PRIOR: Perceptive Learning for Humanoid Locomotion with Reference Gait Priors

Abstract— **Training perceptive humanoid locomotion policies that traverse complex terrains with natural gaits remains an open challenge, typically demanding multi-stage training pipelines, adversarial objectives, or extensive real-world calibration. We present PRIOR, an efficient and reproducible framework built on Isaac Lab that achieves robust terrain traversal with human-like gaits through a simple yet effective design: (i) a parametric gait generator that supplies stable reference trajectories derived from motion capture without adversarial training, (ii) a GRU-based state estimator that infers terrain geometry directly from egocentric depth images via self-supervised heightmap reconstruction, and (iii) terrain-adaptive footstep rewards that guide foot placement toward traversable regions. Through systematic analysis of depth image resolution trade-offs, we identify configurations that maximize terrain fidelity under real-time constraints, substantially reducing perceptual overhead without degrading traversal performance. Comprehensive experiments across terrains of varying difficulty—including stairs, boxes, and gaps—demonstrate that each component yields complementary and essential performance gains, with the full framework achieving a 100% traversal success rate. We will open-source the complete PRIOR framework, including the training pipeline, parametric gait generator, and evaluation benchmarks, to serve as a reproducible foundation for humanoid locomotion research on Isaac Lab.**

I. INTRODUCTION

Deploying humanoid robots in human-centric environments requires locomotion controllers that can negotiate diverse terrain geometries—stairs, boxes, gaps—while preserving the natural bipedal gaits essential for safe and predictable coexistence with people and infrastructure. Reinforcement learning (RL) has emerged as the dominant paradigm for training such controllers, with recent work demonstrating impressive results in either terrain-robust locomotion through perceptive policies [1], [2] or natural gait synthesis through motion priors [3], [4]. Achieving both capabilities simultaneously, however, typically incurs substantial system complexity: multi-stage teacher–student distillation, adversarial discriminators for style enforcement, or extensive sim-to-real calibration. These requirements hinder reproducibility and raise the barrier to entry for locomotion research. In this work, we ask: *can a single-stage RL pipeline, without adversarial training or distillation, produce perceptive humanoid locomotion that is both terrain-robust and natural?*

We answer this question with **PRIOR**, a framework whose design is guided by three observations drawn from recent literature.

(i) LiDAR-based elevation mapping, while geometrically precise, relies on odometric integration that accumulates drift during extended locomotion; egocentric depth images, by contrast, provide a self-contained terrain signal that is inherently free of such drift.

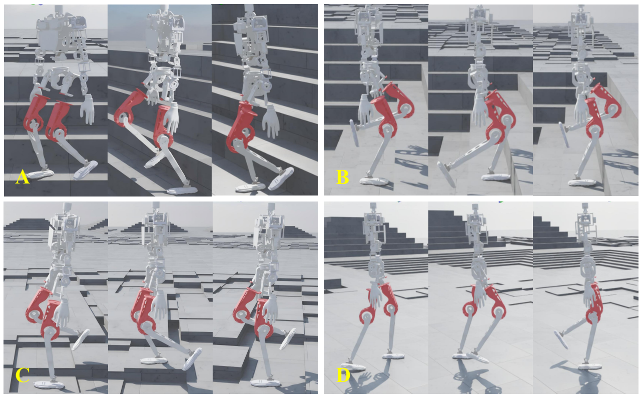


Fig. 1: The proposed PRIOR framework was simulated and demonstrated on the ZERITH Z1 model. (A)–(D) illustrate the robot traversing four representative terrain types.

(ii) Adversarial motion priors, the prevailing mechanism for imposing gait style, suffer from well-documented training pathologies—mode collapse, reward ambiguity, and hyperparameter sensitivity — that are exacerbated on challenging terrains where the policy must deviate significantly from reference motions; a parametric gait generator can offer comparable stylistic guidance through deterministic supervision, sidestepping these instabilities entirely.

(iii) The ongoing transition from Isaac Gym to Isaac Lab as the community-standard simulation platform calls for training pipelines natively built on the newer stack; we develop PRIOR entirely within Isaac Lab, incorporating systematic optimizations that yield a 3× speedup over the vanilla Isaac Lab baseline.

Concretely, PRIOR trains a single locomotion policy end-to-end through three mutually reinforcing mechanisms. A *parametric gait generator* produces phase-conditioned joint trajectories by dynamically blending motion capture primitives, supplying the policy with velocity-adaptive motion targets that replace adversarial style losses. A *GRU-based state estimator* fuses proprioceptive history with egocentric depth observations and is trained via self-supervised auxiliary objectives—heightmap reconstruction and linear velocity prediction—to distill local terrain geometry into a compact latent representation, requiring neither external localization nor manual annotation. *Terrain-adaptive footstep rewards* bias swing-leg placement toward geometrically favorable contact regions, enabling reliable footholds on discontinuous surfaces. Complementing these components, we conduct a depth resolution study that characterizes the Pareto frontier between reconstruction fidelity and computational cost, revealing that perceptual overhead can be substantially reduced with negligible impact on traversal performance.

We validate PRIOR on simulated terrains of progressive difficulty—flat ground, boxes, staircases, and gaps. Controlled ablations confirm that each component is individually necessary and that their combination produces synergistic gains beyond any subset, with the complete system reaching a 100% traversal success rate. To lower the barrier for future work, we will publicly release the full code. Our contributions are summarized as follows:

- We present PRIOR, a single-stage RL framework that unifies depth-based terrain perception, parametric gait generation, and terrain-adaptive footstep rewards to achieve robust and natural humanoid locomotion—eliminating the need for adversarial objectives, teacher–student distillation, or multi-stage training.
- We provide a depth resolution analysis that maps the Pareto frontier between terrain reconstruction quality and computational throughput, together with Isaac Lab-specific training optimizations that yield a 3× training speedup over the vanilla implementation.
- We conduct comprehensive ablations that isolate the necessity and synergy of each component, and commit to open-sourcing the complete framework as a reproducible baseline on Isaac Lab.

II. Related Work

A. Perception-Driven Robot Locomotion

Early research on legged robot locomotion focused primarily on "blind walking" strategies based on proprioception. By constructing closed-loop feedback control using internal sensors such as joint encoders and IMUs, these methods achieved a degree of robust walking on unknown terrains. Studies such as [5], [6], [7], [8], [9] have demonstrated that strong terrain adaptability can be attained relying solely on internal sensing. Based on this, several works have introduced motion prior constraints to generate more agile and holistic locomotion patterns [10], [11]. However, due to the lack of direct environmental observation, these methods struggle with precise, proactive gait planning when encountering significant non-local terrain variations.

As task complexity has increased, recent research has gradually shifted toward end-to-end locomotion learning frameworks. One category of methods introduces two-stage teacher-student distillation frameworks with privileged information [1], [12], [13], [14]. Some studies have incorporated temporal structures and attention mechanisms to enhance the robustness and consistency of terrain representations [2], [15], [16], while some other work has explored the integration of gait optimization [17]. Despite the significant progress made by these methods in complex environments, their motion generation often relies heavily on meticulous reward engineering or multi-stage training pipelines. This results in high system complexity and leaves room for improvement regarding the naturalness of the generated motion.

B. Motion-Prior-Based Robot Locomotion

Research on motion priors primarily focuses on enhancing the human-likeness of robot movements and can be broadly categorized into explicit motion generation and stylistic constraints. One prominent category follows the imitation learning paradigm. Some approaches achieve complex skill synthesis and multi-action switching through explicit trajectory constraints [18], [19]. Others utilize latent variables or probabilistic models to model high-dimensional motion distributions in order to generate diverse human-like movements [4], [20]. Recently, diffusion models have emerged as a significant tool for constructing motion priors, demonstrating remarkable advantages [21], [22]. Despite significantly improving gait naturalness, insufficient modeling of environmental perception and online feedback limits its adaptability and robustness in complex terrain.

Another category of methods utilizes reference motions as reward signals or discriminative criteria to guide the policy toward human-like styles in a weakly supervised manner. This approach enhances motion flexibility by relaxing strict trajectory matching constraints. Representative works include adversarial imitation for physical character control [23] and its applications in legged robot locomotion control [10], [24], [25]. Furthermore, some studies have utilized latent space modeling or Mixture-of-Experts (MoE) structures to enhance gait diversity while improving the continuity of motion generation [11], [17]. However, due to insufficient integration of environmental perception, learned human-like styles struggle to maintain consistency in real-world or unstructured environments.

III. Method

In this section, we provide a detailed description of the implementation of the PRIOR framework. PRIOR utilizes an estimator that fuses temporal depth information with proprioception to estimate both terrain features and robot proprioceptive states. Furthermore, reference gaits derived from processed natural motion data are integrated as inputs, providing soft constraints on robot locomotion via gait-aware rewards. Our discussion is organized into four primary components: the overall perceptive locomotion framework, the generation and learning of human-like motions, the high-throughput training infrastructure, and the specific implementation details of the training process.

A. Perception-Driven Locomotion Framework

1) Asymmetric Actor–Critic Architecture: As illustrated in Fig. 2, the proposed PRIOR framework adopts an asymmetric actor–critic architecture and is trained in a single-stage, end-to-end manner that enables the perception module and the control policy to co-evolve [2]. This design avoids error amplification issues commonly encountered in two-stage teacher–student distillation-based training pipelines.

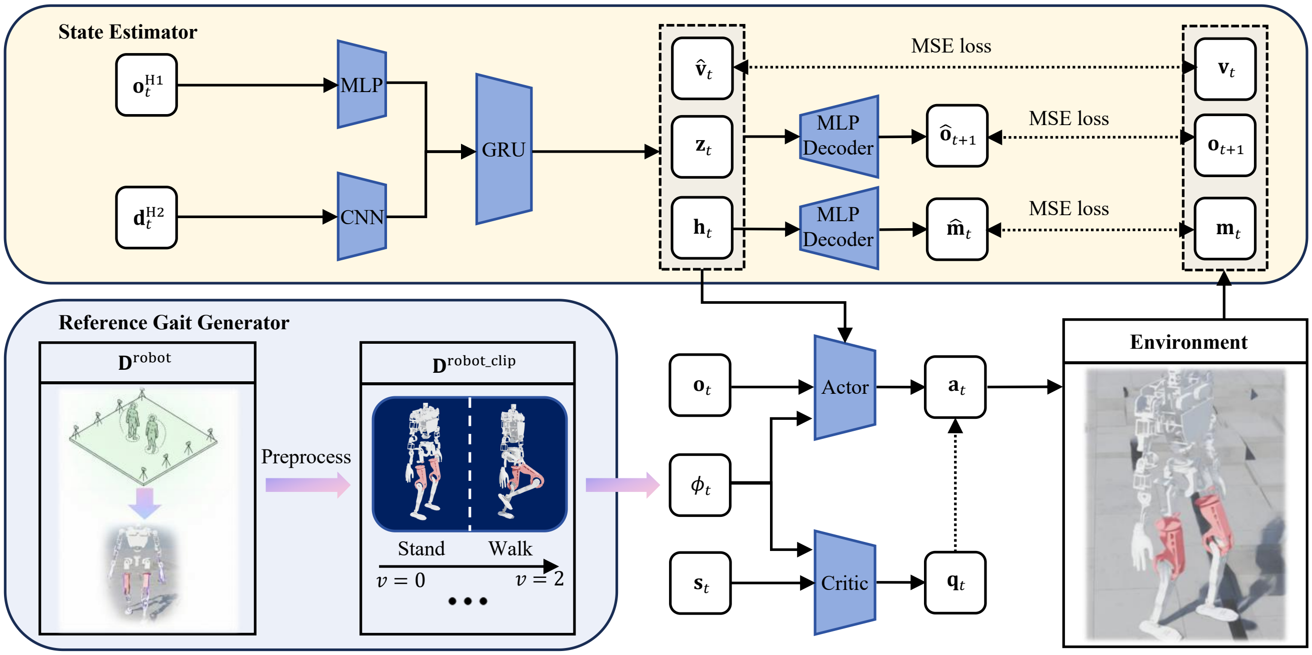


Fig. 2: Overview of the proposed PRIOR framework. The framework comprises three components: (a) Asymmetric actor-critic architecture for reinforcement learning. (b) **State Estimator** (yellow): Fuses multimodal latent features for policy driving and performs self-supervised regression for velocity estimation, terrain reconstruction, and state prediction. (c) **Reference Gait Generator** (blue): Synthesizes physics-consistent reference trajectories via phase normalization and velocity-driven weighted interpolation, while constraining locomotion through gait-aware rewards.

The policy is optimized using Proximal Policy Optimization (PPO) [26].

Actor network. The actor network takes (i) a 45-dimensional current proprioceptive observation $\mathbf{o}_t$, (ii) a 163-dimensional state estimator output $\mathbf{e}_t$ which is detailed in Section III-A.2, and (iii) a 1-dimensional gait phase signal ϕ_t as input. The current proprioceptive observation is defined as follows:

$$\mathbf{o}_t = \left[\boldsymbol{\omega}_t,\ \mathbf{g}_t,\ \mathbf{c}_t,\ \boldsymbol{\theta}_t,\ \dot{\boldsymbol{\theta}}_t,\ \mathbf{a}_{t-1}\right]^\top. \tag{1}$$

where $\boldsymbol{\omega}_t$ is the body angular velocity, $\mathbf{g}_t$ is the gravity direction vector expressed in the body frame, $\mathbf{c}_t$ is the velocity command, $\boldsymbol{\theta}_t$ and $\dot{\boldsymbol{\theta}}_t$ are the joint positions and velocities, respectively, and $\mathbf{a}_{t-1}$ is the action applied at the previous time step.

Critic Network. The critic network receives the noise-free base linear velocity $\mathbf{v}_t$ from the simulation environment, the proprioceptive observation $\mathbf{o}_t$, the height map scan information $\mathbf{m}_t$ provided by the RayCaster, and the reference gait phase ϕ_t. The input is defined as:

$$\mathbf{s}_t = \left[\mathbf{v}_t,\ \mathbf{o}_t,\ \mathbf{m}_t\ \right]^\top. \tag{2}$$

Reward and Action Space. We categorize the reward functions into four groups: task tracking, stability, smoothness, and safety. Specific rewards for foot-terrain interaction are crucial for humanoid balance and obstacle negotiation. These foot-related components, summarized in Table I, ensure gait rhythm, ground clearance, and precise landing on complex environmental features.

TABLE I: LANDING STATE REWARD COMPONENTS

Reward	Description	Weight (w_i)
r_{air}	Promotes gait rhythm.	1.25
r_{slide}	Minimizes ground slipping.	-0.10
$r_{\text{dbl-air}}$	Penalizes walking on one leg.	-1.00
r_{swing}	Ensures leg lift height.	-20.0
r_{stumble}	Prevents foot-obstacle tripping.	-30.0
$r_{\text{edge}}^{L/R}$	Encourages safe foot placement.	-2.00

The actor network outputs a 12-dimensional action vector $\mathbf{a}_t$ corresponding to the leg joints of the humanoid robot. This action serves as modulation to the default standing joint configuration $\boldsymbol{\theta}^{\text{default}}$, yielding the target joint positions $\boldsymbol{\theta}_t^{\text{target}}$,which is defined as:

$$\boldsymbol{\theta}_t^{\text{target}} = \boldsymbol{\theta}^{\text{default}} + \mathbf{a}_t. \tag{3}$$

The desired joint positions are then fed into a low-level PD controller to calculate the target joint torques $\boldsymbol{\tau}_t$, which is defined as:

$$\boldsymbol{\tau}_t = K_p(\boldsymbol{\theta}_t^{\text{target}} - \boldsymbol{\theta}_t) - K_d\dot{\boldsymbol{\theta}}_t. \tag{4}$$

where the stiffness K_p and damping K_d are set to 60.0 and 2.0, respectively, to match the hardware specifications of ZERITH Z1.

2) State Estimator: The state estimator fuses proprioceptive sensing and depth visual perception. Compared with estimation methods that rely solely on visual information, this multimodal fusion architecture improves the real-time performance of state feedback and reduces estimation bias caused by noise in a single visual modality. The two input modalities consist of proprioceptive observations $\mathbf{o}_t^{\mathrm{H1}}$ with a stacking horizon of H1 = 10, and temporally stacked depth images $\mathbf{d}_t^{\mathrm{H2}}$ with a stacking horizon of H2 = 2, where each depth frame has a cropped resolution of [36,64]. The proprioceptive observation $\mathbf{o}_t^{\mathrm{H1}}$ is processed by a multilayer perceptron (MLP) encoder to extract a 128-dimensional proprioceptive state feature, while the temporally stacked depth images $\mathbf{d}_t^{\mathrm{H2}}$ are processed by a convolutional neural network (CNN) encoder to extract a 128-dimensional depth feature. The encoded features of the two modalities are then concatenated and fed into a single-layer gated recurrent unit (GRU) to generate a memory representation of the proprioceptive state and the terrain state. The output of the memory module is a 163-dimensional vector $\mathbf{e}_t$, which serves as an input to the actor network, composed of a 3-dimensional base velocity $\hat{\mathbf{v}}_t$, a 32-dimensional latent vector $\mathbf{z}_t$ and a 128-dimensional height map latent vector $\mathbf{h}_t$.

$$\mathbf{e}_t = \left[\hat{\mathbf{v}}_t,\ \mathbf{z}_t,\ \mathbf{h}_t\right]^\top . \tag{5}$$

The height map latent vector $\mathbf{h}_t$ is decoded by an MLP to obtain the estimated terrain state $\hat{\mathbf{m}}_t$, while $\mathbf{e}_t$ is decoded by another MLP to predict the next-step proprioceptive state $\hat{\mathbf{o}}_{t+1}$. Using the privileged information available in the critic observations, self-supervised learning is applied to $\hat{\mathbf{v}}_t$, as well as the decoded results $\hat{\mathbf{m}}_t$ and $\hat{\mathbf{o}}_{t+1}$, by minimizing the mean squared error (MSE) loss. The overall loss function of the state estimator is defined as follows:

$$\mathcal{L} = \mathrm{MSE}(\hat{\mathbf{v}}_t, \mathbf{v}_t) + \mathrm{MSE}(\hat{\mathbf{o}}_{t+1}, \mathbf{o}_{t+1}) + \mathrm{MSE}(\hat{\mathbf{m}}_t, \mathbf{m}_t) \tag{6}$$

B. Reference Gait Priors

Inspired by the Parameterized Motion Generator (PMG) framework [27], we present a data-efficient motion prior extraction method designed to achieve humanoid control on complex terrains using a minimal set of high-fidelity motion templates. Distinguishing itself from conventional imitation learning approaches that necessitate hours of large-scale motion capture data, our method extracts structural features from core gait cycles to construct a continuous and physics-consistent reference gait space. This not only substantially lowers the overhead of data acquisition and preprocessing but also provides robust kinematic guidance for policy convergence in non-stationary terrain environments.

1) Motion Data Preprocessing: We utilize a high-precision motion capture system to collect human movement data, ranging from static postures to various forward velocities, denoted as $\mathbf{D}^{\mathrm{human}}$. This data is then retargeted to our humanoid robot, ZERITH Z1, using optimization-based algorithms [28], [29], resulting in the robot dataset $\mathbf{D}^{\mathrm{robot}}$:

$$\mathbf{D}^{\mathrm{robot}} = \{\boldsymbol{\theta}, \dot{\boldsymbol{\theta}}, \mathbf{v}, \boldsymbol{\omega}, \mathbf{c}\} \tag{7}$$

$\boldsymbol{\theta}$ represents the joint angles, $\dot{\boldsymbol{\theta}}$ represents the joint angular velocities, $\mathbf{v}$ represents the base linear velocity, $\boldsymbol{\omega}$ represents the base angular velocity, $\mathbf{c} = \{\mu, \sigma\}$ represents the foot contact information, including the contact center and range.

Since the raw retargeted dataset $\mathbf{D}^{\mathrm{robot}}$ contains initial transitions, terminal decelerations, and measurement noise inherent in the motion capture process, we designed a preprocessing pipeline to extract high-fidelity, periodic, and smooth reference motion segments. To construct reference trajectories with strict periodicity, we utilize the foot contact information $\mathbf{c}$ to segment the motion data at various velocities. We employ a clip_range mechanism to extract a single, stable gait cycle T from the original long sequences. In practice, we typically select the second or third cycle from the sequence to avoid the non-stationary dynamics associated with acceleration or deceleration phases. Furthermore, to eliminate high-frequency noise introduced by the motion capture system and ensure the smoothness of joint commands, we apply a 1D Gaussian filter to the raw joint sequences. The resulting processed robot dataset is represented as:

$$\mathbf{D}^{\mathrm{robot\_clip}} = \{\boldsymbol{\theta}, \dot{\boldsymbol{\theta}}, \mathbf{v}, \boldsymbol{\omega}, \mathbf{c}, T\} \tag{8}$$

where T denotes the duration of a single gait cycle.

2) Reference Gait Generation: Given the commanded base velocity $\mathbf{v} = [v_x, v_y, \omega]$ and the gait phase $\phi \in [0, 1)$, the reference joint trajectory is synthesized through a unified weighted interpolation framework.

For each velocity channel $x \in \{v_x, v_y, \omega\}$, we determine an interpolation coefficient based on the magnitude of the commanded velocity. Assume that the dataset contains a set of motion templates $\{\theta_i(\phi), T_i\}$, where $\theta_i(\phi)$ denotes the phase-dependent joint trajectory and T_i is the corresponding gait period.

Given a commanded velocity u_x, we select its two neighboring nominal velocities u_l and u_u, and define the interpolation factor as

$$\alpha = \mathrm{clip}\left(\frac{|u_x| - u_l}{u_u - u_l + \varepsilon}, 0, 1\right) \tag{9}$$

where ε is a small constant for numerical stability.

The commanded gait period is obtained via linear interpolation:

$$T_u = (1-\alpha)T_l + \alpha T_u. \tag{10}$$

The gait phase is then updated according to the normalized phase progression rule:

$$\phi_{t+1} = \left(\phi_t + \frac{\Delta t}{T_u}\right) \bmod 1. \tag{11}$$

After obtaining the updated phase ϕ, the reference joint trajectory is synthesized by blending the neighboring motion templates:

$$\theta_d(\phi) = (1-\alpha)\,\theta_l(\phi) + \alpha\,\theta_u(\phi). \tag{12}$$

To handle near-zero velocity commands, we introduce a standing threshold v_{th}:

$$\theta_d(\phi) = \begin{cases} \theta_{\mathrm{stand}}, & \|v\| \le v_{\mathrm{th}}, \\ (1-\alpha)\theta_l(\phi) + \alpha\theta_u(\phi), & \text{otherwise.} \end{cases} \tag{13}$$

TABLE II: GAIT-AWARE REWARD COMPONENTS

Reward	**Equation** (e_i)	**Weight** (w_i)
r_{pos}	$\|\boldsymbol{\theta} - \boldsymbol{\theta}_d(\phi)\|^2$	0.10
r_{vel}	$\|\mathbf{v}_b - \mathbf{v}\|^2$	0.05
r_{Δ}	$\|\Delta\boldsymbol{\theta} - \Delta\boldsymbol{\theta}_d(\phi)\|_1$	0.05
r_{ankle}	$\sum_{j\in\{L,R\}} \left(\theta_j - \theta_{d,j}(\phi)\right)^2$	0.05

TABLE III: PARALLEL SCALE UNDER DIFFERENT MEMORY STRATEGIES

GPU	**Max** N_{env}	**Storage**	**VRAM**
4090 (24G)	512	GPU	∼24G
4090 (24G)	1024	CPU	∼24G
4090 (48G)	1536	CPU	∼48G

In addition, a velocity-dependent stance ratio $\rho(\mathbf{v})$ is defined to construct phase-based contact indicators $r^L(\phi)$ and $r^R(\phi)$, which are later used for contact supervision and reward formulation.

This velocity-conditioned interpolation strategy ensures smooth transitions across different commanded speeds and maintains temporal consistency via unified phase evolution.

3) Gait-aware Reward Design: To encourage the policy to maintain consistency with the reference gait over complex terrains, we construct a set of exponential tracking reward terms based on the target joint trajectories and commanded base velocities generated by the reference gait module. Each reward term follows a unified exponential form:

$$r_i = \exp(-\lambda_i e_i), \tag{14}$$

where e_i denotes the tracking error of the corresponding physical quantity, and λ_i is a scaling coefficient. All gait-related reward terms are combined as a weighted summation:

$$r_{\text{gait}} = \sum_i w_i r_i, \tag{15}$$

where w_i represents the weight of each component.

The detailed definitions of each gait reward term are summarized in Table II.

These reward components constrain the policy from four complementary aspects, including pose consistency, velocity matching, dynamic motion trend tracking, and key support joint stabilization. This design allows the robot to preserve the periodic structure imposed by the reference gait generator while maintaining adaptability to complex environments.

C. High-Throughput Training Infrastructure

To address the computational overhead and GPU memory (VRAM) bottlenecks caused by high-dimensional depth perception in massively parallel environments, we developed a systematic engineering optimization framework on the NVIDIA Isaac Sim and Isaac Lab platform. Based on our self-developed ZERITH Z1 humanoid robot model, we constructed the training environment and improved overall system throughput from two perspectives: memory management and rendering strategy.

1) Heterogeneous Observation Buffer Management: VRAM capacity is a primary factor limiting the degree of parallelism (i.e., N_{env}) in reinforcement learning. To address the high-dimensional tensor storage pressure introduced by depth images, we propose a heterogeneous memory management scheme designed to decouple physics computation from data caching. Under this mechanism, VRAM serves only as a transient buffer for rendering outputs, while the generated observation tensors are asynchronously transferred to host memory (RAM) for storage and indexing.

Experimental results demonstrate that this strategy significantly frees GPU computational resources. As shown in Table III, on an RTX 4090 (24 GB) platform, the maximum number of parallel environments for vision-based tasks increases from the baseline of 512 to 1024. In a 48 GB VRAM configuration, the parallel scale further extends to 1536 environments, substantially improving sample efficiency during training.

2) Render-time Pre-processing Optimization: To optimize the vision processing pipeline, we first define the core perceptual requirement: the system must reliably distinguish terrain features with a minimum height of 5 cm. We derive the spatial resolution based on the geometric camera model:

$$\mathrm{r}_v = \frac{z_0 \sin(-\delta)}{\sin(\alpha)\sin(\alpha+\delta)} = \frac{z_0 \sin(\frac{\beta}{h})}{\sin(\alpha)\sin(\alpha - \frac{\beta}{h})} \tag{16}$$

where z_0 represents the mounting height, α is the pitch angle, β denotes the vertical Field of View (FOV), h is the image height in pixels, and auxiliary variable $\delta = -\beta/h$.

Under the configuration of $z_0 = 0.8\,\text{m}$, $\beta = 58°$, and $\alpha = 45°$, an effective vertical resolution of $h = 36\,\text{px}$ yields a spatial resolution of $r_v = 0.0463\,\text{m/pix}$ at a typical measurement distance of $1.13\,\text{m}$. Since $0.0463\,\text{m} < 0.05\,\text{m}$, this configuration satisfies the critical constraint for sensing 5 cm terrain variations while minimizing the input dimensionality.

Based on this analysis, we directly render a low-resolution depth buffer of 45×80 instead of high-resolution frames. A 36×64 region is extracted via center cropping, followed by stochastic depth perturbations to improve robustness. This render-time optimization reduces perception overhead and GPU memory bandwidth usage, enabling the policy to process visual observations at a significantly higher frequency, which is crucial for maintaining stable locomotion on highly irregular terrains.. The overall pipeline is illustrated in Fig. 3.

D. Training Details

1) Training Curriculum: To prevent policy instability caused by overly challenging terrains during the early stage of training, we adopt an adaptive terrain curriculum strategy based on the traveled distance of the robot, following the approach in [30]. In simulation, curriculum training is conducted simultaneously over four types of terrain: Pyramid Stairs, Inverted Stairs, Boxes, and Plane. All four terrain types are native terrain generators provided by Isaac Lab. Detailed terrain configurations and curriculum settings are illustrated in Fig. 4.

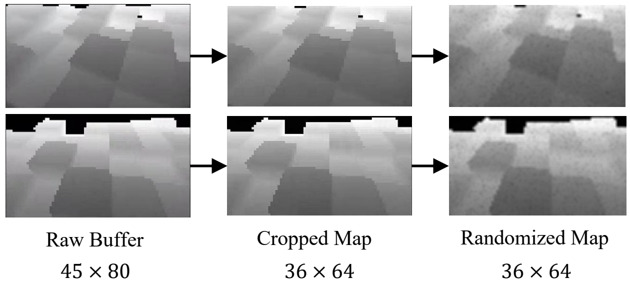
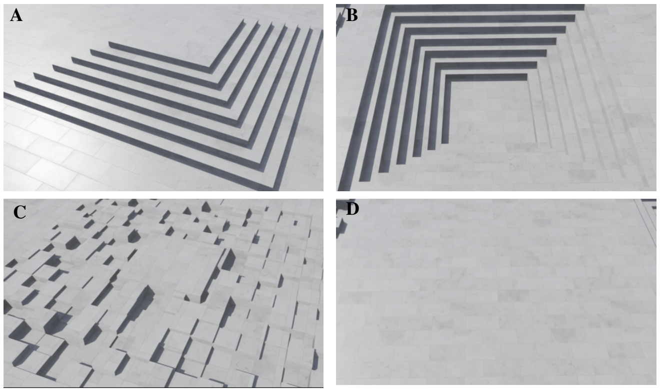


Fig. 3: Depth data flow: Images in the top and bottom rows represent samples from different parallel environments.



E	A	B	C	D
Terrain Type	Pyramid Stairs	Inverted Stairs	Boxes	Plane
Range (/m)	$[0.05, 0.23]$	$[0.05, 0.23]$	$[0.05, 0.20]$	–
Parameter	step height	step height	obstacle height	flat surface
Weight	0.2	0.2	0.2	0.1

Fig. 4: Overview of the terrain curriculum for training. (A)–(D) represent distinct terrain types posing various physical challenges. Table (E) details the parameter ranges and the distribution (weight) for each terrain type.

2) Domain Randomization: To enhance robustness and sim-to-real transfer, we apply domain randomization across physical dynamics, initial conditions, and visual inputs. The corresponding perturbation ranges are listed in Table IV.

IV. EXPERIMENT

A. Experimental Setting

Based on the optimization architecture described in Section III-C, we conduct high-throughput policy training on a single NVIDIA RTX 4090 (24GB) GPU. The system supports 1024 parallel environments simultaneously performing depth perception and physics simulation, significantly improving data collection efficiency while maintaining stable simulation dynamics. The policy converges after approximately 12,000 training iterations.

The depth images are updated at 30 Hz, while the control policy is executed at 50 Hz, ensuring a balance between perception refresh rate and control stability.

The trained policy is exported via ONNX and deployed directly on the onboard computing unit of the ZERITH Z1 humanoid robot. The ZERITH Z1 model has 23 degrees of freedom (DoF), including 6 DoF per leg, 3 DoF at the waist, and 4 DoF per arm.

TABLE IV: DOMAIN RANDOMIZATION RANGES

Parameter	Randomization range	Unit
Base payload	$[-5.0, 5.0]$	kg
Link mass factor	$[0.8, 1.2]$	–
Center of mass shift	$[-0.15, 0.15]$	m
Friction coefficient	$[0.2, 1.5]$	–
K_p factor	$[0.9, 1.1]$	–
K_d factor	$[0.9, 1.1]$	–
Joint armature	$[2 \times 10^{-3}, 2 \times 10^{-2}]$	kg·m^2
Initial base position (x, y)	$[-0.5, 0.5]$	m
Initial base orientation (yaw)	$[-\pi, \pi]$	rad
Initial base linear velocity	$[-0.5, 0.5]$	m/s
Initial base angular velocity	$[-0.5, 0.5]$	rad/s
Initial joint position scale	$[0.5, 1.5]$	–
Depth image bias	$[-0.04, 0.04]$	m
Depth image noise (σ)	0.02	m
Depth hole probability	0.03	–

B. Simulation Ablation Studies

1) Experimental Configurations: To systematically analyze the contribution of each component in the PRIOR framework, we design a staged ablation study to separately evaluate the perception-driven locomotion architecture and the reference gait prior.

First, four ablated variants are compared against the configuration without the reference gait prior ("PRIOR w/o reference gait") to investigate the impact of architectural components within the perception-motion framework. Subsequently, the "PRIOR w/o reference gait" model is compared with the full PRIOR framework to quantify the contribution of the reference gait prior to locomotion stability and behavioral quality.

The detailed configurations are as follows:

- **PRIOR (Ours)**: The complete PRIOR framework.
- **PRIOR w/o reference gait**: The PRIOR framework without the humanoid reference gait constraint.
- **PRIOR w/o $\hat{\mathbf{m}}_t$**: The PRIOR framework without explicit terrain estimation and elevation map supervision.
- **PRIOR w/o $\mathbf{d}_t^{\mathrm{H2}}$**: The PRIOR framework where temporally stacked depth observations are removed from the estimator input.
- **PRIOR with shorter H1**: The PRIOR framework with a reduced proprioceptive history length (H1 = 6).
- **PRIOR w/o landing state reward**: The PRIOR framework without the designed landing state reward terms.

2) Evaluation Procedure: All six policies are trained in the Isaac Lab simulator under the same terrain curriculum setting. We evaluate the policies using the following quantitative metrics:

- **Curriculum capability**: The mean terrain level achieved (mean level) and the maximum successfully traversed level across different terrain types (Pyramid Stairs, Inverted Stairs, Boxes, and Plane).

TABLE V: ABLATION STUDY RESULTS OF THE PRIOR FRAMEWORK ON TERRAIN ADAPTABILITY

Method	Mean Level	Pyramid Stairs	Inverted Stairs	Boxes	Plane	Mean Reward
PRIOR (ours)	5.7533	1.0	1.0000	1.0000	1.0	**26.3462**
PRIOR w/o reference gait	**5.7735**	1.0	1.0000	1.0000	1.0	23.7233
PRIOR w/o $\hat{\mathbf{m}}_t$	5.7672	1.0	0.7734	1.0000	1.0	13.1775
PRIOR w/o $\mathbf{d}_t^{\mathrm{H2}}$	5.4627	1.0	0.3750	0.9687	1.0	10.1463
PRIOR with H1 = 6	5.7417	1.0	1.0000	1.0000	1.0	19.3234
PRIOR w/o landing state reward	5.7403	1.0	1.0000	1.0000	1.0	22.6262

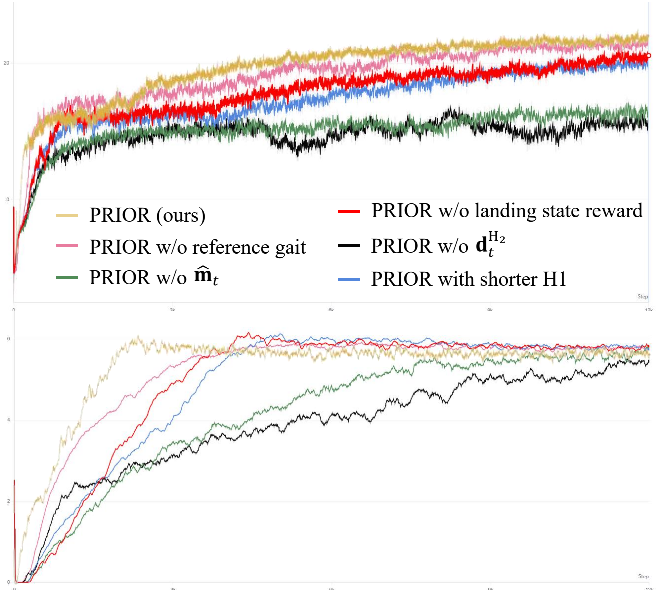


Fig. 5: The upper panel illustrates the training curves for mean reward, while the lower panel displays the mean terrain level achieved over training iterations.

- **Training performance**: Total reward, convergence speed, and training stability.

3) Results and Discussion: As shown in Table V and Fig. 5, the ablation results demonstrate the effectiveness of each component.

Reward–stability correlation: Although the “PRIOR w/o reference gait” variant achieves a slightly higher mean curriculum level (5.7735), its average reward (23.7233) is approximately 10% lower than that of the full framework. This suggests that traversal capability alone does not guarantee motion quality or efficiency.

Behavioral rationality: The reference gait prior contributes not only to performance but also to motion quality. Empirically, policies without gait prior tend to exploit unstable or high-frequency oscillatory motions to maximize traversal success. With the humanoid gait constraint, the robot maintains near-perfect terrain success rates (all terrain metrics reaching 1.0) while achieving significantly smoother and more energy-efficient locomotion, as reflected by higher reward values. This property is crucial for real-world deployment.

Explicit terrain estimation and supervision (PRIOR w/o $\hat{\mathbf{m}}_t$): Removing explicit terrain estimation reduces the average reward to 13.1775. Although it performs better than the variant without temporal depth information, its performance on complex terrains such as Inverted Stairs (0.7734) remains substantially lower than the full model. This indicates that explicit terrain representation and elevation supervision significantly enhance the policy’s understanding of geometric structures.

Temporal depth information (PRIOR w/o $\mathbf{d}_t^{\mathrm{H2}}$): This variant exhibits the worst performance, with an average reward of 10.1463 and a significant drop in success rate on Inverted Stairs (0.375). Without integrating temporal depth information, the robot loses the ability to anticipate terrain variations, leading to unstable foothold planning during dynamic locomotion.

Proprioceptive history length (H1): Reducing the proprioceptive history length results in an average reward of 19.3234. A longer history window (H1 = 10) enables the system to implicitly estimate physical properties such as ground friction and center-of-mass deviation, thereby improving robustness under unknown disturbances.

Landing state reward: As illustrated in Fig. 6, the removal of the landing state reward results in less stable foot placements, leading to a decrease in mean reward to 22.6262. This qualitative comparison reinforces that our fine-grained reward design effectively optimizes the behavior during the landing transient.

V. CONCLUSIONS

This paper presents **PRIOR**, an efficient and reproducible single-stage reinforcement learning framework developed in the Isaac Lab environment for perception-aware humanoid locomotion. The proposed method addresses the challenges of integrating perception and control for humanoid robots operating over complex terrains. **PRIOR** combines a GRU-based explicit terrain reconstruction state estimator with a parameterized gait generator within a unified learning pipeline. This design enables the ZERITH Z1 humanoid model to traverse diverse challenging terrains with high precision and robustness. Experimental results demonstrate that PRIOR achieves high average rewards and maintains a 100% traversal success rate across various complex terrains. Furthermore, the learned policy consistently handles terrains of relatively high difficulty, highlighting the effectiveness and generalization capability of the proposed framework.

Despite these advantages, our current work still has limitations in real-world deployment experiments. Future research will focus on developing more generalizable dynamics adap-

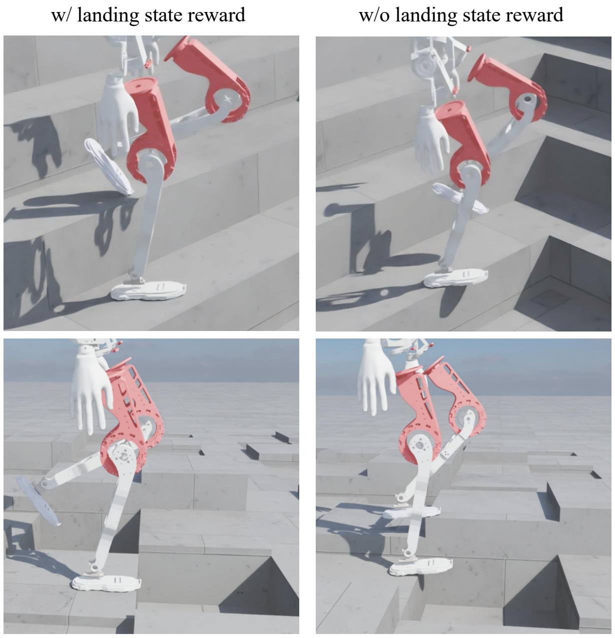


Fig. 6: Comparison of foot-placement behavior with and without the landing state reward.

tation algorithms to enable seamless sim-to-real transfer from Isaac Lab to the physical ZERITH Z1 platform.

References

[1] Z. Zhuang, S. Yao, and H. Zhao, "Humanoid parkour learning," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=fs7ia3FqUM

[2] S. Luo, S. Li, R. Yu, Z. Wang, J. Wu, and Q. Zhu, "Pie: Parkour with implicit-explicit learning framework for legged robots," *IEEE Robotics and Automation Letters*, vol. 9, no. 11, pp. 9986–9993, 2024.

[3] X. B. Peng, E. Coumans, T. Zhang, T.-W. Lee, J. Tan, and S. Levine, "Learning agile robotic locomotion skills by imitating animals," *arXiv preprint arXiv:2004.00784*, 2020.

[4] X. B. Peng, Y. Guo, L. Halper, S. Levine, and S. Fidler, "Ase: large-scale reusable adversarial skill embeddings for physically simulated characters," *ACM Transactions on Graphics*, vol. 41, no. 4, p. 1–17, July 2022. [Online]. Available: http://dx.doi.org/10.1145/3528223.3530110

[5] I. M. Aswin Nahrendra, B. Yu, and H. Myung, "Dreamwaq: Learning robust quadrupedal locomotion with implicit terrain imagination via deep reinforcement learning," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 5078–5084.

[6] A. Kumar, Z. Fu, D. Pathak, and J. Malik, "RMA: Rapid Motor Adaptation for Legged Robots," in *Proceedings of Robotics: Science and Systems*, Virtual, July 2021.

[7] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Science Robotics*, vol. 5, no. 47, Oct. 2020. [Online]. Available: http://dx.doi.org/10.1126/scirobotics.abc5986

[8] G. Ji, J. Mun, H. Kim, and J. Hwangbo, "Concurrent training of a control policy and a state estimator for dynamic and robust legged locomotion," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, p. 4630–4637, Apr. 2022. [Online]. Available: http://dx.doi.org/10.1109/LRA.2022.3151396

[9] J. Hwangbo, J. Lee, A. Dosovitskiy, D. Bellicoso, V. Tsounis, V. Koltun, and M. Hutter, "Learning agile and dynamic motor skills for legged robots," *Science Robotics*, vol. 4, no. 26, Jan. 2019. [Online]. Available: http://dx.doi.org/10.1126/scirobotics.aau5872

[10] J. Wu, G. Xin, C. Qi, and Y. Xue, "Learning robust and agile legged locomotion using adversarial motion priors," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 4975–4982, 2023.

[11] J. Wu, Y. Xue, and C. Qi, "Learning multiple gaits within latent space for quadruped robots," *arXiv preprint arXiv:2308.03014*, 2023.

[12] A. Agarwal, A. Kumar, J. Malik, and D. Pathak, "Legged locomotion in challenging terrains using egocentric vision," in *6th Annual Conference on Robot Learning*, 2022. [Online]. Available: https://openreview.net/forum?id=Re3NjSwf0WF

[13] X. Cheng, K. Shi, A. Agarwal, and D. Pathak, "Extreme parkour with legged robots," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 11 443–11 450.

[14] H. Wang, Z. Wang, J. Ren, Q. Ben, T. Huang, W. Zhang, and J. Pang, "Beamdojo: Learning agile humanoid locomotion on sparse footholds," in *Robotics: Science and Systems (RSS)*, 2025.

[15] J. He, C. Zhang, F. Jenelten, R. Grandia, M. Bächer, and M. Hutter, "Attention-based map encoding for learning generalized legged locomotion," *Science Robotics*, vol. 10, no. 105, p. eadv3604, 2025. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.adv3604

[16] J. Sun, G. Han, P. Sun, W. Zhao, J. Cao, J. Wang, Y. Guo, and Q. Zhang, "Dpl: Depth-only perceptive humanoid locomotion via realistic depth synthesis and cross-attention terrain reconstruction," *arXiv preprint arXiv:2510.07152*, 2025.

[17] D. Wang, X. Wang, X. Liu, J. Shi, Y. Zhao, C. Bai, and X. Li, "More: Mixture of residual experts for humanoid lifelike gaits learning on complex terrains," *arXiv preprint arXiv:2506.08840*, 2025.

[18] X. B. Peng, P. Abbeel, S. Levine, and M. van de Panne, "Deepmimic: example-guided deep reinforcement learning of physics-based character skills," *ACM Trans. Graph.*, vol. 37, no. 4, July 2018. [Online]. Available: https://doi.org/10.1145/3197517.3201311

[19] P. Xu, X. Shang, V. Zordan, and I. Karamouzas, "Composite motion learning with task control," *ACM Trans. Graph.*, vol. 42, no. 4, July 2023. [Online]. Available: https://doi.org/10.1145/3592447

[20] H. Zhang, L. Zhang, Z. Chen, L. Chen, Y. Wang, and R. Xiong, "Natural humanoid robot locomotion with generative motion prior," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025, pp. 6622–6629.

[21] A. Serifi, R. Grandia, E. Knoop, M. Gross, and M. Bächer, "Robot motion diffusion model: Motion generation for robotic characters," in *SIGGRAPH Asia 2024 Conference Papers*, ser. SA '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3680528.3687626

[22] Q. Liao, T. E. Truong, X. Huang, Y. Gao, G. Tevet, K. Sreenath, and C. K. Liu, "Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion," *arXiv preprint arXiv:2508.08241*, 2025.

[23] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa, "Amp: adversarial motion priors for stylized physics-based character control," *ACM Transactions on Graphics*, vol. 40, no. 4, p. 1–20, July 2021. [Online]. Available: http://dx.doi.org/10.1145/3450626.3459670

[24] T. Peng, L. Bao, J. Humphreys, A. M. Delfaki, D. Kanoulas, and C. Zhou, "Learning bipedal walking on a quadruped robot via adversarial motion priors," in *Towards Autonomous Robotic Systems*, M. N. Huda, M. Wang, and T. Kalganova, Eds. Cham: Springer Nature Switzerland, 2025, pp. 118–129.

[25] J. Shi, X. Liu, D. Wang, O. Lu, S. Schwertfeger, C. Zhang, F. Sun, C. Bai, and X. Li, "Adversarial locomotion and motion imitation for humanoid policy learning," *arXiv preprint arXiv:2504.14305*, 2025.

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

[27] C. Han, Y. Min, Z. Huang, A. Hong, H. Liu, Y. Cheng, and H. Liu, "Pmg: Parameterized motion generator for human-like locomotion control," 2026. [Online]. Available: https://arxiv.org/abs/2602.12656

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

[29] J. P. Araujo, Y. Ze, P. Xu, J. Wu, and C. K. Liu, "Retargeting matters: General motion retargeting for humanoid motion tracking," *arXiv preprint arXiv:2510.02252*, 2025.

[30] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," in *Conference on Robot Learning, 8-11 November 2021, London, UK*, ser. Proceedings of Machine Learning Research, A. Faust, D. Hsu, and G. Neumann, Eds., vol. 164. PMLR, 2021, pp. 91–100. [Online]. Available: https://proceedings.mlr.press/v164/rudin22a.html