body center of mass (COM). For each leg, let O_L be the origin of its leg frame (coincident with the hip joint location), and ${}^B P_{tip}$ the vector from O_b to O_L . The vector from O_b to the leg's tip is denoted ${}^B P_{tip}$, and similarly we denote by ${}^{L_i} P_{tip}$ the vector from the leg frame origin O_L to the tip (thus ${}^B P_{tip} = {}^B P_{Li} + {}^{L_i} P_{tip}$ in coordinates of Σ_B). When the lander contacts the ground, a ground reaction force $\mathbf{F}_g$ will act on the tip of each contacting leg (pointing upward from the ground on the tip). Each leg contains active joint actuators that behave as tunable spring–damper units (by compliance control), characterized by spring stiffness coefficients K_a, K_t, K_s and damping coefficients B_a, B_t, B_s for the abduction, thigh, and rocker joints respectively. In addition, each leg has a passive linear spring (in the parallel five-bar mechanism of the knee) with stiffness K_p .

Using Newton–Euler equations, the translational and rotational equations of motion for the body can be written as:

$$m_b \mathbf{a}_b = m_b \mathbf{g} + \sum_{i=1}^4 \mathbf{F}_{gi} \quad (1)$$

$$\frac{d}{dt}(I_b \boldsymbol{\omega}_b) = \sum_{i=1}^4 ((\mathbf{r}_{com} + \mathbf{r}_{ti}) \times \mathbf{F}_{gi}) \quad (2)$$

where $\mathbf{a}_b$ and $\boldsymbol{\omega}_b$ are the linear acceleration and angular velocity of the body in Σ_B , and $\mathbf{g}$ is the gravitational acceleration vector (in Σ_B coordinates). The first equation is simply $F = ma$ for the body center of mass, stating that the net force (gravity plus all ground contact forces $\mathbf{F}_{gi}$ on the legs) equals the total mass times acceleration. The second equation is the angular momentum balance about the body's center of mass (incorporating the gyroscopic term $\boldsymbol{\omega}_b \times I_b \boldsymbol{\omega}_b$) and states that the sum of moments due to all ground reaction forces about O_b equals the time rate of change in angular momentum of the body. In these equations, $\mathbf{r}_{com}$ and $\mathbf{r}_{ti}$ are expressed in the body frame.

When the lander is in flight (no ground contact), $\mathbf{F}_{gi} = 0$. During landing, as soon as a leg's foot contacts the ground, the corresponding $\mathbf{F}_{gi}$ becomes nonzero and the leg begins to compress, causing $\mathbf{r}_{ti}$ (the tip position relative to the body) to change with time. Thus, to solve the above body equations for the landing dynamics, we first need the leg tip trajectories $\mathbf{r}_{ti}$, which are obtained by solving the leg dynamic model as derived next.

We derive the dynamic equations of a single leg of the Lunar Quadruped Robot using d'Alembert's principle in the form of virtual work. Let the leg's generalized coordinates be $q = [\theta_a, \theta_t, \theta_s]^T$, where θ_a is the hip abduction (or swing) angle, θ_t is the thigh (upper leg) rotation angle, and θ_s is the shank (lower leg) rotation angle. These angles uniquely define the leg configuration. We consider infinitesimal virtual displacements $\delta\theta_a$, $\delta\theta_t$, and $\delta\theta_s$ in each of these coordinates. Correspondingly, each link j (for $j = 1, \dots, 5$) of the leg experiences a virtual linear displacement $\delta\mathbf{r}_{cj}$ of its center of mass (COM) and a virtual angular displacement $\delta\theta_{cj}$ about its COM.

In applying d'Alembert's principle, we introduce the inertial force and inertial torque for each link j :

$$\mathbf{F}_{Ij} = -m_j \mathbf{a}_{cj} \quad \mathbf{T}_{Ij} = -I_j \boldsymbol{\alpha}_{cj} \quad (3)$$

where m_j is the mass of link j , I_j its moment of inertia about the relevant axis, $\mathbf{a}_{cj}$ the linear acceleration of its COM, and $\boldsymbol{\alpha}_{cj}$ its angular acceleration. The negative signs indicate these inertial forces/torques act opposite to the actual accelerations. The leg is also subjected to applied forces/torques: (i) T_a , T_t , T_s are the torques applied by the actuators at the joints corresponding to θ_a , θ_t , θ_s , respectively (positive in the direction of increasing each coordinate); (ii) a passive linear spring (with an inline damper) in the leg's parallelogram mechanism exerts an internal spring force F_p along its length L_p ; and (iii) an external ground reaction force $\mathbf{F}_g = [F_{ex}; F_{ey}; F_{ez}]^T$ acts on the foot (tip). Using the principle of virtual work, the total virtual work done by all these forces and torques (including inertial forces) must sum to zero for any set of virtual displacements $(\delta\theta_a, \delta\theta_t, \delta\theta_s)$. To account for joint friction at the three actuated joints $(\theta_a, \theta_t, \theta_s)$, we introduce friction torques T_{fa} , T_{ft} , T_{fs} at the hip abduction, thigh, and shank joints, respectively. These torques act in opposition to the motion at each joint. Incorporating these into the principle of virtual work adds additional negative work terms for each DOF. Therefore, we can write the equilibrium of virtual work as:

$$\begin{aligned} F_p^T \delta L_p + T_a^T \delta\theta_a + T_t^T \delta\theta_t + T_s^T \delta\theta_s - T_{fa}^T \delta\theta_a - T_{ft}^T \delta\theta_t - T_{fs}^T \delta\theta_s + \dots \\ \dots + \sum_{j=1}^5 (\mathbf{F}_{Ij}^T \delta\mathbf{r}_{cj} + \mathbf{T}_{Ij}^T \delta\theta_{cj}) + \mathbf{F}_g^T \delta\mathbf{r}_{tip} = 0 \end{aligned} \quad (4)$$

Each term in Equation (4) corresponds to the virtual work done by a force or torque: for example, $T_a \delta \theta_a$ is the work by the hip actuator torque through the virtual rotation $\delta \theta_a$, $F_{Ij}^T \delta \mathbf{r}_{cj}$ is the work by the inertial force of link j through its virtual COM displacement, etc. We now proceed to expand each term in detail.

The passive spring (of length L_p) is connected between the thigh and shank via the parallelogram linkage. Let L_a and L_b be the fixed link lengths from the spring's mounting points to the hip and shank joints, respectively, and let p be a constant geometric offset angle in the spring's placement. The length L_p is a function of the joint angles θ_t and θ_s (and p). In fact, from the law of cosines one can derive:

$$L_p = \sqrt{[L_a^2 + L_b^2 - 2L_a L_b c(\theta_s - \theta_t)]} \quad (5)$$

Differentiating this expression, the virtual change in spring length δL_p is obtained as:

$$\delta L_p = \frac{L_a L_b}{L_p} s(\pi + \theta_s - \theta_t) \cdot (-\delta \theta_t + \delta \theta_s) \quad (6)$$

The spring's virtual work contribution in Equation (4) is then

$$F_p \delta L_p = F_p \frac{L_a L_b}{L_p} s(\pi + \theta_s - \theta_t) \cdot (-\delta \theta_t + \delta \theta_s) \quad (7)$$

Next we derive $\delta \mathbf{r}_{cj}$ for each link j . For clarity, let us define shorthand notations $c_x := \cos x$ and $s_x := \sin x$ for any angle x . All coordinates and vectors are expressed in the leg's local coordinate frame Σ_{Li} (attached at the leg's base). The kinematic structure of the leg of Lunar Quadruped Robot is as follows: link 1 is the hip housing (rotating about the body with angle θ_a about a nearly horizontal axis), link 2 is the thigh (rotating by θ_t relative to the hip), link 3 is a rocker arm in the four-bar knee mechanism (rotating by θ_s relative to the hip, in concert with the shank), link 4 is the connecting link of the four-bar (pinned between the thigh and rocker), and link 5 is the shank (which holds the foot). Each link j has a known length or offset to its center of mass, denoted r_j (for $j = 1, \dots, 5$). Additionally, links 3 and 5 have small COM offset angles u_3 and u_5 , respectively (accounting for the fact that their centers of mass are not located exactly along the geometric link line). Links 1, 2, 4 have COMs on their symmetry axis so any offset angle for them is zero or neglected. Using the leg geometry, we can write the position vectors of each link's COM in the R_L frame as functions of $\theta_a, \theta_t, \theta_s$. Differentiating those will give the virtual displacements $\delta \mathbf{r}_{cj}$. We list the results here:

$$\delta \mathbf{r}_{c1} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T \quad (8)$$

$$\delta \mathbf{r}_{c2} = \begin{bmatrix} -r_2 s \theta_t \delta \theta_t \\ r_2 c \theta_t c \theta_a \delta \theta_t - r_2 s \theta_t s \theta_a \delta \theta_a \\ r_2 c \theta_t s \theta_a \delta \theta_t + r_2 s \theta_t c \theta_a \delta \theta_a \end{bmatrix} \quad (9)$$

$$\delta \mathbf{r}_{c3} = \begin{bmatrix} -r_3 s(\theta_s - \varphi_3) \delta \theta_s \\ r_3 c(\theta_s - \varphi_3) \delta \theta_s c \theta_a - r_3 s(\theta_s - \varphi_3) \delta \theta_s s \theta_a \\ r_3 c(\theta_s - \varphi_3) \delta \theta_s s \theta_a + r_3 s(\theta_s - \varphi_3) \delta \theta_s c \theta_a \end{bmatrix} \quad (10)$$

$$\delta \mathbf{r}_{c4} = \begin{bmatrix} -L_a s \theta_s \delta \theta_s - r_4 s \theta_t \delta \theta_t \\ (L_a c \theta_s \delta \theta_s + r_4 c \theta_t \delta \theta_t) c \theta_a - (L_a s \theta_s + r_4 s \theta_t) s \theta_a \delta \theta_a \\ (L_a c \theta_s \delta \theta_s + r_4 c \theta_t \delta \theta_t) s \theta_a - (L_a s \theta_s + r_4 s \theta_t) c \theta_a \delta \theta_a \end{bmatrix} \quad (11)$$

$$\delta \mathbf{r}_{c5} = \begin{bmatrix} -L_t s \theta_t \delta \theta_t + r_5 s(\theta_s - \varphi_5) \delta \theta_s \\ [L_t c \theta_t \delta \theta_t - r_5 c(\theta_s - \varphi_5) \delta \theta_s] c \theta_a \\ -[L_t s \theta_t - r_5 s(\theta_s - \varphi_5)] s \theta_a \delta \theta_a \\ [L_t c \theta_t \delta \theta_t - r_5 c(\theta_s - \varphi_5) \delta \theta_s] s \theta_a \\ +[L_t s \theta_t - r_5 s(\theta_s - \varphi_5)] c \theta_a \delta \theta_a \end{bmatrix} \quad (12)$$

We must also account for the virtual rotations $\delta \theta_{cj}$ of each link's COM. Each link's small rotation can be expressed in terms of the virtual changes $\delta \theta_a$, $\delta \theta_t$, $\delta \theta_s$. Importantly, θ_a is a rotation about the leg's x -axis, while θ_t and θ_s are rotations taking place in the plane that has been rotated by θ_a . A small rotation of link j by $\delta \theta_a$ corresponds to an angular displacement $\delta \theta$ about the x -axis (i.e., $\delta \theta_x = \delta \theta_a$, $\delta \theta_y = 0$, $\delta \theta_z = 0$ in R_L coordinates). A small rotation by $\delta \theta_t$ (of those links that depend on θ_t , namely links 2 and 4) corresponds to an angular displacement about an axis perpendicular to the x -axis. When $\theta_a = 0$, the θ_t rotation axis is along the z -axis of R_L ; for a general θ_a , that axis is rotated by θ_a about x , yielding a unit direction vector $(0, -s \theta_a, c \theta_a)$ in the R_L frame. Similarly, a small rotation $\delta \theta_s$ (for links 3 and 5 which depend on θ_s) is about an axis initially along the z -axis (when $\theta_a = 0$) and thus along $(0, -s \theta_a, c \theta_a)$ after rotation by θ_a .

Using these observations, we can write the virtual rotation of each link's COM as a 3-component vector. For example,

$$\begin{cases} \delta \theta_{c1} = [\delta \theta_a & 0 & 0]^T \\ \delta \theta_{c2} = [\delta \theta_a & -s \theta_a \delta \theta_t & c \theta_a \delta \theta_t]^T \\ \delta \theta_{c3} = [\delta \theta_a & -s \theta_a \delta \theta_s & c \theta_a \delta \theta_s]^T \\ \delta \theta_{c4} = \delta \theta_{c2} \\ \delta \theta_{c5} = \delta \theta_{c3} \end{cases} \quad (13)$$

Finally, the virtual displacements of F_p can be given as

$$\delta \mathbf{r}_{\text{tip}} = \begin{bmatrix} -L_t s \theta_t \delta \theta_t + L_s s \theta_s \delta \theta_s \\ (L_t c \theta_t \delta \theta_t - L_s c \theta_s \delta \theta_s) c \theta_a - (L_t s \theta_t - L_s s \theta_s) s \theta_a \delta \theta_a \\ 0(L_t c \theta_t \delta \theta_t - L_s c \theta_s \delta \theta_s) s \theta_a + (L_t s \theta_t - L_s s \theta_s) c \theta_a \delta \theta_a \end{bmatrix} \quad (14)$$

The above equations can be recognized as the dynamic equilibrium conditions (force/moment balance) for the θ_a , θ_t , and θ_s motions, including all inertial and applied effects. While these expanded equations are correct, it is helpful to express them in a more compact form. We can identify in each equation the contributions from inertial forces (which will be associated with acceleration terms $\ddot{\theta}_a, \ddot{\theta}_t, \ddot{\theta}_s$), from velocity-dependent effects (Coriolis and centrifugal terms, associated with $\dot{h}$ products), from gravity and spring forces (potential forces), and from external forces. In fact, the above equations can be written in matrix–vector form as:

$$\mathbf{M}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}}) + \mathbf{G}(\mathbf{q}) = \boldsymbol{\tau} \quad (15)$$

where $\mathbf{M}(\mathbf{q}) \in \mathbb{R}^{3 \times 3}$ is the symmetric mass/inertia matrix of the leg, $\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})$, $\dot{\mathbf{q}} \in \mathbb{R}^3$ is the vector of Coriolis and centrifugal terms, $\mathbf{G}(\mathbf{q}) \in \mathbb{R}^3$ is the gravity and spring force generalized force vector (including weight of links as well as the passive spring force, which acts like an elastic potential), and $\boldsymbol{\tau} \in \mathbb{R}^3$ is the vector of generalized actuator forces (here, the three actuator torques $[T_a, T_t, T_s]^T$). Equation (15) is the standard form of the leg's equations of motion. In the absence of external contact force ($\mathbf{F}_g = \mathbf{0}$), $\boldsymbol{\tau}$ would equal the actual motor torques required to produce the motion $\mathbf{q}(t)$. When an external foot force $\mathbf{F}_g$ is present (such as during landing impact), its effect appears on the right-hand side of Equation (15) as well, typically through an additional term $\mathbf{J}^T(\mathbf{q})\mathbf{F}_g$ (where $\mathbf{J}$ is

the foot Jacobian). In other words, the actuator torques must also counteract the external force's influence.

2.2. Lunar Sloped-Terrain Landing as a Hybrid Control Problem

In this study, the gravitational acceleration is explicitly set to the lunar gravity level, i.e., approximately one-sixth of Earth's gravity:

$$g_{\text{moon}} \approx 1.62\text{--}1.63 \text{ m/s}^2. \quad (16)$$

Unless otherwise specified, all references to gravity in this paper correspond to this lunar-gravity setting. The robot is modeled as a free-floating multibody system with a floating base and twelve actuated leg joints, interacting with the environment through intermittent multi-point contacts at the feet. Let the system state be denoted by

$$\mathbf{x} = (\mathbf{q}, \dot{\mathbf{q}}) \quad (17)$$

where $\mathbf{q}$ includes the base position and orientation as well as joint coordinates, and $\dot{\mathbf{q}}$ denotes the corresponding generalized velocities.

Unlike flat-ground locomotion or quasi-static stance control, the landing process is inherently hybrid, involving both continuous-time dynamics and discrete contact events. During an episode, the system evolves through a sequence of modes characterized by different contact configurations, including free-flight, partial foot contact, full multi-leg support, and possible transient detachment or rebound. Each contact event introduces discontinuities in acceleration and constraint activation, resulting in nonsmooth dynamics.

The presence of a sloped terrain fundamentally alters the structure of contact interactions. Gravity is no longer aligned with the principal support direction, and normal and tangential contact forces become strongly coupled. Under this weak-gravity condition $g = g_{\text{moon}}$, the flight phase is prolonged, and the transition from flight to contact becomes highly sensitive to initial conditions.

From a control perspective, the landing impact mitigation problem thus constitutes a high-dimensional hybrid control problem with strong nonlinearities, mode-dependent constraints, and pronounced transient phenomena concentrated in short time intervals around contact events.

2.3. Control Objectives and Engineering Constraints

The ultimate goal of lunar landing impact mitigation is not merely to bring the robot to rest, but to do so in a manner that satisfies strict engineering safety and stability requirements. Accordingly, the control objectives are defined at both instantaneous and episode levels.

2.3.1. Impact Mitigation and Peak Suppression

From a structural safety standpoint, the primary risk during landing arises from short-duration impact peaks rather than average loads. Let $\mathbf{a}_{\text{lin}}$ and $\boldsymbol{\alpha}$ denote the linear and angular accelerations of the robot base, respectively. To isolate impact-induced effects, the gravity-compensated excess linear acceleration is defined as

$$\mathbf{a}_{\text{ex}} = \mathbf{a}_{\text{lin}} - \mathbf{g}_{\text{moon}} \quad (18)$$

where g_{moon} denotes the constant lunar gravitational acceleration used throughout both the modeling and experimental evaluations. In both simulation and experimental setups, the gravitational acceleration is consistently set to g_{moon} . All gravity-related quantities in the controller design and evaluation metrics are defined with respect to this value,

including excess acceleration computation and the interpretation of landing dynamics under weak-gravity conditions.

The corresponding impact intensities are quantified by the norms

$$a_{\text{ex}} = \|\mathbf{a}_{\text{ex}}\|, \quad \alpha = \|\boldsymbol{\alpha}\| \quad (19)$$

The control objective is to minimize the maximum values of these quantities over the entire landing episode

$$\min \max_{t \in [0, T]} (a_{\text{ex}}(t), \alpha(t)) \quad (20)$$

subject to system dynamics and contact constraints. This extremum-oriented objective reflects the fact that peak overloads dominate failure risk for mechanical structures, joint transmissions, and onboard instruments.

2.3.2. Posture Stability and Slip Suppression

In addition to impact mitigation, the robot must maintain posture stability during and after touchdown. Excessive roll or pitch excursions can lead to rollover, while tangential velocities along the slope may violate friction constraints and induce foot slip. Therefore, posture angles, angular velocities, and tangential motion components must be regulated within admissible bounds throughout the landing process.

2.3.3. Sustained Stability and Effective Buffering

A successful landing requires not only instantaneous satisfaction of stability conditions, but also their persistence over a finite time interval. Moreover, the leg mechanisms must actively engage to dissipate kinetic energy through compression. Superficial strategies that momentarily satisfy velocity or posture thresholds without meaningful energy absorption are unacceptable from an engineering standpoint. Consequently, success is defined by the combination of sustained stability and sufficient utilization of the available buffering stroke.

Together, these requirements impose a set of state, input, and trajectory-level constraints on the control policy, including limits on accelerations, body orientation, ground clearance, joint torques, and contact conditions. These constraints must be respected despite uncertainty in initial velocities and variations in system mass due to payload configuration or propellant consumption.

2.4. Limitations of End-to-End Reinforcement Learning

A natural approach to the above problem is to apply end-to-end reinforcement learning, treating the entire landing process as a black-box Markov decision process. However, the hybrid and extremum-dominated nature of lunar landing introduces several fundamental limitations to naive end-to-end formulations, as shown in Figure 3.

First, different phases of the landing process are governed by distinct physical objectives. During flight and early contact, the primary concern is establishing consistent contact without inducing numerical or physical artifacts, whereas during buffering, peak impact suppression and energy dissipation dominate. In later stages, sustained stability becomes the main criterion. When these heterogeneous objectives are combined into a single undifferentiated reward signal, gradients associated with different phases tend to interfere with each other, weakening learning signals at critical transients such as touchdown impacts.

Second, standard step-wise cumulative rewards are ill-suited for extremum-oriented objectives. Short but hazardous impact peaks may contribute negligibly to the total return when averaged over long episodes, leading the policy to underestimate their importance. As a result, learned behaviors may exhibit opportunistic strategies that shift or concentrate impacts rather than genuinely suppressing them.

Third, hybrid events such as contact establishment and rebound introduce nonsmooth dynamics that violate the assumptions underlying many reinforcement learning algorithms. Without appropriate problem structuring, policies may exploit simulator artifacts near contact transitions or develop unstable behaviors that are not physically meaningful.

Finally, purely end-to-end formulations often struggle to incorporate engineering constraints in a principled manner. Hard termination alone leads to sparse learning signals, while overly aggressive penalties may destabilize training or bias the policy toward overly conservative behaviors.

These limitations motivate a more structured problem formulation that preserves the advantages of end-to-end learning while embedding physical consistency and engineering semantics into the control problem. In the following sections, we introduce a phase-structured modeling framework, terrain-aware state and control representations, and extremum-oriented performance modeling that collectively address the above challenges without resorting to explicit phase switching or multi-policy architectures.

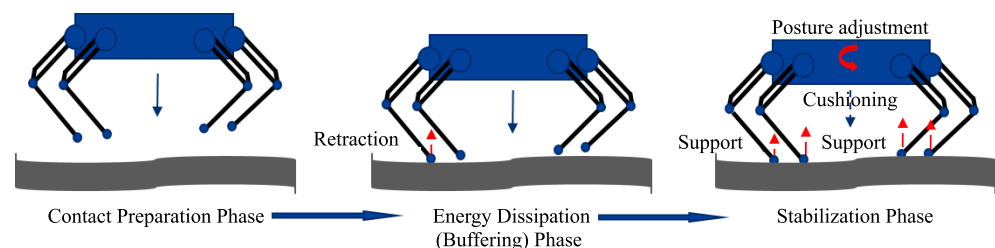


Figure 3. Phase-structured decomposition of the sloped-terrain landing process: contact preparation, energy dissipation (buffering), and stabilization.

3. Phase-Structured Landing Framework

The block diagram of the proposed terrain-aware learning-based control design for sloped-terrain landing is illustrated in Figure 4.

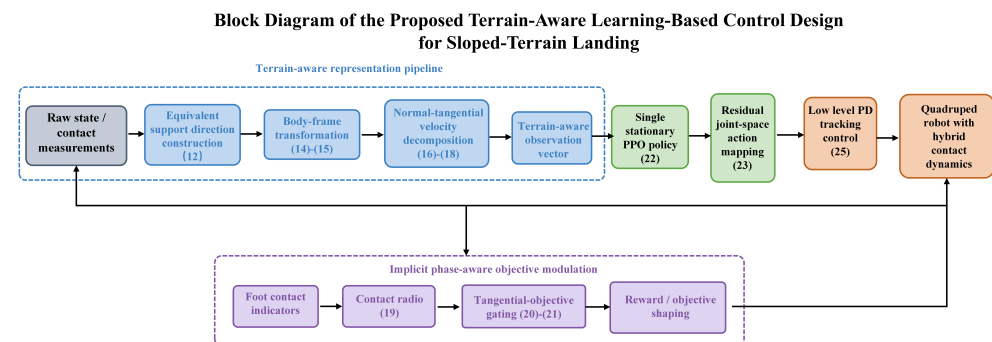


Figure 4. Block diagram of the proposed reinforcement learning-based landing control framework.

3.1. Implicit Phase Decomposition of the Landing Process

The landing impact mitigation process of a quadruped robot on sloped terrain exhibits strongly time-varying dynamics, in which the dominant physical mechanisms and control requirements change substantially over the course of a single episode. Rather than treating landing as a homogeneous control problem, we adopt a phase-structured perspective, in which the overall process is decomposed into a sequence of conceptually distinct phases according to the evolution of contact conditions, energy flow, and stability characteristics.

Importantly, the proposed phase decomposition is introduced for modeling and analysis purposes only. No explicit phase variable, phase classifier, or mode-switching controller is assumed to be available to the policy. Instead, phases are defined implicitly by physical events and state evolution.

From a control and dynamics standpoint, the landing process can be decomposed into three consecutive phases:

1. Contact Preparation Phase;
2. Energy Dissipation (Buffering) Phase;
3. Stabilization Phase.

Let $t \in [0, T]$ denote the episode time horizon. The three phases are not separated by fixed time instants, but rather by state-dependent transitions governed by contact establishment, impact occurrence, and convergence to a quasi-static regime.

The contact preparation phase corresponds to the initial interval during which the robot transitions from flight to consistent multi-leg contact. The primary requirement in this phase is the establishment of a physically consistent contact state, rather than energy dissipation. Let $C(t)$ denote the contact set. A necessary condition is

$$|C(t)| \leq 1 \quad (21)$$

with a preference toward multi-leg support. The dynamics in this phase are dominated by contact activation and geometric consistency, and improper handling may lead to spurious impulsive artifacts that degrade policy learning.

During this phase, the system dynamics are dominated by geometric consistency and constraint activation rather than energy dissipation. The primary concern is to avoid non-physical contact artifacts and numerical inconsistencies that may corrupt subsequent learning signals.

The energy dissipation (buffering) phase begins once effective foot–ground contact is established and significant kinetic energy must be absorbed through leg compression and contact work. This phase is characterized by rapidly varying contact forces and peak-dominated impact responses. The objective can be expressed as

$$E_{\text{kin}}(t) \rightarrow E_{\text{kin}}^{\text{low}}, \quad (22)$$

while minimizing the worst-case impact response

$$\min \max_{t \in T_{\text{buf}}} (a_{\text{ex}}(t), \alpha(t)), \quad (23)$$

where $a_{\text{ex}}(t) = a_{\text{lin}}(t) - g$. This formulation highlights that buffering must occur under peak-load constraints rather than purely minimizing energy. On sloped terrain, tangential velocity must also be regulated to prevent slip, resulting in coupled normal–tangential dynamics.

This phase is characterized by strong nonlinearities, rapidly varying contact forces, and peak-dominated impact responses. Hybrid events such as partial detachment or secondary contact may occur, and the system experiences its largest accelerations during this interval.

Finally, the stabilization phase corresponds to the post-impact regime, in which the system operates near a quasi-static equilibrium. Stability must be maintained over a finite time interval rather than at a single instant. This is enforced through

$$\|\phi(t)\| \leq \bar{\phi}, \quad \|v_t(t)\| \leq \bar{v}_t, \quad \|\omega(t)\| \leq \bar{\omega}. \quad (24)$$

together with a buffering sufficiency condition

$$\Delta h = h_0 - h_{\text{min}}, \quad \Delta h \geq \Delta h_{\text{req}}. \quad (25)$$

Although velocities and accelerations are small, instability may still arise due to accumulated tangential motion, posture drift, or insufficient contact support. Therefore, stability must be evaluated over a finite time window rather than at a single instant.

This phase decomposition clarifies the temporal structure of the landing task and provides a principled basis for defining phase-consistent control objectives, as discussed next.

3.2. Phase-Dominant Control Objectives

Each phase of the landing process is governed by a distinct set of dominant control objectives, reflecting the underlying physical mechanisms and engineering requirements.

3.2.1. Contact Preparation Phase

During the contact preparation phase, the primary objective is contact consistency, rather than impact mitigation or stabilization. Let $\mathcal{C}(t)$ denote the set of feet in contact with the ground at time t . The goal of this phase is to establish a physically meaningful initial contact configuration such that

$$|\mathcal{C}(t)| \geq 1 \quad (26)$$

and preferably multiple contacts, without inducing spurious impulses or artificial slip.

At this stage, minimizing kinetic energy or posture deviation is not yet meaningful. Instead, the dominant requirement is that the initial contact state satisfies geometric and kinematic consistency, ensuring that subsequent impact responses arise from genuine control behavior rather than numerical artifacts. Failure to satisfy this requirement may cause the policy to learn compensatory actions for simulator-induced errors, leading to poor generalization.

3.2.2. Energy Dissipation (Buffering) Phase

Once consistent contact is established, the system enters the buffering phase, during which the dominant objective is to dissipate kinetic energy while suppressing hazardous impact peaks. Let $E_{\text{kin}}(t)$ denote the total kinetic energy of the robot. The buffering process aims to reduce

$$E_{\text{kin}}(t) \rightarrow E_{\text{kin}}^{\text{low}} \quad (27)$$

within the available compression stroke of the legs.

However, due to structural safety considerations, the buffering objective is not simply to minimize energy as quickly as possible. Instead, it is constrained by peak acceleration limits. Let $a_{\text{ex}}(t)$ and $\alpha(t)$ denote the gravity-compensated linear and angular acceleration magnitudes, respectively. The dominant objective of this phase can be expressed as an extremum-oriented criterion

$$\min \max_{t \in \mathcal{T}_{\text{buf}}} (a_{\text{ex}}(t), \alpha(t)) \quad (28)$$

where $\mathcal{T}_{\text{buf}}$ denotes the buffering interval.

On sloped terrain, tangential motion plays a critical role in stability. Excessive tangential velocity may violate friction constraints and induce slip or rollover. Therefore, buffering control must simultaneously suppress impact peaks and attenuate tangential motion, leading to inherently coupled objectives during this phase.

3.2.3. Stabilization Phase

After the majority of kinetic energy has been dissipated, the system transitions into the stabilization phase. In this regime, instantaneous state satisfaction is insufficient to guarantee safe landing. Instead, stability must be maintained over a continuous time interval.

Let $\phi(t)$ denote the body roll–pitch angles, $\mathbf{v}_t(t)$ the tangential velocity component relative to the support plane, and $\omega(t)$ the body angular velocity. Stability is defined through joint satisfaction of bounds,

$$\|\phi(t)\| \leq \bar{\phi}, \quad \|\mathbf{v}_t(t)\| \leq \bar{v}_t, \quad \|\omega(t)\| \leq \bar{\omega} \quad (29)$$

for all t within a finite stability window.

In addition, sufficient utilization of the buffering mechanism must be verified. Let h_0 denote the initial body clearance relative to the slope along the support normal, and $h_{\min}$ the minimum clearance attained during the episode. The effective compression is defined as

$$\Delta h = h_0 - h_{\min} \quad (30)$$

Stabilization is considered meaningful only if

$$\Delta h \geq \Delta h_{\text{req}} \quad (31)$$

where Δh_{req} denotes the minimum required compression stroke. This condition prevents opportunistic strategies that achieve apparent stability without genuine energy dissipation.

3.3. Implicit Phase Encoding Without Policy Switching

While the landing process is naturally phase-structured, explicitly introducing phase indicators or multiple phase-specific policies would significantly increase system complexity and hinder robustness. Instead, this work adopts an implicit phase encoding strategy, preserving a single end-to-end policy while embedding phase awareness through environment design and objective formulation.

The phase boundaries are not defined by fixed time instants or explicit switching logic, but arise implicitly from state evolution and physical events. Therefore, the controller remains a single stationary policy

$$\mathbf{u}_t = \pi_{\theta}(\mathbf{o}_t) \quad (32)$$

and phase-dependent behavior is achieved through state-dependent objective modulation. $\mathbf{o}_t$ denotes the observation vector and π_{θ} is a stationary policy parameterized by θ . No explicit phase variable appears in $\mathbf{o}_t$. For example, tangential regulation is scaled by the contact ratio

$$\sigma(t) = \frac{1}{4} \sum_{i=1}^4 c_i(t), \quad \bar{v}_t^2 = \sigma(t) v_t^2, \quad (33)$$

which increases as support is established.

Phase information is instead conveyed implicitly through physically meaningful state variables and history-dependent quantities. These include contact-related measurements, slope-aligned velocity components, and episode-level statistics such as accumulated peak accelerations and minimum clearance. As the system evolves, these variables naturally reflect the current phase of the landing process, allowing the policy to infer phase-dominant requirements from the observation alone.

Moreover, phase differentiation is reinforced through objective activation patterns rather than discrete switching. Certain penalties or constraints become influential only when their corresponding physical conditions are met. For example, tangential slip suppression is emphasized once sufficient contact support is established, while peak impact penalties become active only when new extreme events occur. This results in a smooth, state-dependent modulation of control priorities across phases.

The key advantage of this implicit phase encoding lies in its balance between structure and flexibility. The policy remains end-to-end and time-invariant, avoiding brittle mode

transitions, while the learning problem retains strong physical interpretability. As a result, the learned controller exhibits phase-appropriate behaviors—contact-consistent initialization, peak-aware buffering, and sustained stabilization—without requiring explicit phase detection or supervisory logic.

This phase-structured yet implicitly encoded framework forms the foundation for the terrain-aware state and control representations introduced in the following section.

4. Terrain-Aware State and Control Representation

Sloped-terrain landing differs fundamentally from flat-ground landing in that the effective support direction associated with contacts is not aligned with a fixed axis of the world frame. Consequently, control variables defined in the world frame (e.g., vertical velocity along the global axis) do not possess consistent physical semantics across different slopes and contact configurations. This mismatch induces reward ambiguity, learning distribution drift, and reduced generalization.

To address this issue without requiring explicit terrain geometry or slope inclination priors, we introduce a terrain-aware state and control representation based on an equivalent support direction constructed from contact-consistent information. All velocity-related variables are expressed in the body frame and further decomposed into normal and tangential components with respect to the equivalent support direction. A contact-gated modulation mechanism is then applied to ensure that tangential objectives are enforced only when effective support is established.

4.1. Equivalent Support Direction Construction

Let the world frame be denoted by $\{W\}$ and the robot body frame by $\{B\}$, as illustrated in Figure 2, which shows the dynamic model of the lunar quadruped robot with all relevant coordinate frames, forces, and kinematic variables. The local terrain geometry is unknown to the controller in realistic lunar missions; therefore, we avoid using explicit slope angles or the true terrain normal as policy inputs. Instead, we construct an equivalent support direction $\mathbf{n}_w \in \mathbb{R}^3$ that represents the dominant support normal formed by the current foot-ground interaction.

In simulation, a physically grounded choice is to aggregate the contact normals (or equivalently, the dominant directions of contact reactions) over all feet that are in contact. Let $\mathcal{C}(t) \subseteq \{1, 2, 3, 4\}$ denote the set of contacting feet at time t , and let $\mathbf{n}_w^{(i)}(t)$ denote the contact normal associated with foot $i \in \mathcal{C}(t)$ (pointing outward from the terrain). The equivalent support direction is defined as the normalized aggregate

$$\tilde{\mathbf{n}}_w(t) = \sum_{i \in \mathcal{C}(t)} w_i(t) \mathbf{n}_w^{(i)}(t), \quad \mathbf{n}_w(t) = \frac{\tilde{\mathbf{n}}_w(t)}{\|\tilde{\mathbf{n}}_w(t)\| + \varepsilon} \quad (34)$$

where $w_i(t) \geq 0$ are nonnegative weights (e.g., uniform weights or functions of normal contact forces), and $\varepsilon > 0$ is a small constant for numerical robustness.

To eliminate sign ambiguity (e.g., the normal pointing into the terrain), a consistency rule is imposed so that $\mathbf{n}_w$ aligns with the physically admissible support half-space. A simple and effective convention is to enforce a positive vertical component:

$$\text{if } \mathbf{e}_z^\top \mathbf{n}_w(t) < 0, \text{ then } \mathbf{n}_w(t) \leftarrow -\mathbf{n}_w(t) \quad (35)$$

where $\mathbf{e}_z = [0, 0, 1]^\top$ is the unit vector of the world z-axis.

The key property of $\mathbf{n}_w(t)$ is that it reflects the support direction implied by contact physics rather than by external terrain priors. For rigid planar slopes, $\mathbf{n}_w$ is approximately

constant within an episode; for mild contact variations, $\mathbf{n}_w(t)$ varies slowly and remains a stable geometric reference for defining control variables.

4.2. Body-Frame Velocity Representation

Directly using world-frame velocities as policy inputs causes distribution drift under varying slopes and body attitudes, because identical physical motions yield different coordinate representations. To reduce this sensitivity, we express velocities in the robot body frame.

Let $\mathbf{R}_{WB} \in SO(3)$ denote the rotation matrix that maps vectors from the body frame $\{B\}$ to the world frame $\{W\}$. The base linear and angular velocities in the world frame are $\mathbf{v}_W \in \mathbb{R}^3$ and $\boldsymbol{\omega}_W \in \mathbb{R}^3$, respectively. Their body-frame representations are

$$\mathbf{v}_B = \mathbf{R}_{WB}^\top \mathbf{v}_W, \quad \boldsymbol{\omega}_B = \mathbf{R}_{WB}^\top \boldsymbol{\omega}_W \quad (36)$$

Similarly, the equivalent support direction is transformed into the body frame as

$$\mathbf{n}_B = \frac{\mathbf{R}_{WB}^\top \mathbf{n}_W}{\|\mathbf{R}_{WB}^\top \mathbf{n}_W\| + \varepsilon} \quad (37)$$

This body-aligned formulation has two advantages. First, it provides a consistent statistical structure for policy learning: forward/lateral/downward motions are always encoded along the body axes regardless of slope orientation. Second, it naturally supports terrain-agnostic generalization because all slope-dependent information enters only through $\mathbf{n}_B$, rather than through explicit slope angles.

4.3. Normal–Tangential Velocity Decomposition

Given $\mathbf{v}_B$ and the equivalent support direction $\mathbf{n}_B$, we decompose the body-frame linear velocity into a normal component (compression/impact direction) and a tangential component (slip direction) relative to the support plane orthogonal to $\mathbf{n}_B$.

The normal velocity scalar captures the compression/penetration tendency along the effective support direction. It is the most relevant velocity component for buffering stroke utilization, rebound risk, and peak impact mitigation.

The tangential velocity vector is the residual after removing the normal projection:

$$\mathbf{v}_t = \mathbf{v}_B - v_n \mathbf{n}_B \quad (38)$$

The slip tendency can be quantified by its norm.

For practical implementation and to provide direction-resolved features, an orthonormal tangential basis $\{\mathbf{t}_1, \mathbf{t}_2\}$ spanning the plane orthogonal to $\mathbf{n}_B$ may be constructed. For example, choose a reference axis $\mathbf{a} \in \{\mathbf{e}_x, \mathbf{e}_y\}$ that is not nearly parallel to $\mathbf{n}_B$, and define

$$\mathbf{t}_1 = \frac{\mathbf{a} - (\mathbf{n}_B^\top \mathbf{a}) \mathbf{n}_B}{\|\mathbf{a} - (\mathbf{n}_B^\top \mathbf{a}) \mathbf{n}_B\| + \varepsilon}, \quad \mathbf{t}_2 = \mathbf{n}_B \times \mathbf{t}_1 \quad (39)$$

The tangential components are then

$$v_{t1} = \mathbf{t}_1^\top \mathbf{v}_B, \quad v_{t2} = \mathbf{t}_2^\top \mathbf{v}_B, \quad v_t = \sqrt{v_{t1}^2 + v_{t2}^2} \quad (40)$$

This decomposition yields a one-to-one correspondence between kinematic variables and physical objectives: v_n governs buffering and impact suppression, while v_t governs anti-slip stabilization. Compared with applying penalties to world-frame components, this

representation avoids objective drift caused by terrain orientation changes and improves interpretability for both reward shaping and evaluation metrics.

4.4. Contact-Gated Variable Modulation

A key subtlety in sloped landing is that tangential velocity suppression should not be enforced uniformly throughout the entire episode. During flight or insufficient support, aggressive tangential regulation may induce non-physical mid-air actions or over-constrain the policy near contact establishment, which can increase touchdown impacts and destabilize learning.

To address this, we introduce a contact-gated modulation mechanism that scales tangential-related objectives according to the current support quality. Let $c_i(t) \in \{0, 1\}$ be a binary indicator of whether foot i is in effective contact at time t . Define the contact ratio

$$\sigma(t) = \frac{1}{4} \sum_{i=1}^4 c_i(t) \in [0, 1] \quad (41)$$

Tangential motion variables (or their associated penalties/constraints) are modulated by $\sigma(t)$. A simple and effective choice is linear gating, in which tangential suppression is strengthened as more feet establish support.

$$\tilde{v}_t^2 = \sigma(t) v_t^2 \quad (42)$$

More generally, one may use a smooth nonlinear gate $g(\sigma)$ satisfying $g(0) = 0$ and $g(1) = 1$, e.g.,

$$g(\sigma) = \sigma^\kappa, \kappa \geq 1, \quad \text{or} \quad g(\sigma) = \text{clip}\left(\frac{\sigma - \sigma_0}{1 - \sigma_0}, 0, 1\right) \quad (43)$$

and apply it as $\tilde{v}_t^2 = g(\sigma)$.

This contact-gated modulation has two important consequences. First, it implicitly differentiates control priorities across phases without introducing explicit phase switching: tangential objectives are naturally down-weighted in the contact preparation stage and become dominant only after stable support emerges. Second, it improves training stability by preventing early exploration from being overly penalized for tangential components that are physically unavoidable before full contact is formed.

In summary, the proposed terrain-aware representation consists of: (i) an equivalent support direction $\mathbf{n}_w$ constructed from contact-consistent information, (ii) body-frame velocity representation $(\mathbf{v}_B, \boldsymbol{\omega}_B)$, (iii) normal–tangential velocity decomposition $(v_n, \mathbf{v}_t)$, and (iv) contact-gated modulation via $\sigma(t)$. Together, these elements provide physically interpretable, terrain-agnostic control variables that significantly reduce coordinate-induced distribution drift and support consistent learning across varying slopes and contact conditions. The complete terrain-aware representation pipeline is illustrated in Figure 4.

5. Learning-Based Control Design

This section presents the learning-based control formulation for sloped-terrain landing impact mitigation. Building upon the phase-structured framework and terrain-aware state representation introduced earlier, we design a single end-to-end reinforcement learning controller that remains stationary in time while exhibiting phase-appropriate behaviors through structured observations, action parameterization, and objective modeling.

5.1. Policy Architecture and Action Parameterization

The control policy is formulated as a stationary stochastic mapping

$$\mathbf{u}_t = \pi_{\theta}(\mathbf{o}_t) \quad (44)$$

where $\mathbf{o}_t \in \mathbb{R}^{n_o}$ denotes the observation vector at control step t , $\mathbf{u}_t \in \mathbb{R}^{12}$ is the policy output, and π_{θ} is parameterized by θ . u_t is a 12-dimensional continuous action representing residual joint-position commands.

Let $\pi_{\theta_{old}}$ denote the policy used to collect rollout data. The probability ratio is defined as

$$r_t(\theta) = \frac{\pi_{\theta}(a_t|o_t)}{\pi_{\theta_{old}}(a_t|o_t)} \quad (45)$$

With $\hat{A}_t$ denoting the estimated advantage, PPO maximizes the clipped objective

$$L^{\text{clip}}(\theta) = \mathbb{E}_t \left[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t) \right]. \quad (46)$$

The critic network is trained to approximate the state-value function $V_{\phi}(o_t)$, and the advantage $\hat{A}_t$ is computed from rollout data using discounted returns or generalized advantage estimation.

This update mechanism improves optimization stability, which is particularly important for landing tasks involving hybrid contact dynamics, peak-dominated transients, and strong nonlinearities. The action space in this work is continuous and bounded, with $u_t \in [-1, 1]^{12}$ representing residual joint commands, making PPO a suitable choice for the present problem.

Directly outputting absolute joint position or torque commands often leads to poor exploration efficiency and unstable behaviors in contact-rich tasks. Instead, we adopt a residual joint-space action parameterization centered around a nominal posture $\mathbf{q}_{\text{nom}} \in \mathbb{R}^{12}$, corresponding to a statically stable standing configuration.

The desired joint positions are defined as

$$\mathbf{q}_{\text{des}}(t) = \mathbf{q}_{\text{nom}} + \alpha \mathbf{u}_t \quad (47)$$

where $\alpha > 0$ is a fixed scaling factor that bounds the magnitude of policy-induced deviations. The resulting commands are clipped to respect joint limits,

$$\mathbf{q}_{\text{des}}(t) \in [\mathbf{q}_{\text{min}}, \mathbf{q}_{\text{max}}] \quad (48)$$

Joint torques are then generated by a low-level proportional–derivative (PD) controller. To further improve clarity, a flowchart illustrating the reinforcement learning pipeline has been added. It summarizes the main steps including domain-randomized environment initialization under lunar gravity, terrain-aware observation construction, PPO policy inference, residual action mapping to joint commands, low-level PD execution in the simulation environment, response collection, reward computation, and policy update.

$$\boldsymbol{\tau}(t) = \mathbf{K}_p(\mathbf{q}_{\text{des}}(t) - \mathbf{q}(t)) - \mathbf{K}_d\dot{\mathbf{q}}(t) \quad (49)$$

with saturation applied to enforce actuator torque limits. This hierarchical structure clarifies how the learned high-level policy interfaces with the physical actuator layer.

This residual formulation offers three advantages:

- (i) It preserves a stable baseline behavior;
- (ii) It reduces the effective action space explored by the policy;

- (iii) It allows the policy to focus on critical adjustments during landing buffering rather than generating full-body motion from scratch.

5.2. Episode-Peak-Based Impact Suppression Modeling

Landing impact mitigation is inherently an extremum-oriented control problem, where structural safety and stability are governed by short-duration peak events rather than by time-averaged quantities. Standard reinforcement learning objectives based on cumulative step-wise costs are poorly suited to this setting.

5.2.1. Peak Acceleration Definition

Let $\mathbf{a}_{\text{lin}}(t)$ and $\boldsymbol{\alpha}(t)$ denote the base linear and angular accelerations, respectively. The gravity-compensated excess linear acceleration is defined as

$$\mathbf{a}_{\text{ex}}(t) = \mathbf{a}_{\text{lin}}(t) - \mathbf{g} \quad (50)$$

The corresponding instantaneous impact intensities are

$$a_{\text{ex}}(t) = \|\mathbf{a}_{\text{ex}}(t)\|, \quad \alpha(t) = \|\boldsymbol{\alpha}(t)\| \quad (51)$$

within each control interval, the step-wise peak values are recorded as

$$a_{\text{max}}^{\text{step}}(k) = \max_{t \in [k\Delta t, (k+1)\Delta t)} a_{\text{ex}}(t) \quad (52)$$

and analogously for angular acceleration.

At the episode level, the historical worst-case peaks are updated according to

$$a_{\text{max}}^{\text{epi}}(k) = \max(a_{\text{max}}^{\text{epi}}(k-1), a_{\text{max}}^{\text{step}}(k)) \quad (53)$$

5.2.2. Peak Growth Modeling

Directly minimizing $a_{\text{max}}^{\text{epi}}$ leads to sparse and non-smooth learning signals due to the max operator. To alleviate this issue, we introduce the peak growth increment

$$\Delta a_{\text{max}}(k) = \max(0, a_{\text{max}}^{\text{step}}(k) - a_{\text{max}}^{\text{epi}}(k-1)) \quad (54)$$

which is nonzero only when a new, more adverse impact occurs.

This formulation converts an episode-level extremum objective into a sequence of incremental penalties, enabling the policy to learn how specific actions contribute to the creation of new impact peaks.

5.2.3. Barrier-Based Safety Modeling

Engineering constraints impose explicit upper bounds on admissible impact levels. Let $\bar{a}$ and $\bar{\alpha}$ denote safety thresholds for linear and angular accelerations, respectively. A soft barrier penalty is applied when step-wise peaks approach or exceed these limits,

$$\ell_{\text{bar}}(k) = \beta_a \left[\max(0, a_{\text{max}}^{\text{step}}(k) - \bar{a}) \right]^2 + \beta_\alpha \left[\max(0, \alpha_{\text{max}}^{\text{step}}(k) - \bar{\alpha}) \right]^2 \quad (55)$$

where $\beta_a, \beta_\alpha > 0$ are weighting coefficients.

For severe violations beyond higher abort thresholds, the episode is terminated to ensure numerical and physical safety. The combination of soft barriers and hard aborts balances learnability with strict safety enforcement.

5.3. Reward Function Design

The overall reward function integrates multiple complementary objectives corresponding to survival, stability, control regularization, and impact suppression. At each control step k , the reward is expressed as

$$r_k = r_{\text{alive}} + r_{\text{state}} + r_{\text{ctrl}} + r_{\text{impact}} + r_{\text{terminal}} \quad (56)$$

5.3.1. State Stabilization Terms

Using the terrain-aware decomposition introduced in Section 4, let $v_n(k)$ and $v_t(k)$ denote the normal and tangential velocity components, respectively. Dense stabilization penalties are defined as

$$r_v(k) = -w_n v_n(k)^2 - w_t \sigma(k) v_t(k)^2 \quad (57)$$

where $\sigma(k) \in [0, 1]$ is the contact ratio used to gate tangential suppression.

Posture and angular velocity stabilization are enforced via

$$r_{\text{att}}(k) = -w_\phi \|\phi(k)\|^2 - w_\omega \|\omega(k)\|^2 \quad (58)$$

$$r_h(k) = -w_h [\max(0, h_{\min} - h(k))]^2 \quad (59)$$

where $h(k)$ is the minimum body–terrain clearance along the support direction.

5.3.2. Control Regularization

To ensure executability and avoid aggressive oscillations, control effort and action rate are penalized:

$$r_{\text{ctrl}}(k) = -w_\tau \|\tau(k)\|^2 - w_{\Delta u} \|\mathbf{u}_k - \mathbf{u}_{k-1}\|^2 \quad (60)$$

5.3.3. Impact Peak Penalties

Impact-related penalties consist of step-wise peak costs, peak growth costs, and barrier terms:

$$r_{\text{impact}}(k) = -w_s \left(a_{\max}^{\text{step}}(k) \right)^2 - w_g (\Delta a_{\max}(k))^2 - \ell_{\text{bar}}(k) \quad (61)$$

Terminal rewards further incorporate episode-level peak suppression when success is achieved.

5.4. Success Criteria with Stability Window and Buffering Sufficiency

Defining success for landing impact mitigation requires more than instantaneous threshold satisfaction. Two complementary criteria are enforced: sustained stability and buffering sufficiency.

5.4.1. Stability Window Criterion

Let $\phi(k)$, $v_n(k)$, $v_t(k)$, $\omega(k)$, and $h(k)$ denote posture, velocity, angular velocity, and clearance variables at step k . The instantaneous stability condition is defined as

$$S(k) = \begin{cases} 1, & \|\phi(k)\| \leq \bar{\phi}, |v_n(k)| \leq \bar{v}_n, v_t(k) \leq \bar{v}_t \\ & \|\omega(k)\| \leq \bar{\omega}, h(k) \geq \bar{h}, \sigma(k) \geq \bar{\sigma} \\ 0, & \text{otherwise} \end{cases} \quad (62)$$

Let N_s denote the number of steps in the stability window. Stability is declared only if

$$\sum_{i=k-N_s+1}^k S(i) = N_s \quad (63)$$

This temporal requirement eliminates pseudo-stability caused by transient zero-crossings or brief posture recovery.

5.4.2. Buffering Sufficiency Criterion

To ensure that genuine energy dissipation has occurred, a buffering sufficiency condition is imposed. Let h_0 be the initial clearance and $h_{\min}$ the minimum clearance attained during the episode. The effective compression is

$$\Delta h = h_0 - h_{\min} \quad (64)$$

Buffering is considered sufficient if

$$\Delta h \geq \Delta h_{\text{req}} \quad (65)$$

where Δh_{req} is a prescribed minimum compression threshold.

5.4.3. Overall Success Definition

The landing is deemed successful if and only if both criteria are satisfied:

$$\text{Success} = (\text{Stability Window Satisfied}) \wedge (\text{Buffering Sufficiency Satisfied}) \quad (66)$$

This joint criterion prevents opportunistic behaviors that achieve apparent stability without meaningful buffering and aligns the terminal objective of reinforcement learning with the true engineering semantics of soft landing.

6. Training Configuration and Implementation Details

This section describes the training configuration and implementation details adopted to obtain a stable and generalizable policy for sloped-terrain landing impact mitigation. The task involves long-horizon hybrid dynamics, peak-dominated transients, and multi-objective reward shaping. Consequently, careful configuration of the reinforcement learning algorithm, observation/action spaces, and domain randomization strategy is required to balance sample efficiency, optimization stability, and generalization.

6.1. PPO Hyperparameter Configuration

The policy is trained using Proximal Policy Optimization (PPO), an on-policy policy-gradient method well suited for continuous control tasks with high-dimensional observations and stochastic dynamics. PPO is adopted in this work because the landing task involves high-dimensional continuous control, hybrid contact transitions, and strongly coupled objectives including impact mitigation, buffering, and stabilization. Compared with algorithms that are more sensitive to value-estimation errors or replay-buffer distribution drift, PPO provides a favorable balance between optimization stability, implementation simplicity, and robustness in contact-rich robotic systems.

6.2. Observation and Action Space Specification

6.2.1. Observation Space

At each control step t , the policy receives an observation vector

$$\mathbf{o}_t \in \mathbb{R}^{n_o} \quad (67)$$

which aggregates kinematic, dynamic, contact, and episode-level information. The observation is updated at the control frequency and normalized online to improve numerical conditioning.

The observation vector includes the following components:

1. Terrain-aligned geometric variables
 - Minimum body–terrain clearance along the equivalent support direction $h(t)$.
 - Equivalent support direction expressed in the body frame $\mathbf{n}_B(t)$.
2. Velocity and posture states
 - Normal and tangential velocity components $(v_n(t), v_{t1}(t), v_{t2}(t))$.
 - Body roll and pitch angles $\phi(t)$.
 - Body-frame angular velocity $\omega(t)$.
3. Joint states
 - Joint positions $\mathbf{q}(t)$ and velocities $\dot{\mathbf{q}}(t)$ for all actuated joints.
4. Contact-related information
 - Binary foot contact indicators $c_i(t)$.
 - Accumulated normal contact forces $f_n^{(i)}(t)$.
5. Episode-level peak statistics
 - Historical peak linear and angular accelerations $(a_{\max}^{\text{epi}}(t), \alpha_{\max}^{\text{epi}}(t))$.

The inclusion of episode-level peak statistics effectively augments the Markov state with partial history information, allowing the policy to adapt its behavior based on previously experienced extreme events. This state expansion mitigates non-Markovian effects introduced by extremum-based objectives.

6.2.2. Action Space

The observation vector includes terrain-aligned geometric variables, body-frame normal and tangential velocities, posture and angular velocity, joint states, contact indicators, contact forces, and episode-level peak statistics. The reward function is constructed as a combination of survival, stabilization, control regularization, and impact peak suppression, including step-wise peak penalties, peak-growth terms, and barrier functions.

Actions are linearly scaled and mapped to desired joint positions relative to a nominal posture, as described in Section 5.1.

$$\mathbf{u}_t \in [-1, 1]^{12} \quad (68)$$

This action representation ensures bounded exploration, preserves actuator feasibility, and provides a smooth interface between the high-level policy and the low-level joint-space PD controller.

6.3. Domain Randomization Strategy

To enhance robustness and generalization across uncertain lunar landing conditions, domain randomization is applied at the beginning of each training episode. The objective is to expose the policy to a distribution of plausible operating conditions rather than a single nominal model.

6.3.1. Mass and Inertia Randomization

Variations in payload configuration and propellant consumption lead to uncertainty in the robot's mass and inertia. To model this effect, the base mass m_{mm} is randomly sampled from a bounded interval

$$m \sim \mathcal{U}(m_{\min}, m_{\max}) \quad (69)$$

and the corresponding rotational inertia tensor I is scaled proportionally

$$\mathbf{I} = \frac{m}{m_0} \mathbf{I}_0 \quad (70)$$

where m_0 and $\mathbf{I}_0$ denote the nominal mass and inertia, respectively. This preserves physical consistency while introducing significant variation in dynamic response.

6.3.2. Initial Velocity Randomization

Uncertainty in pre-touchdown conditions is modeled by randomizing the initial base velocity. Horizontal and vertical components are sampled independently within predefined bounds,

$$v_x \sim \mathcal{U}(-\bar{v}_x, \bar{v}_x) \quad (71)$$

$$v_y \sim \mathcal{U}(-\bar{v}_y, \bar{v}_y) \quad (72)$$

$$v_z \sim \mathcal{U}(v_{z,\min}, v_{z,\max}) \quad (73)$$

where negative v_z indicates downward motion. This setting covers landing scenarios with lateral drift and varying descent speeds, encouraging the policy to learn robust tangential slip suppression and impact buffering strategies.

6.3.3. Contact-Consistent Initialization

To avoid corrupting training signals with non-physical artifacts, environment initialization enforces contact consistency before velocity injection. The robot is first placed in a configuration with stable multi-leg contact and zero velocity, followed by a short settling period. Initial base velocity is then applied while joint velocities are adjusted to maintain zero foot linear velocity in the world frame. This procedure ensures that early impact responses arise from genuine control actions rather than from inconsistent kinematic initialization.

7. Simulation-Based Training

7.1. Simulation Environment and Training Configuration

All experiments are conducted in a physics-based simulation environment using the MuJoCo engine. The simulation and control timing parameters are summarized in Table 2.

Table 2. Simulation and control timing parameters used in training.

Parameter	Symbol/Name	Value
Simulation timestep	(Δt_{sim})	0.002 s
Control frequency	(f_{ctrl})	50 Hz
Control period	(Δt_{ctrl})	0.02 s
Episode duration	(T_{ep})	8.0 s
Episode length (steps)	-	400 control steps

The control policy is executed under a zero-order-hold assumption at 50 Hz, while the simulator internally integrates dynamics at 2 ms resolution. The quadruped robot has 12 actuated joints. A nominal joint configuration ($\mathbf{q}_{\text{nom}} \in \mathbb{R}^{12}$) is defined and reused consistently in action parameterization, reset, and reward computation:

$$\mathbf{q}_{\text{nom}} = \begin{bmatrix} 0 & -0.25 & 0.729 \\ 0 & -0.25 & 0.729 \\ 0 & -0.841 & 0.686 \\ 0 & -0.841 & 0.686 \end{bmatrix} \quad (74)$$

The initial joint configuration at reset is identical:

$$\mathbf{q}(0) = \mathbf{q}_{\text{nom}} \quad (75)$$

The robot base is initialized above an inclined plane with a minimum slope-normal clearance.

$$h_{\min}(0) \geq z_{\min} = 0.10 \text{ m} \quad (76)$$

The base linear velocity at reset is randomized to emulate uncertain landing conditions:

$$v_z(0) \sim \text{U}(0.3, 1.1) \text{ m/s} \quad (77)$$

$$v_x(0), v_y(0) \sim \text{U}(-0.4, 0.4) \text{ m/s} \quad (78)$$

To preserve physically consistent ground contact, joint velocities are solved such that the linear velocities of all four foot sites are zero in the world frame at reset:

$$\mathbf{p}_{\text{foot},i}(0) = \mathbf{0}, \quad i = 1, \dots, 4 \quad (79)$$

with a numerical velocity clip of 10 m/s applied for stability.

Joint-level tracking is implemented via a PD controller with parameters summarized in Table 3.

Table 3. Low-Level Control Parameters.

Parameter	Symbol	Value
Proportional gain	(K_p)	210.0
Derivative gain	(K_d)	21.0
Torque limit	$(\tau_{\max})$	$\pm 140 \text{ N}\cdot\text{m}$

The policy outputs normalized actions ($\mathbf{a} \in [-1, 1]^{12}$), which are mapped to desired joint positions via a residual formulation:

$$\mathbf{q}_{\text{des}} = \mathbf{q}_{\text{nom}} + \alpha \mathbf{a}, \quad \alpha = 0.60 \quad (80)$$

Safety is enforced through both soft penalties and hard termination thresholds. Excess linear acceleration is defined

$$\mathbf{a}_{\text{excess}} = \mathbf{a}_{\text{lin}} - \mathbf{g} \quad (81)$$

where gravity is removed. Thresholds are summarized in Table 4.

Table 4. Safety Thresholds.

Quantity	Limit	Abort
Linear acceleration	2.0 m/s^2	10.2 m/s^2
Angular acceleration	12.0 rad/s^2	30.0 rad/s^2
Body tilt	20°	termination

The total reward at time step t is expressed as

$$r_t = r_{\text{alive}} + r_{\text{acc}} + r_{\text{vel}} + r_{\text{att}} + r_{\text{height}} + r_{\text{effort}} + r_{\text{contact}} + r_{\text{terminal}} \quad (82)$$

All reward weights are listed exhaustively in Table 5.

Table 5. Reward and Penalty Weights.

Category	Parameter	Value
Alive penalty	(r_{alive})	−0.01
Linear acc barrier	(w_{acc})	50.0
Angular acc barrier	(w_{angacc})	5.0
Step peak lin acc	$(w_{step,lin})$	0.002
Step peak ang acc	$(w_{step,ang})$	0.001
Peak growth lin	$(w_{grow,lin})$	0.05
Peak growth ang	$(w_{grow,ang})$	0.02
Terminal peak lin	$(w_{term,lin})$	1.0
Terminal peak ang	$(w_{term,ang})$	0.3
Normal velocity	(w_{v_n})	1.2
Tangential velocity	(w_{v_t})	2.0
Attitude	(w_{att})	30.0
Angular velocity	(w_{ω})	0.10
Height barrier	(w_h)	350.0
Joint deviation	(w_q)	0.50
Joint velocity	$(w_{\dot{q}})$	0.03
Torque effort	(w_{τ})	1×10^{-4}
Action smoothness	$(w_{\Delta a})$	1×10^{-3}
Contact ratio	(w_c)	0.15
Normal force variance	(w_{f_n})	1×10^{-4}
Failure penalty	—	−300
Success bonus	—	+250

A compression gate is enforced to ensure effective impact buffering: compression ≥ 0.20 m.

A landing is considered successful if all conditions in Table 6 are satisfied continuously over a sliding window of 0.9 s.

Table 6. Stability Window Criteria.

Metric	Threshold
Roll pitch	$\leq 2^\circ$
Normal velocity	≤ 0.04 m/s
Tangential velocity	≤ 0.05 m/s
Angular velocity norm	≤ 0.125 rad/s
Minimum clearance	≥ 0.15 m
Contact ratio	≥ 0.75 (≥ 3 feet)
Window duration	0.9 s

The controller is trained using Proximal Policy Optimization (PPO). All hyperparameters are summarized in Table 7.

Table 7. PPO Training Parameters.

Parameter	Value
Algorithm	PPO
Policy network	MLP
Actor critic layers	[512, 512]
Learning rate	2×10^{-4}
Rollout length	2048
Batch size	2048
Epochs per update	5
Discount factor (γ)	0.99
GAE (λ)	0.95
Clip range	0.2
Target KL	0.12
Entropy coefficient	0.0
Value coefficient	0.5
Gradient norm	0.5
Parallel envs	16
Total steps	5×10^6
Device	GPU (CUDA)

Observations are normalized online, while rewards remain unnormalized to preserve physical meaning.

7.2. Task Definition and Evaluation Metrics

Task definition. The task is a terrain-landing/buffering control problem in which the quadruped experiences an initial impact velocity and must (i) suppress peak impact-induced accelerations, (ii) realize sufficient buffering through leg compression, and (iii) achieve a stable post-impact equilibrium within a finite time window. The landing process is treated as successful only when safety bounds and stability conditions are simultaneously satisfied.

To evaluate learning progress and convergence behavior, the following metrics are recorded during training and evaluation:

- Episodic return: the cumulative reward over an episode, reflecting the composite objective combining safety, stability, and efficiency;
- Evaluation return: computed in a separate evaluation environment to assess generalization and detect overfitting.

These metrics are used to analyze whether improved task performance correlates with safer and more stable physical behavior. To interpret the physical behavior induced by the learned policy, representative episodes are analyzed using time-domain trajectories of key variables, including:

- Minimum base height and compression;
- Step-wise and episode-level peak accelerations;
- Angular velocity norm;
- Roll and pitch angles.

These trajectories allow the landing process to be decomposed into impact, buffering, and stabilization phases, and provide direct insight into how control effort and contact forces evolve over time. To evaluate robustness under parametric uncertainty, the learned policy is tested over a grid of base masses and impact velocities. For each configuration, multiple rollouts are performed and distribution-level statistics are computed. In particular, the empirical cumulative distribution function (CDF) of the peak linear acceleration,

$$F_a(x) = \Pr(a_{\text{peak}} \leq x) \quad (83)$$

is used to assess tail behavior and safety margins. Narrow CDFs without heavy tails indicate consistent and predictable impact mitigation, while shifts in the CDF across conditions reveal how the policy adapts to changing mass and velocity.

7.3. Experimental Scenarios and Protocol

Policies are trained using PPO with vectorized environments. Per-episode metrics are accumulated online and logged at episode termination into a structured CSV, including episode return, episode length, termination flags, termination reason, peak accelerations, compression usage, end-of-episode velocities, contact indicators, and randomized parameters such as base mass. In parallel, rolling-window aggregates (e.g., mean return and success rate) are recorded to the SB3 logger to support consistent training-curve visualization and multi-seed statistical analysis.

To rigorously assess robustness across physically meaningful operating conditions, we adopt a grid-based evaluation protocol over (i) base mass and (ii) impact speed magnitude. Let m denote the base mass and v_0 denote the initial speed magnitude at reset. The evaluation grid is specified by finite sets

$$m \in \mathbf{M} = \{m_1, \dots, m_{N_m}\}, \quad v_0 \in \mathbf{V} = \{v_1, \dots, v_{N_v}\} \quad (84)$$

For each grid cell (m, v_0) , we perform N independent rollouts (episodes) with identical reset parameters and compute empirical statistics of the metrics defined in Section 7.2.

Initial velocity is constructed by splitting v_0 into horizontal and vertical components using a fixed angle θ (with $v_z < 0$ indicating downward motion):

$$v_{x,0} = v_0 \sin \theta, \quad v_{y,0} = 0, \quad v_{z,0} = -v_0 \cos \theta \quad (85)$$

This design standardizes impact direction across evaluations and enables controlled variation between predominantly vertical impacts ($\theta \approx 0$) and more oblique landings ($\theta > 0$).

We compare the learned controller against a PD-only baseline that disables policy modulation by setting $a(t) \equiv 0$, which reduces the command to tracking the nominal posture $q_{\text{des}} = q_{\text{nom}}$. Importantly, both RL and baseline controllers share the same inner-loop gains, torque saturation, observation preprocessing (including frozen VecNormalize during evaluation), and identical success/termination criteria. This ensures that observed improvements are attributable to learned action modulation rather than implementation differences.

7.4. Results Analysis

Figure 5 presents the training and evaluation curves of the proposed reinforcement learning controller. Overall, the results demonstrate stable learning progress, effective impact mitigation behavior, and acceptable optimization dynamics throughout most of the training horizon.

From a learning performance perspective, the episodic return exhibits a clear and consistent increasing trend over training timesteps (as shown in Figure 5b). Both the raw episode return and its moving-average counterpart show monotonic improvement, indicating that the policy continuously refines its landing behavior rather than oscillating between suboptimal strategies. The evaluation return follows a similar trend and remains stable after convergence, suggesting that the learned policy generalizes well beyond the stochasticity encountered during training. Although short-term fluctuations are observed,

particularly in the evaluation return, these variations remain bounded and do not indicate performance collapse.

The evolution of the compression-related metric further provides insight into the learned landing strategy (as shown in Figure 5a). At early training stages, the compression usage is relatively low, implying insufficient utilization of the available stroke during impact. As training progresses, the compression steadily increases and stabilizes around a consistent level, indicating that the policy learns to actively exploit the compliant mechanism to dissipate impact energy. The subsequent slight decrease and stabilization suggest a transition from overly aggressive compression toward a more balanced trade-off between impact absorption and post-landing stability.

The physical behavior induced by the learned policy is illustrated through a representative landing episode (as shown in Figure 5c). The evolution of the minimum foot height and compression clearly demonstrates active utilization of the compliant stroke. Upon initial contact, the minimum height decreases rapidly while the compression increases smoothly and saturates at a stable level, indicating that the impact energy is absorbed progressively rather than impulsively. After the peak compression phase, the system transitions into a recovery regime, during which the minimum height slightly rebounds and converges to a steady value, reflecting successful stabilization without secondary impacts.

This buffering behavior is closely coupled with the acceleration response (as shown in Figure 5d). The step-wise peak linear acceleration rises sharply during the initial contact phase and reaches its maximum within a short time window, while the episode-level peak remains bounded thereafter. The rapid decay of acceleration following the peak indicates that the policy effectively dissipates impact energy and avoids prolonged high-load conditions. The angular velocity norm exhibits a similar pattern (as shown in Figure 5e), with an initial transient peak followed by a smooth decay toward near-zero values, demonstrating effective attenuation of rotational disturbances induced by asymmetric contact or terrain interaction.

Attitude responses further confirm this stabilization mechanism (as shown in Figure 5f). Both roll and pitch remain within small angular bounds throughout the episode. After initial transient deviations immediately following touchdown, the attitude converges smoothly without sustained oscillations, indicating that the policy implicitly coordinates leg forces to regulate body orientation rather than relying on abrupt corrective actions.

Contact quality metrics provide additional insight into force distribution during landing (as shown in Figure 5g). The contact ratio rapidly approaches and maintains a high level once ground contact is established, while the summed normal force increases to a peak during maximum compression and then gradually decreases as the system settles. This evolution suggests that load sharing among legs is maintained throughout the landing process, avoiding abrupt force redistribution that could trigger slipping or tipping.

From a control-effort standpoint, the RMS of the action remains relatively small, while the torque RMS exhibits a pronounced peak during the impact phase followed by a gradual decay (as shown in Figure 5h). This indicates that the policy concentrates control effort during the critical energy dissipation interval and subsequently reduces actuation intensity as stability is achieved. Such behavior is consistent with an efficient control strategy that avoids sustained high-torque operation after the landing task is completed.

Collectively, these results demonstrate that the proposed policy exhibits well-structured learning dynamics, physically interpretable landing behavior, and robust scaling properties. The controller not only converges to high-performance solutions but also achieves impact mitigation through coordinated compression, force regulation, and attitude stabilization, while maintaining reasonable control effort across a broad range of operating conditions.

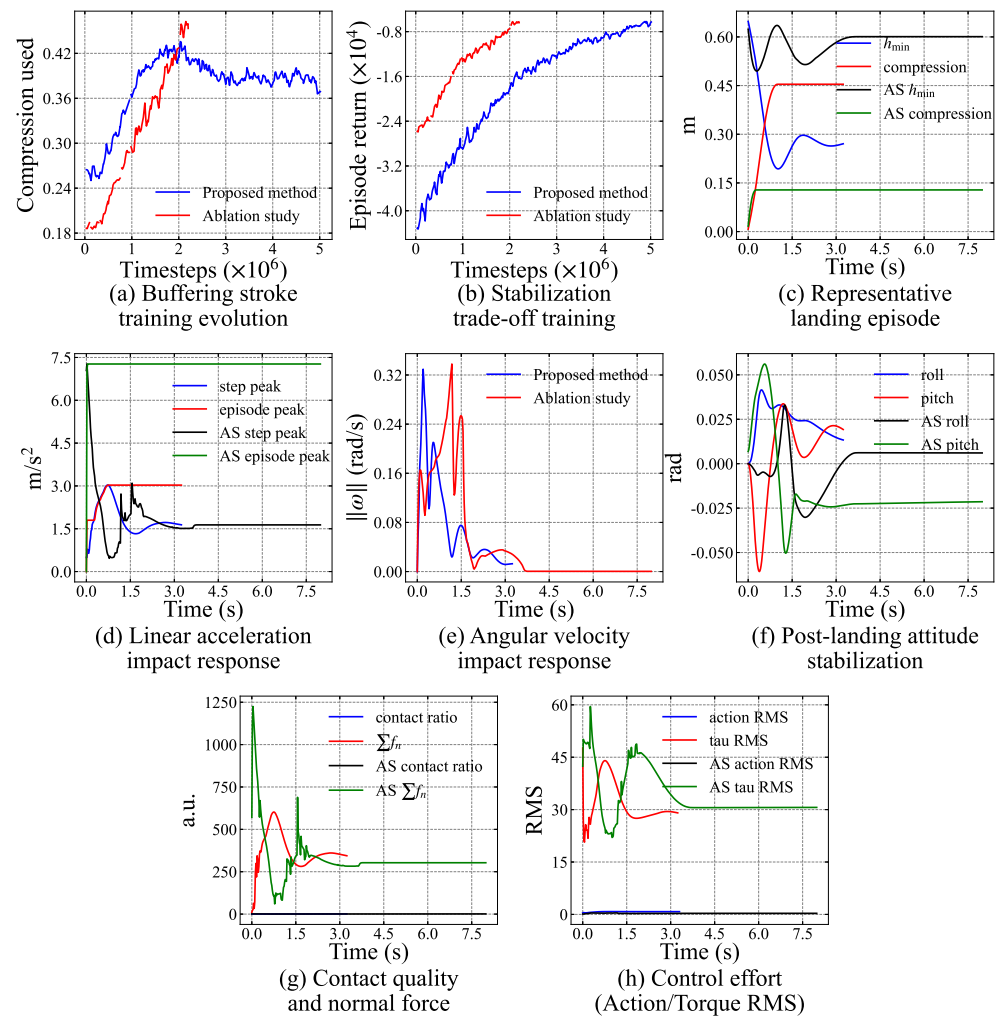


Figure 5. Training evolution of buffering-stroke utilization (compression metric) and stabilization trade-off.

The robustness and safety characteristics of the learned controller are further examined through distribution-level statistics, reward decomposition (as shown in Figure 6), and qualitative landing snapshots.

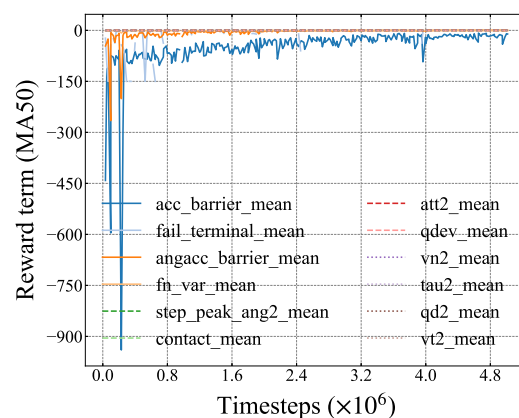


Figure 6. Reward decomposition during training: evolution of major reward terms.

Figure 7 illustrates the empirical cumulative distribution functions (CDFs) of the peak linear acceleration for different impact velocities under three representative base masses. For each fixed mass, the CDF curves shift monotonically to the right as the impact velocity

increases, indicating a predictable and physically consistent increase in peak acceleration. Importantly, the distributions remain narrow for all tested velocities, with no evidence of heavy tails or outliers, suggesting that the policy avoids rare but extreme impact events. As the base mass increases, the entire distribution shifts slightly leftward, reflecting the combined effect of larger inertia and longer effective impact duration. This behavior indicates that the learned controller adapts its force regulation strategy across different mass regimes rather than relying on a single, mass-specific tuning.

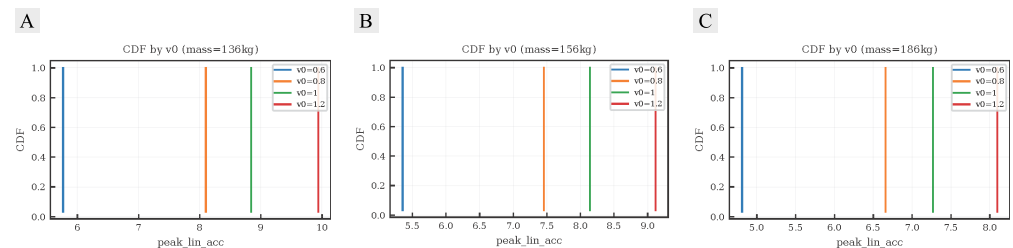


Figure 7. Distributional robustness of impact loads: CDFs of peak linear acceleration across impact velocities and base masses. (A) Base mass = 136 kg; (B) Base mass = 156 kg; (C) Base mass = 186 kg.

To interpret how the policy achieves such behavior, the evolution of individual reward components is examined through reward decomposition. Early in training, safety-related terms—including acceleration barriers, angular acceleration penalties, and failure-related penalties—dominate the overall reward signal, enforcing feasibility and preventing catastrophic landings. As training progresses, these terms converge toward zero, indicating that constraint violations become increasingly rare. Meanwhile, velocity-related penalties and contact-related rewards gradually diminish in magnitude, reflecting improved impact mitigation and more consistent ground engagement. In later stages, smoothness- and effort-related terms, such as joint velocity and torque penalties, become comparatively more influential, suggesting that the policy refines its behavior toward reduced control effort and smoother actuation once basic safety and stability are achieved. This staged evolution confirms that the composite reward function effectively guides the policy from coarse survivability toward fine-grained performance optimization.

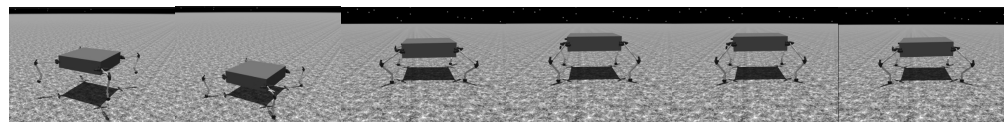
Qualitative landing snapshots provide an intuitive visualization that complements the quantitative analysis. Across a sequence of representative landings, the robot consistently exhibits coordinated leg deployment, progressive compression upon ground contact, and smooth recovery into a stable stance. Even under higher impact velocities, the legs engage nearly simultaneously, and the body orientation remains well regulated without excessive pitching or rolling. These visual observations align with the measured contact ratio, force distribution, and attitude responses, reinforcing the conclusion that the learned policy produces physically plausible and robust landing behaviors rather than exploiting artifacts of the simulation. This is illustrated in Figure 8.

Taken together, the distributional analysis, reward decomposition, and qualitative visualizations demonstrate that the learned controller achieves a favorable balance between safety, robustness, and efficiency. The policy not only performs well on average but also exhibits controlled tail behavior, interpretable reward-driven learning dynamics, and consistent qualitative execution across a wide range of operating conditions. These results further substantiate the suitability of the proposed approach for autonomous legged landing tasks under uncertain mass and impact velocity variations.

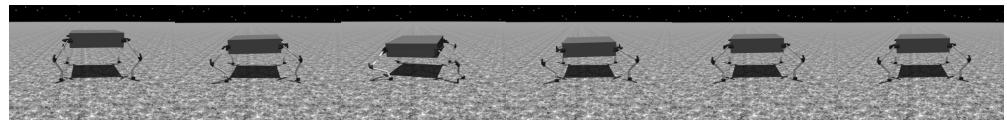
A further comparison with representative previous approaches provides additional context. The reviewer-suggested reference concerns a two-wheeled robot and is therefore not directly comparable to the present quadruped landing problem. Nevertheless, inspired by its benchmarking logic, we compare our results with a recent RL-based quadruped

landing study that reports overload acceleration, final residual velocity, and post-landing attitude criteria. In that reference, soft landing is evaluated using terminal conditions such as pitch angle below 15° , final velocity below 0.05 m/s, and maximum final foot-ground distance below 0.10 m, while the authors also report occasional rebound that causes soft-landing failure in some cases. By contrast, the present framework enforces a stricter stability-window-based criterion with roll/pitch, normal velocity, tangential velocity, angular velocity, minimum clearance, and contact ratio constraints that must be satisfied continuously over a finite horizon. Therefore, the proposed method demonstrates stronger post-impact stabilization capability and a more rigorous notion of successful landing than the compared quadruped RL approach.

A Condition 1: Vertical landing velocity, slope angle 0°



B Condition 2: Vertical landing velocity + horizontal landing velocity, slope angle 0°



C Condition 3: Vertical landing velocity + horizontal landing velocity, slope angle 8°

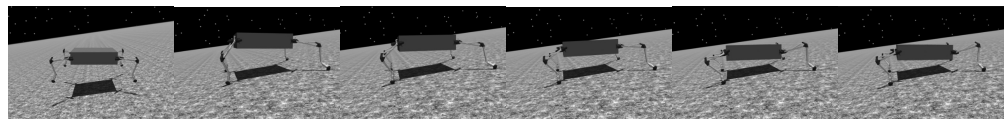


Figure 8. MuJoCo simulation snapshots of the lunar quadruped during RL training (representative landing and stabilization behaviors).

8. Hardware Experiments on a Quadruped Robot

The physical experimental system consists of a lunar quadruped robot, a host computer, a motion capture system (three-dimensional accuracy ± 0.1 mm, maximum frame rate 1000 FPS), an inertial measurement unit (IMU) with gyroscope bias of $1^\circ/\text{h}$ and angular random walk not exceeding $\leq 0.125^\circ/\sqrt{\text{h}}$, accelerometers with bias stability of $10\ \mu\text{g}$, six-axis force sensors, a simulated lunar regolith terrain, and an overhead cable-suspended low-gravity offloading platform, as illustrated in Figure 9. The host computer runs a multi-leg coordinated impedance control algorithm, which generates joint torque commands to drive the motors and achieve precise motion control of the robot. An onboard IMU monitors body attitude variations, while optical markers mounted on the robot surface are tracked in real time by the motion capture system to provide global pose information, forming a closed-loop feedback system that maintains overall stability during landing experiments.

To emulate the low-gravity environment of the lunar surface, an overhead cable-suspended gravity offloading platform is employed. The platform consists of an aluminum frame, low-friction horizontal guide rails, a wire-rope suspension mechanism, and counterweights, enabling partial gravity compensation on the robot body.

Based on the lunar polar soil parameters listed in Table 8, the simulated lunar soil used in the experiments is precisely prepared and mechanically calibrated to ensure that its physical properties match those of real lunar soil. This terrain simulates the soft and uneven mechanical characteristics of the lunar surface, which significantly influence the robot's sinkage behavior, traction performance, and overall locomotion stability, thereby providing highly realistic experimental conditions for validating the foot-ground interaction model and the control algorithm.

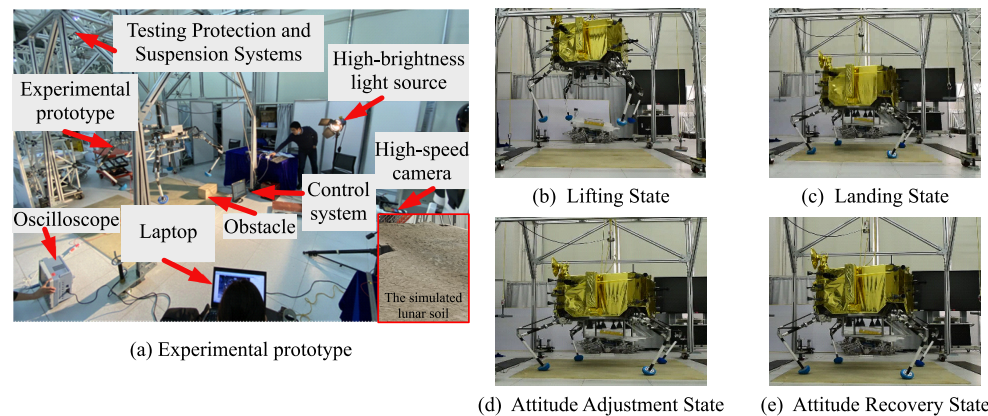


Figure 9. Experimental prototype and landing buffering process of the lunar quadruped robot.

Table 8. Key parameters of the simulated lunar soil.

No.	Parameter	Test Value
1	Bulk density (g/cm^3)	1.4–2.3
2	Deformation index	0.97–1.32
3	Cohesion modulus (kN/m^{n+1})	1.46–4.89
4	Friction modulus (kN/m^{n+2})	281–652
5	Shear deformation modulus (cm)	1.09–1.66
6	Equivalent stiffness modulus (kPa/m^n)	840–2800
7	Contact stiffness (N)	3.2×10^5 – 2.0×10^7 (0–16 wt%)
8	Cohesion (kPa)	34.4–41.9
9	Internal friction angle ($^\circ$)	44.1–48.3
10	Thermal conductivity ($\text{W}/(\text{m} \cdot \text{k})$)	0.0773–0.935 (0–15 wt%)
11	Albedo	0.25–0.65

8.1. Test Condition 1: Total Mass 200 kg, Vertical Landing Velocity 0.3 m/s, Slope Angle 0° , Without Surface Protrusions

Joint response consistency among the four legs is first analyzed using telemetry data from the ground controller. The motor speed and joint torque measurements for all legs, summarized in Tables 9 and 10 and their corresponding plots (as shown in Figure 10a–f), show that the inter-leg motor speed differences remain within 200 rpm, and the joint torque differences remain within $25 \text{ N} \cdot \text{m}$. These results indicate good consistency of leg behavior during flat-ground landing. The observed motor speed differences among the legs arise from a combination of factors, including the initial landing configuration, body attitude angles, and angular velocities at touchdown. Due to different orientations of the leg bases with respect to the gravity direction, slight variations exist in the initial landing configurations under the effective one-sixth gravity condition, while body attitude and angular velocity further influence individual leg touchdown velocities.

Table 9. Peak joint-motor speeds of each leg during buffering under test condition 1.

Buffering Mobile Leg	Motor Speed (rpm)		
	Hip Roll Joint	Hip Pitch Joint	Knee Pitch Joint
M01-1	73	911	545
M01-2	91	721	399
M01-3	3	780	461
M01-4	73	882	527
Maximum	91	911	545

Table 10. Peak joint torques of each leg during buffering under test condition 1.

Buffering Mobile Leg	Joint Torque (N · m)		
	Hip Roll Joint	Hip Pitch Joint	Knee Pitch Joint
M01-1	9.66	75.69	31.82
M01-2	16.73	84.51	31.43
M01-3	3.92	101.5	30.96
M01-4	17.36	96.05	35.87
Maximum	17.36	101.5	35.87

Differences in joint torques among the legs are mainly attributed to gravity-induced disturbance torques. The overall mass balancing of the robot is performed based on the nominal initial configuration of the buffering legs. During the landing buffering process, compression of the buffering legs causes a shift in the system center of mass, generating gravity disturbance torques. As indicated by the experimental configuration shown in Figure 10, the influence of these disturbance torques decreases with increasing installation height of the buffering legs in the inertial frame, resulting in the observed trend of “leg 3 > leg 4 > leg 2 > leg 1.” This effect also influences the magnitude and direction of the hip roll joint torques. The flat-ground experimental data are consistent with expectations, with joint torques—particularly hip pitch torques—of legs 1 and 2 being smaller than those of legs 3 and 4. Similar characteristics are observed in test conditions 2 through 3.

During the landing buffering experiments, the maximum measured motor speed reaches 911 rpm, with a speed rise time of approximately 58 ms. This peak occurs on leg 1 and is related to the landing configuration, body attitude, and angular velocity at touchdown. The maximum joint torque measured is 101.5 N · m, occurring on leg 3 and associated with gravity disturbance torques. Both values satisfy the technical specification limits, including the maximum allowable motor speed of 3600 rpm, and meet the success criteria defined in the experimental protocol.

Minimum ground clearance is analyzed using data from the ground controller. The variation of foot height for each leg during the landing buffering process is summarized in the corresponding figure, as shown in Figure 10g. The minimum foot height during buffering is measured as 789.51 mm. Given the current mechanical design, the distance from the detector base plate to the hip roll axis of the buffering leg is 75.6 mm, resulting in a minimum ground clearance of 713.91 mm between the detector base plate and the ground, which exceeds the required minimum of 200 mm. Throughout the experiment, the safety support rods do not contact the landing surface, indicating that the ground clearance constraint is satisfied and that the experimental success criteria are met.

Touchdown detection signals are analyzed based on telemetry data from the ground controller. The touchdown switch signals, as summarized in Table 11 and Figure 10h, confirm that all touchdown switches successfully detect ground contact and exhibit trigger durations exceeding 10 ms. Due to the rigid contact between the rigid foot pads and the rigid landing surface, brief foot-end rebounds occur at touchdown, resulting in short intervals of “off-on-off” behavior of the touchdown switches, with durations of approximately 4–5 ms. Simulation analyses predict a single-touch trigger duration of approximately 1.6–3.5 ms for a triggering stroke of 2 mm, which is consistent with the experimental observations. Torque-based detection signals are triggered when the touchdown switch remains continuously engaged, indicating correct switching of control parameters during the landing phase.

Load characteristics at key structural locations are evaluated using six-axis force sensor data. The summarized force measurements for each leg during landing buffering, expressed in the leg-base coordinate frame and shown in Table 12 and Figure 10i, indicate that the

maximum vertical force is 332.34 N and the maximum lateral moment is 188.11 N·m. The structural loads at all buffering leg mounting points remain below the specified limit loads, satisfying both technical requirements and experimental success criteria.

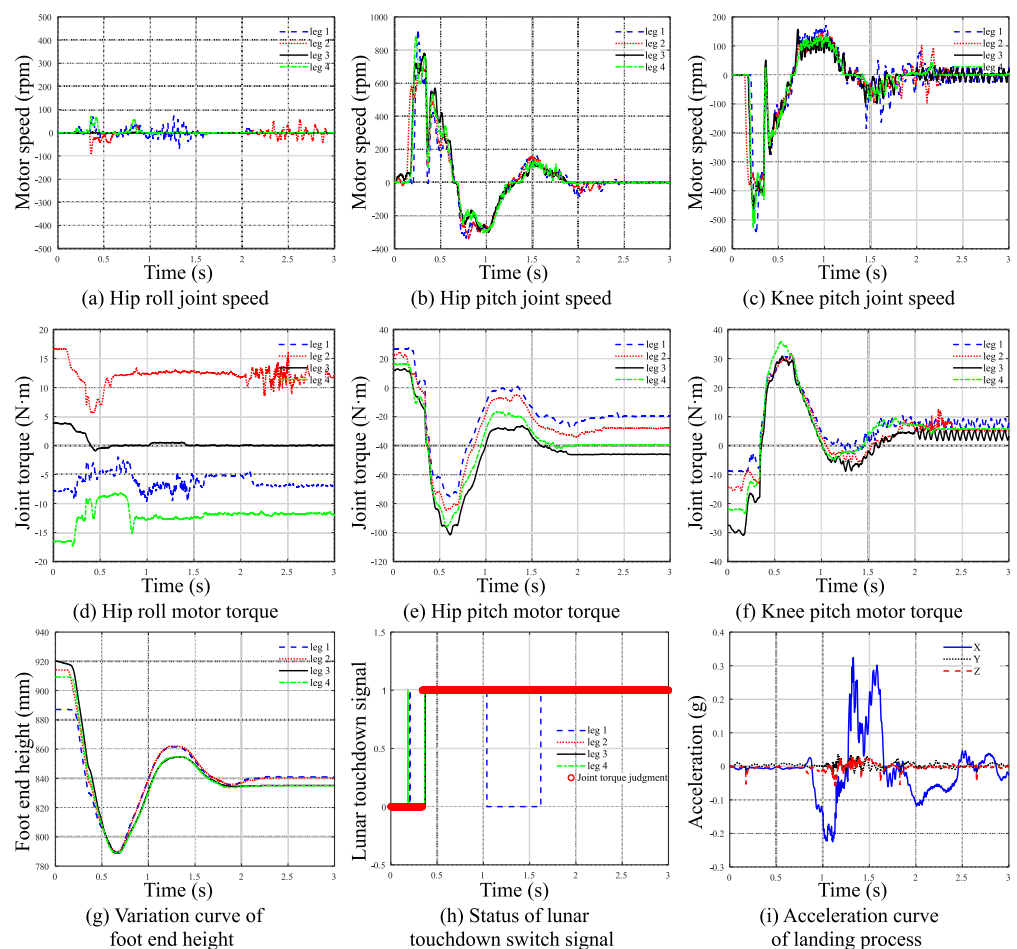


Figure 10. Consistency curves of joint responses of each leg and key measurement results during the landing buffering phase under test condition 1.

Table 11. Touchdown-switch trigger-duration statistics under test condition 1 (including rebound re-triggers).

Buffering Mobile Leg	Duration of the First Touchdown	Duration of the Second Touchdown	Duration of the Third Touchdown
Leg 1	5 ms	667 ms	Duration
Leg 2	Duration	/	/
Leg 3	Duration	/	/
Leg 4	4 ms	Duration	/

Table 12. Six-axis force/torque at buffering-leg mounts (leg-base frame) under test condition 1.

Number	FX (N)	FY (N)	FZ (N)	MX (N·m)	MY (N·m)	MZ (N·m)
Leg 1	36.66	−24.19	273.20	−8.67	−170.60	−9.07
Leg 2	−27.38	16.08	313.29	33.13	−180.58	10.19
Leg 3	−77.62	22.62	312.68	−7.26	−184.85	10.74
Leg 4	−56.51	−35.80	332.34	−11.93	−188.11	−13.75
Maximum absolute value	77.62	35.80	332.34	33.13	188.11	13.75

Acceleration responses during the landing buffering process are obtained from the IMU. The measured acceleration profiles show that the maximum acceleration in the vertical direction reaches 0.32 g, well below the 2 g limit, while the maximum accelerations in the other two directions are 0.06 g and 0.04 g, respectively, both below the 1 g threshold. These results meet the technical specifications and experimental success criteria.

Body attitude variations during landing buffering are also analyzed using IMU data, as shown in Table 13. The measured attitude deviations relative to the initial landing state indicate that roll and pitch angles in all directions remain within 30°, satisfying the stability requirements defined in the experimental protocol.

Table 13. Body-attitude evolution and maximum deviation (pitch/roll/yaw) under test condition 1.

Category	Pitch Angle	Roll Angle	Yaw Angle
Initial landing (°)	22.44	8.70	39.02
Maximum deflection (°)	22.42	8.12	39.73
Landing completion (°)	22.37	8.15	39.65
Maximum deviation (°)	0.07	0.58	0.71

Finally, simulation-to-experiment consistency is assessed by comparing data from leg 1 with corresponding simulation results. Motor speed and joint torque trajectories are presented in Figure 10a–f, and quantitative comparisons are summarized in Table 14. The deviations between simulated and measured peak motor speeds and peak joint torques are within 30%, demonstrating good agreement between simulation predictions and experimental observations. This level of consistency confirms the validity of the proposed real-to-sim modeling approach and satisfies the success criteria specified in the experimental outline.

Table 14. Sim-to-experiment comparison of peak motor speed and joint torque for key joints (hip pitch, knee pitch) under test condition 1.

Joint		Motor Speed (rpm)	Joint Torque (N·m)
Hip pitch joint	Measured maximum value	911	75.69
	Simulated maximum value	884	81.73
	Deviation	26	6.04
Knee pitch joint	Measured maximum value	545	31.82
	Simulated maximum value	551	39.21
	Deviation	7	7.39

8.2. Test Condition 2: Total Mass 250 kg, Vertical Landing Velocity 0.3 m/s, Slope Angle 0°, Without Surface Protrusions

Joint response consistency among the four legs is first examined using telemetry data from the ground controller. The comparative motor speed and joint torque results for all legs, summarized in Tables 15 and 16 together with the corresponding plots (as shown in Figure 11a–f), show that the inter-leg motor speed differences are approximately 200 rpm, while the joint torque differences are approximately 25 N·m. These results indicate good consistency of leg behavior during flat-ground landing under the increased mass condition.

According to single-leg experimental characterization, the expected maximum motor speed during landing buffering should fall within the range of 800–1200 rpm. In the present test, the maximum motor speed recorded during the landing buffering process is 1175 rpm, with a rise time of approximately 55 ms. This peak occurs on leg 3 and is related to the landing configuration, initial posture, body attitude, and angular velocity at touchdown. The maximum joint torque measured is 114.19 N·m, also occurring on leg 3 and primarily attributed to gravity-induced disturbance torque. These values are consistent with expectations and satisfy the technical specification limits, including the

maximum allowable motor speed of 3600 rpm, as well as the success criteria defined in the experimental protocol.

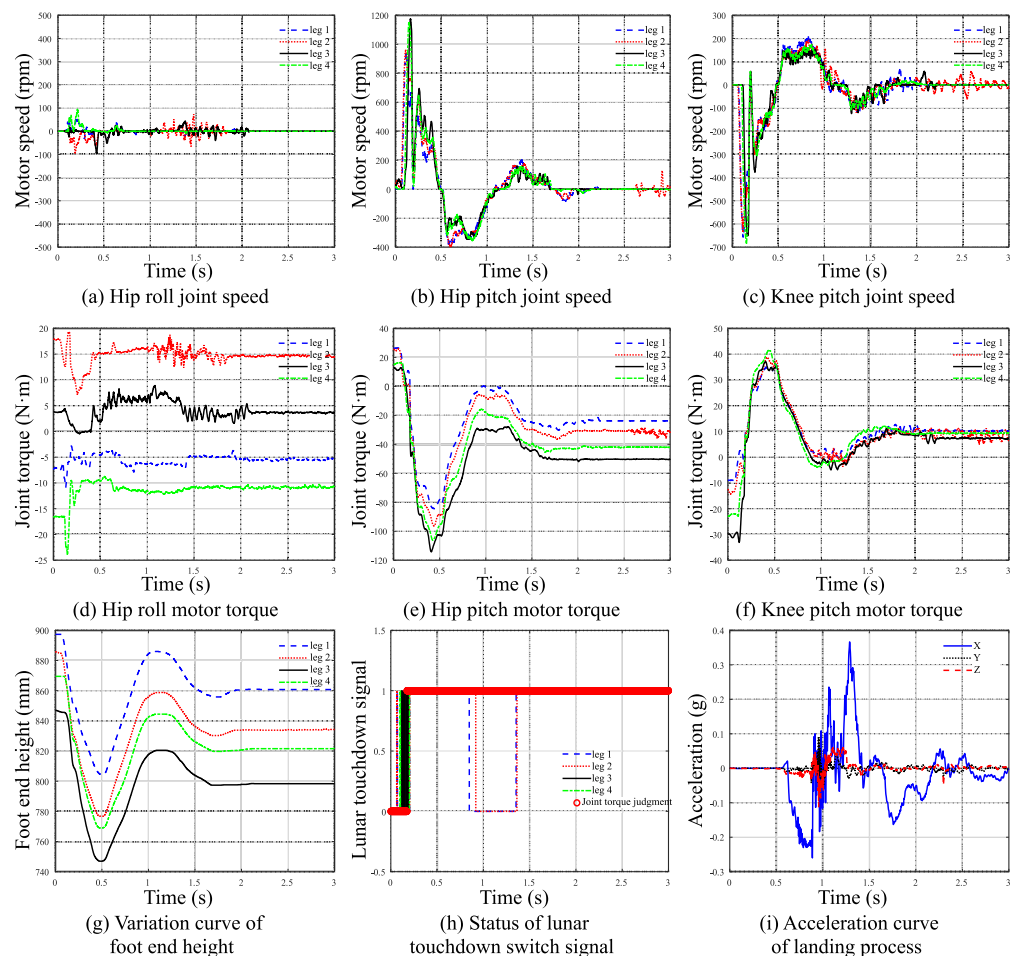


Figure 11. Consistency curves of joint responses of each leg and key measurement results during the landing buffering phase under test condition 2.

Table 15. Peak joint-motor speeds of each leg during buffering under test condition 2.

Buffering Mobile Leg	Motor Speed (rpm)		
	Hip Roll Joint	Hip Pitch Joint	Knee Pitch Joint
M01-1	51	944	659
M01-2	95	966	637
M01-3	98	1175	651
M01-4	95	1153	684
Maximum	98	1175	684

Table 16. Peak joint torques of each leg during buffering under test condition 2.

Buffering Mobile Leg	Joint Torque (N·m)		
	Hip Roll Joint	Hip Pitch Joint	Knee Pitch Joint
M01-1	10.91	84.44	35.96
M01-2	19.35	96.4	38.65
M01-3	8.91	114.19	37.15
M01-4	24.02	106.1	41.53
Maximum	24.02	114.19	41.53

Minimum ground clearance is analyzed based on telemetry data from the ground controller. The variations of foot height for all legs during landing buffering are summarized in the corresponding figure, as shown in Figure 11g. The minimum foot height observed during buffering is 746.98 mm. Given the current mechanical design, where the distance between the detector base plate and the hip roll axis of the buffering leg is 75.6 mm, the minimum ground clearance of the detector base plate is calculated as 671.38 mm, which exceeds the required minimum clearance of 200 mm. Throughout the experiment, the safety support rods do not contact the landing surface, confirming compliance with both technical requirements and experimental success criteria.

Touchdown detection signals are analyzed using data from the ground controller. The touchdown switch signals, as summarized in Table 17 and Figure 11h, show that all touchdown switches successfully detect ground contact and exhibit trigger durations of at least 10 ms. Due to the rigid contact between the rigid foot pads and the rigid landing surface, brief foot-end rebounds occur, resulting in short “off-on-off” transitions of the touchdown switches. The torque-based detection signal is triggered when the touchdown switch remains continuously engaged, indicating correct switching of control parameters during the landing phase.

Table 17. Touchdown-switch trigger-duration statistics under test condition 2 (including rebound re-triggers).

Buffering Mobile Leg	Duration of the First Touchdown	Duration of the Second Touchdown	Duration of the Third Touchdown	Duration of the Fourth Touchdown	Duration of the Fifth Touchdown
Leg 1	4 ms	644 ms	Duration	/	/
Leg 2	4 ms	713 ms	Duration	/	/
Leg 3	4 ms	4 ms	18 ms	3 ms	Duration
Leg 4	4 ms	Duration	/		

Load characteristics at key structural locations are evaluated using six-axis force sensor data. The summarized measurements for each leg during landing buffering, expressed in the leg-base coordinate frame and reported in Table 18, show that the structural loads at all buffering leg mounting points remain below the specified limit loads. These results satisfy the technical requirements and the success criteria of the experimental outline.

Table 18. Six-axis force/torque at buffering-leg mounts (leg-base frame) under test condition 2.

Number	FX (N)	FY (N)	FZ (N)	MX (N·m)	MY (N·m)	MZ (N·m)
Leg 1	40.47	−26.1	317.7	−9.55	−187.47	−8.02
Leg 2	35.84	−22.89	353.34	25.62	−197.34	−9.99
Leg 3	−94.09	33.34	345.5	7.98	−200.17	13.08
Leg 4	−72.14	−41.04	372.29	−17.94	−202.79	−15.09
Maximum absolute value	94.09	41.04	372.29	25.62	202.79	15.09

Acceleration responses during the landing buffering process are obtained from the inertial navigation system, as shown in Figure 12. The measured acceleration profiles indicate that the maximum acceleration in the vertical direction reaches 0.37 g, which is well below the 2 g limit, while the maximum accelerations in the other two directions are 0.11 g and 0.09 g, respectively, both below the 1 g threshold. These results meet the technical specifications and the experimental success criteria.

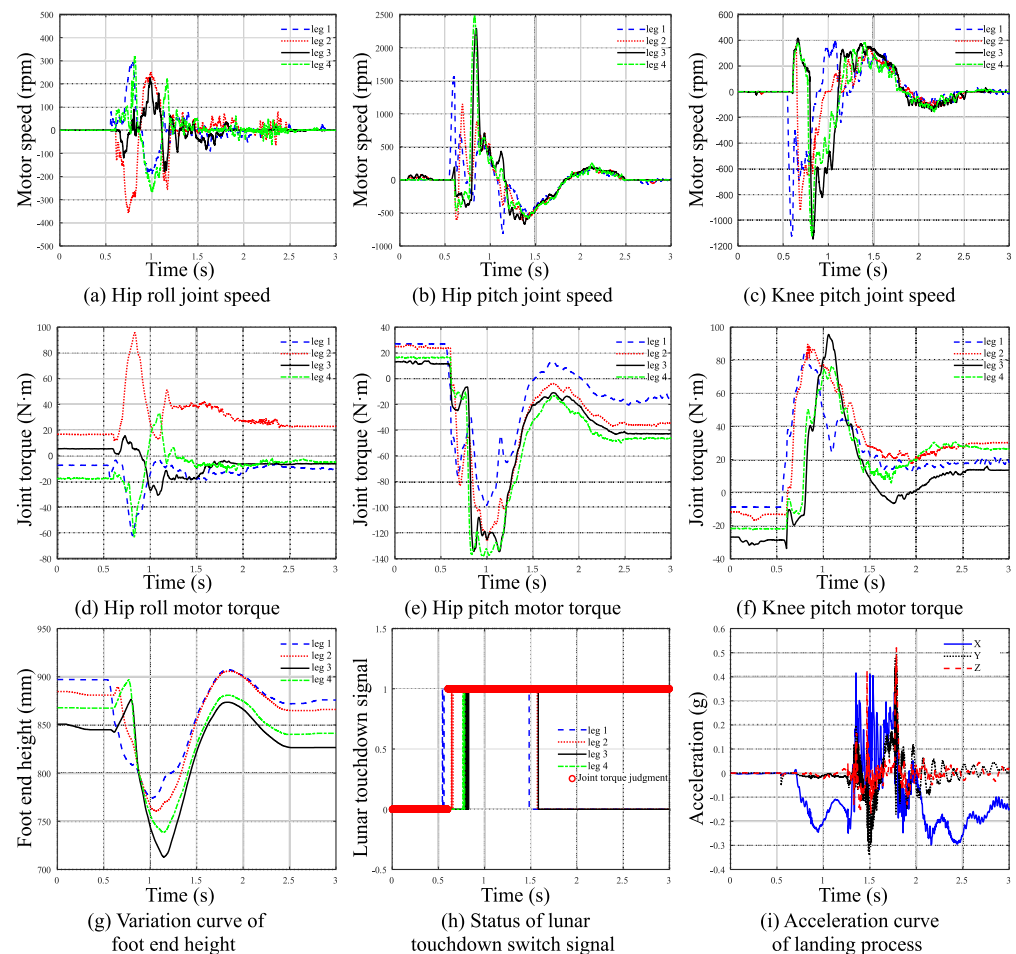


Figure 12. Consistency curves of joint responses of each leg and key measurement results during the landing buffering phase under test condition 3.

Body attitude variations during landing buffering are also analyzed using inertial measurement data, as shown in Table 19. The measured attitude deviations relative to the initial landing state show that roll and pitch angles in all directions remain within 30° , satisfying the stability requirements defined in the experimental protocol.

Table 19. Sim-to-experiment comparison of peak motor speed and joint torque for key joints (hip pitch, knee pitch) under test condition 2.

Joint		Motor Speed (rpm)	Joint Torque (N·m)
Hip pitch joint	Measured maximum value	944	84.44
	Simulated maximum value	1063	72.81
	Deviation	120	11.63
Knee pitch joint	Measured maximum value	659	35.96
	Simulated maximum value	647	22.26
	Deviation	11	13.69

Finally, simulation-to-experiment consistency is evaluated by comparing data from leg 1 in Test Condition 1 with the corresponding simulation results. Motor speed and joint torque trajectories are presented in Figure 11a–f, and quantitative comparisons are summarized in Table 20. The deviations between simulated and measured peak motor speeds and peak joint torques are within 30%, demonstrating good agreement between simulation predictions and experimental observations. This level of consistency confirms that the simulation model provides a reliable representation of the physical system and meets the success criteria specified in the experimental outline.

Table 20. Body-attitude evolution and maximum deviation (pitch/roll/yaw) under test condition 2.

Category	Pitch Angle	Roll Angle	Yaw Angle
Initial landing (°)	23.00	9.13	30.63
Maximum deflection (°)	22.96	8.24	31.31
Landing completion (°)	22.96	8.34	31.31
Maximum deviation (°)	0.04	0.89	0.68

8.3. Test Condition 3: Total Mass 250 kg, Vertical Landing Velocity 0.8 m/s, Horizontal Landing Velocity 0.2 m/s, Slope Angle 8°, Without Surface Protrusions

Under this condition, leg 1 is located on the uphill side of the slope, while leg 3 is located on the downhill side.

Joint response characteristics of all legs are analyzed using telemetry data from the ground controller. The comparative motor speed and joint torque results, summarized in Tables 21 and 22 together with the corresponding plots (as shown in Figure 12a–f), indicate that leg 1 establishes ground contact first during landing. The maximum motor speed recorded during the landing buffering process reaches 2482 rpm, with a rise time of approximately 52 ms. This peak occurs on leg 4, which is located on the downhill side of the slope. The maximum joint torque measured is 138.06 N·m, occurring on leg 3, also on the downhill side. All measured values satisfy the technical specification limits, including the maximum allowable motor speed of 3600 rpm, and meet the success criteria defined in the experimental protocol.

Table 21. Peak joint-motor speeds of each leg during buffering under test condition 3.

Buffering Mobile Leg	Motor Speed (rpm)		
	Hip Roll Joint	Hip Pitch Joint	Knee Pitch Joint
M01-1	300	1585	1124
M01-2	355	1149	919
M01-3	227	2288	1146
M01-4	318	2482	1124
Maximum	355	2482	1146

Table 22. Peak joint torques of each leg during buffering under test condition 3.

Buffering Mobile Leg	Joint Torque (N·m)		
	Hip Roll Joint	Hip Pitch Joint	Knee Pitch Joint
M01-1	62.1	98.58	85.56
M01-2	95.75	126.15	89.69
M01-3	30.71	138.06	95.65
M01-4	63.1	134.47	76.31
Maximum	95.75	138.06	95.65

Minimum ground clearance is evaluated based on telemetry data from the ground controller. The variations of foot-end height for all legs during landing buffering are summarized in the corresponding figure, as shown in Figure 12g. The minimum foot height observed during buffering is 712.96 mm. Given the current mechanical design, in which the distance between the detector base plate and the hip roll axis of the buffering leg is 75.6 mm, the minimum ground clearance of the detector base plate is calculated as 637.36 mm, which exceeds the required minimum clearance of 200 mm. Throughout the experiment, the safety protection support rods do not contact the landing surface, confirming compliance with technical requirements and experimental success criteria.

Touchdown detection signals are analyzed using data from the ground controller. The touchdown switch signals, summarized in Table 23 and Figure 12h, show that all touchdown switches successfully detect ground contact and exhibit trigger durations of at least 10 ms. Due to rigid contact between the rigid foot pads and the rigid landing surface, brief foot-end rebounds occur at touchdown, resulting in short “off–on–off” transitions of the touchdown switches. After completion of the landing process, the touchdown switches of legs 1, 2, and 3 are not triggered. The torque-based detection signal is activated when the touchdown switch remains continuously engaged, indicating correct switching of control parameters during the landing phase.

Table 23. Touchdown-switch trigger-duration statistics under test condition 3 (including rebound re-triggers).

Buffering Mobile Leg	Duration of the First Touchdown	Duration of the Second Touchdown	Duration of the Third Touchdown	Duration of the Fourth Touchdown	Duration of the Fifth Touchdown
Leg 1	3 ms	921 ms	/	/	/
Leg 2	4 ms	912 ms	/	/	/
Leg 3	4 ms	3 ms	4 ms	754 ms	/
Leg 4	5 ms	Duration	/	/	/

Acceleration responses during the landing buffering process are obtained from the inertial navigation system, as shown in Figure 12i. The measured acceleration profiles indicate that the maximum acceleration in the vertical direction reaches 0.42 g, which is well below the 2 g limit, while the maximum accelerations in the other two directions are 0.49 g and 0.52 g, respectively, both below the 1 g threshold. These results satisfy the technical specifications and experimental success criteria.

Load characteristics at key structural locations are evaluated using six-axis force sensor data. The summarized measurements for each leg during landing buffering, expressed in the leg-base coordinate frame and reported in Table 24, show that the structural loads at all buffering leg mounting points remain below the specified limit loads. These results meet the technical requirements and the success criteria defined in the experimental outline.

Table 24. Six-axis force/torque at buffering-leg mounts (leg-base frame) under test condition 3.

Number	FX (N)	FY (N)	FZ (N)	MX (N·m)	MY (N·m)	MZ (N·m)
Leg 1	−315.73	110.64	528.84	82.06	−216.11	49.02
Leg 2	−260.42	−208.21	549.78	63.32	−227.44	−102.15
Leg 3	−317.58	222.36	539.44	54.66	−288.84	75.16
Leg 4	−227.97	283.05	596.76	83.51	−337.47	−101.14
Maximum absolute value	317.58	283.05	596.76	63.32	337.47	102.15

Body attitude variations during landing buffering are analyzed using inertial measurement data, as shown in Table 25. The measured attitude deviations relative to the initial landing state show that roll and pitch angles in all directions remain within 30°, satisfying the stability requirements specified in the experimental protocol.

Finally, simulation-to-experiment consistency is evaluated by comparing data from leg 1 in Test Condition 1 with the corresponding simulation results. Motor speed and joint torque trajectories are presented in Figure 12a–f, and quantitative comparisons are summarized in Table 25. The deviations between simulated and measured peak motor speeds and peak joint torques are within 30%, demonstrating good agreement between

simulation predictions and experimental observations. This level of consistency confirms that the simulation model provides a reliable representation of the physical system and meets the success criteria specified in the experimental outline.

Table 25. Body-attitude evolution and maximum deviation (pitch/roll/yaw) under test condition 3.

Category	Pitch Angle	Roll Angle	Yaw Angle
Initial landing (°)	21.82	9.27	−94.98
Maximum deflection (°)	21.86	2.22	−92.35
Landing completion (°)	22.38	3.15	−92.89
Maximum deviation (°)	0.56	7.05	2.63

9. Conclusions

This paper presents a reinforcement learning-based landing impact mitigation and stabilization control framework for a lunar quadruped robot operating under complex terrain conditions. To address challenges arising from weak gravity, large mass variation, strict structural constraints, and uncertain contact properties, a real-to-sim dynamic modeling pipeline is established and validated through experimental data. Based on this foundation, a terrain-agnostic control formulation is developed using equivalent support–direction-based velocity decomposition, together with episode-level peak impact modeling, contact-gated anti-slip regulation, and stability-window-based success criteria, enabling physically interpretable and robust end-to-end policy learning without explicit phase switching.

Extensive simulations and full-scale hardware experiments demonstrate that the proposed approach effectively suppresses landing impact peaks, maintains posture stability, and satisfies engineering safety constraints across a wide range of masses, initial velocities, and slope conditions. The close agreement between simulation and experimental results further confirms the reliability of the real-to-sim modeling and sim-to-real transfer strategy, indicating that the proposed method provides a practical and deployable solution for safe lunar landing of quadruped robots.

Future work will focus on extending the proposed framework toward landing-walking integration, enabling a unified control architecture that seamlessly bridges the impact mitigation phase and subsequent planetary exploration. A key direction is the development of whole-body control strategies that coordinate landing buffering, posture stabilization, and terrain-adaptive locomotion within a single reinforcement learning framework, eliminating the need for mode switching between distinct task phases. Additionally, we plan to incorporate more comprehensive domain randomization, including variable friction and deformable terrain models, to further enhance robustness. The integration of proprioceptive and exteroceptive sensing, such as vision and LiDAR, will also be explored to enable terrain-aware landing site selection and adaptive gait generation in unstructured lunar environments.

Author Contributions: Conceptualization, J.L. and S.S.; methodology, J.L.; software, J.L.; validation, J.L. and Z.L.; formal analysis, J.L. and Y.Y.; investigation, J.L. and Y.Y.; resources, J.L. and Y.Y.; data curation, J.L. and Y.Y.; writing—original draft preparation, J.L.; writing—review and editing, J.L.; visualization, J.L.; supervision, J.L.; project administration, J.L.; funding acquisition, S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by Civil Aerospace Technology Advanced Research Program during the 14th Five-Year Plan Period (No. D010107), the National Natural Science Foundation of China (No. U22B2080), and the Heilongjiang Provincial Natural Science Foundation of China (No. JJ2024LH0935).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sang, H.; Wang, S. Lunar Leap Robot: 3M Architecture-Enhanced Deep Reinforcement Learning Method for Quadruped Robot Jumping in Low-Gravity Environment. *J. Aerosp. Eng.* **2024**, *37*, 04024076. [\[CrossRef\]](#)
2. Rudin, N.; Kolvenbach, H.; Tsounis, V.; Hutter, M. Cat-Like Jumping and Landing of Legged Robots in Low Gravity Using Deep Reinforcement Learning. *IEEE Trans. Robot.* **2022**, *38*, 317–328. [\[CrossRef\]](#)
3. Lee, J.; Hwangbo, J.; Wellhausen, L.; Koltun, V.; Hutter, M. Learning quadrupedal locomotion over challenging terrain. *Sci. Robot.* **2020**, *5*, eabc5986. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Dong, Y.; Ding, J.; Wang, C.; Wang, H.; Liu, X. Soft landing stability analysis of a Mars lander under uncertain terrain. *Chin. J. Aeronaut.* **2022**, *35*, 377–388. [\[CrossRef\]](#)
5. Kim, Y.B.; Jeong, H.J.; Park, S.M.; Lim, J.H.; Lee, H.H. Prediction and Validation of Landing Stability of a Lunar Lander by a Classification Map Based on Touchdown Landing Dynamics' Simulation Considering Soft Ground. *Aerospace* **2021**, *8*, 380. [\[CrossRef\]](#)
6. Zhu, J.; Ma, J.; Chen, J.; Wang, C.; Li, Y.; Fan, Z.; Lu, C. Improving landing stability and terrain adaptability in Lunar exploration with biomimetic lander design and control. *Acta Astronaut.* **2025**, *226*, 860–875. [\[CrossRef\]](#)
7. Xin, G.; Zeng, F.; Qin, K. Loco-Manipulation Control for Arm-Mounted Quadruped Robots: Dynamic and Kinematic Strategies. *Machines* **2022**, *10*, 719. [\[CrossRef\]](#)
8. Ji, S.; Liang, S. DEM-FEM-MBD coupling analysis of landing process of lunar lander considering landing mode and buffering mechanism. *Adv. Space Res.* **2021**, *68*, 1627–1643. [\[CrossRef\]](#)
9. Lynch, D.J.; Lynch, K.M.; Umbanhowar, P.B. The Soft-Landing Problem: Minimizing Energy Loss by a Legged Robot Impacting Yielding Terrain. *IEEE Robot. Autom. Lett.* **2020**, *5*, 3658–3665. [\[CrossRef\]](#)
10. Kiefer, J.; Ward, M.; Costello, M. Rotorcraft Hard Landing Mitigation Using Robotic Landing Gear. *J. Dyn. Syst. Meas. Control* **2016**, *138*, 031003. [\[CrossRef\]](#)
11. You, Y.; Yang, Z.; Zou, T.; Sui, Y.; Xu, C.; Zhang, C.; Xu, H.; Zhang, Z.; Han, J. A New Trajectory Tracking Control Method for Fully Electrically Driven Quadruped Robot. *Machines* **2022**, *10*, 292. [\[CrossRef\]](#)
12. Ding, Y.; Pandala, A.; Li, C.; Shin, Y.H.; Park, H.W. Representation-Free Model Predictive Control for Dynamic Motions in Quadrupeds. *IEEE Trans. Robot.* **2021**, *37*, 1154–1171. [\[CrossRef\]](#)
13. Van Hauwermeiren, T.; Coene, A.; Crevecoeur, G. Tactile Force Sensing for Admittance Control on a Quadruped Robot. *Machines* **2025**, *13*, 426. [\[CrossRef\]](#)
14. Garaffa, L.C.; Basso, M.; Konzen, A.A.; de Freitas, E.P. Reinforcement Learning for Mobile Robotics Exploration: A Survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 3796–3810. [\[CrossRef\]](#)
15. Liang, J.; Tang, S.; Jia, B. Control of Parallel Quadruped Robots Based on Adaptive Dynamic Programming Control. *Machines* **2024**, *12*, 875. [\[CrossRef\]](#)
16. Wang, J.; Hu, C.; Zhu, Y. CPG-Based Hierarchical Locomotion Control for Modular Quadrupedal Robots Using Deep Reinforcement Learning. *IEEE Robot. Autom. Lett.* **2021**, *6*, 7193–7200. [\[CrossRef\]](#)
17. Hwangbo, J.; Lee, J.; Dosovitskiy, A.; Bellicoso, D.; Tsounis, V.; Koltun, V.; Hutter, M. Learning agile and dynamic motor skills for legged robots. *Sci. Robot.* **2019**, *4*, eaau5872. [\[CrossRef\]](#)
18. Shao, Y.; Jin, Y.; Liu, X.; He, W.; Wang, H.; Yang, W. Learning Free Gait Transition for Quadruped Robots Via Phase-Guided Controller. *IEEE Robot. Autom. Lett.* **2022**, *7*, 1230–1237. [\[CrossRef\]](#)
19. Aractingi, M.; Léziart, P.A.; Flayols, T.; Perez, J.; Silander, T.; Souères, P. Controlling the solo12 quadruped robot with deep reinforcement learning. *Sci. Rep.* **2023**, *13*, 11945. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Huang, S.; Xiao, Z.; Zheng, M.; Shi, W. Hierarchical reinforcement learning for enhancing stability and adaptability of hexapod robots in complex terrains. *Biomim. Intell. Robot.* **2025**, *5*, 100231. [\[CrossRef\]](#)
21. Qi, J.; Gao, H.; Su, H.; Huo, M.; Yu, H.; Deng, Z. Reinforcement Learning and Sim-to-Real Transfer of Reorientation and Landing Control for Quadruped Robots on Asteroids. *IEEE Trans. Ind. Electron.* **2024**, *71*, 14392–14400. [\[CrossRef\]](#)
22. Qi, J.; Gao, H.; Su, H.; Han, L.; Su, B.; Huo, M.; Yu, H.; Deng, Z. Reinforcement learning-based stable jump control method for asteroid-exploration quadruped robots. *Aerosp. Sci. Technol.* **2023**, *142*, 108689. [\[CrossRef\]](#)
23. Morente-Molinera, J.A.; Wang, Y.; Gong, Z.W.; Morfeq, A.; Al-Hmouz, R.; Herrera-Viedma, E. Reducing Criteria in Multicriteria Group Decision-Making Methods Using Hierarchical Clustering Methods and Fuzzy Ontologies. *IEEE Trans. Fuzzy Syst.* **2022**, *30*, 1585–1598. [\[CrossRef\]](#)

24. Scorsoglio, A.; D'Ambrosio, A.; Ghilardi, L.; Gaudet, B.; Curti, F.; Furfaro, R. Image-Based Deep Reinforcement Meta-Learning for Autonomous Lunar Landing. *J. Spacecr. Rocket.* **2022**, *59*, 153–165. [\[CrossRef\]](#)
25. Scorsoglio, A.; Gaudet, B.; Ghilardi, L.; Furfaro, R. Meta-reinforcement learning guidance, navigation, and control for autonomous lunar landing with safe site selection. *Neural Comput. Appl.* **2025**, *37*, 17311–17340. [\[CrossRef\]](#)
26. Xiao, H.; Gong, Y.; Mei, J.; Wu, Z.; Ma, G.; Wu, W. Residual-learning-based landing control with gravity estimation for quadruped robot in low-gravity scenarios. *Astrodynamics* **2026**, 1–14. [\[CrossRef\]](#)
27. Yang, X.; Wen, T.; Zhang, K.; Yu, Y.; Qiao, D.; Zeng, X. Landing Dynamics of Telescopic-Legged Bionic Rover on Asteroid Gravel Surface Using Discrete Element Method. *J. Field Robot.* **2026**, *43*, 1091–1110. [\[CrossRef\]](#)
28. Wang, L.; Meng, F.; Kang, R.; Sato, R.; Chen, X.; Yu, Z.; Ming, A.; Huang, Q. Design and Implementation of Symmetric Legged Robot for Highly Dynamic Jumping and Impact Mitigation. *Sensors* **2021**, *21*, 6885. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Hoseinifard, S.M.; Sadedel, M. Standing balance of single-legged hopping robot model using reinforcement learning approach in the presence of external disturbances. *Sci. Rep.* **2024**, *14*, 32036. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Tanaka, T.; Malki, H.; Cescon, M. Linear Quadratic Tracking With Reinforcement Learning Based Reference Trajectory Optimization for the Lunar Hopper in Simulated Environment. *IEEE Access* **2021**, *9*, 162973–162983. [\[CrossRef\]](#)
31. Chen, Z.; Shen, S.; Cui, H.; Tian, Y. Robust adaptive guidance for autonomous asteroid landing via search-based meta-reinforcement learning. *Acta Astronaut.* **2025**, *236*, 723–734. [\[CrossRef\]](#)
32. Panichi, E.; Ding, J.; Atanassov, V.; Yang, P.; Kober, J.; Pan, W.; Santina, C.D. On-the-Fly Jumping With Soft Landing: Leveraging Trajectory Optimization and Behavior Cloning. *IEEE/ASME Trans. Mechatron.* **2025**, *30*, 3142–3151. [\[CrossRef\]](#)
33. Li, J.; Zhao, W.; Chen, L.; Liu, Z.; Sun, S. Reinforcement Learning-Based Locomotion Control for a Lunar Quadruped Robot Considering Space Lubrication Conditions. *Mathematics* **2026**, *14*, 848. [\[CrossRef\]](#)
34. Chen, R.; Yang, H.; Feng, Q.; Bai, L.; Liu, L.; Yuan, Z.; Wang, H.; Liang, L.; Jiang, P.; Luo, J. Human-Inspired Foot-Spine Coordination Control for Stable Landing of Jumping Robots. *Chin. J. Mech. Eng.* **2025**, *38*, 178. [\[CrossRef\]](#)
35. Sun, Z.; Zhao, J.; Li, Y.; Teng, L. Robotic leaping enhanced by thrust-induced hypogravity, achieving precise, predictable, and extended jumps. *Nat. Commun.* **2026**, *17*, 2523. [\[CrossRef\]](#)
36. Zhao, W.; Liu, H.; Lewis, F.L. Robust Formation Control for Cooperative Underactuated Quadrotors via Reinforcement Learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 4577–4587. [\[CrossRef\]](#)
37. Fan, R.; Chen, X.; Liu, M.; Cao, X. Attitude-orbit coupled sliding mode tracking control for spacecraft formation with event-triggered transmission. *ISA Trans.* **2022**, *124*, 338–348. [\[CrossRef\]](#)
38. Zhu, A.; Ai, H.; Chen, L. A Fuzzy Logic Reinforcement Learning Control with Spring-Damper Device for Space Robot Capturing Satellite. *Appl. Sci.* **2022**, *12*, 2662. [\[CrossRef\]](#)
39. Jendoubi, I.; Bouffard, F. Multi-agent hierarchical reinforcement learning for energy management. *Appl. Energy* **2023**, *332*, 120500. [\[CrossRef\]](#)
40. Qi, J.; Gao, H.; Yu, H.; Huo, M.; Feng, W.; Deng, Z. Integrated attitude and landing control for quadruped robots in asteroid landing mission scenarios using reinforcement learning. *Acta Astronaut.* **2023**, *204*, 599–610. [\[CrossRef\]](#)
41. Choi, S.; Ji, G.; Park, J.; Kim, H.; Mun, J.; Lee, J.H.; Hwangbo, J. Learning quadrupedal locomotion on deformable terrain. *Sci. Robot.* **2023**, *8*, eade2256. [\[CrossRef\]](#) [\[PubMed\]](#)
42. Shi, Y.; He, X.; Zou, W.; Yu, B.; Yuan, L.; Li, M.; Pan, G.; Ba, K. Multi-Objective Optimal Torque Control with Simultaneous Motion and Force Tracking for Hydraulic Quadruped Robots. *Machines* **2022**, *10*, 170. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.