

Walk the PLANC: Physics-Guided RL for Agile Humanoid LocomotioN on Constrained Footholds

Min Dai¹, William D. Compton², Junheng Li¹, Lizhi Yang¹, and Aaron D. Ames^{1,2}

Project Website: <https://caltech-amber.github.io/planc/>

Abstract—Bipedal humanoid robots must precisely coordinate balance, timing, and contact decisions when locomoting on constrained footholds such as stepping stones, beams, and planks—even minor errors can lead to catastrophic failure. Classical optimization and control pipelines handle these constraints well but depend on highly accurate mathematical representations of terrain geometry, making them prone to error when perception is noisy or incomplete. Meanwhile, reinforcement learning has shown strong resilience to disturbances and modeling errors, yet end-to-end policies rarely discover the precise foothold placement and step sequencing required for discontinuous terrain. These contrasting limitations motivate approaches that guide learning with physics-based structure rather than relying purely on reward shaping. In this work, we introduce a locomotion framework in which a reduced-order stepping planner supplies dynamically consistent motion targets that steer the RL training process via Control Lyapunov Function (CLF) rewards. This combination of structured footstep planning and data-driven adaptation produces accurate, agile, and hardware-validated stepping-stone locomotion on a humanoid robot, substantially improving reliability compared to conventional model-free reinforcement-learning baselines. The entire open-sourced code base is available at https://anonymous.4open.science/r/robot_rl-E4FF, and the supplementary video is provided at https://youtu.be/D_CZ_FRYjtc.

I. INTRODUCTION

Achieving robust and adaptive bipedal locomotion over discrete footholds such as stepping stones remains a fundamental challenge in legged robotics. These conditions are critically relevant for humanoid robots deployed in disaster zones or cluttered industrial environments, where viable contact regions are sparse, discontinuous, and bounded by strict safety margins. Unlike traversing continuous rough terrain, where small foot placement errors can often be tolerated, stepping-stone environments impose hard constraints on contact location and timing. A single misstep or localized slip can lead to unrecoverable failure, making full-body motion coordination significantly more challenging.

Traditionally, terrain-aware locomotion has been approached through model-based control strategies that prioritize physical consistency and safety constraints. In one school of thought, the challenge of adapting to discrete terrain is delegated to a dedicated footstep planner with quasi-static walking, often built on reduced-order models (RoMs) such as the Linear Inverted Pendulum (LIP) [1] and centroidal dynamics [2], [3]. Within this hierarchical structure, terrain reasoning can be explicitly encoded as a mixed-integer



Fig. 1. Model-guided RL traversing constrained footholds on the Unitree G1 humanoid robot. The policy is trained with environmentally consistent references generated dynamically during training by a momentum-regulated linear inverted pendulum for precise footstepping and center of mass regulation, and is deployed successfully on hardware.

optimization problem with explicit kinematic and foothold constraints [4], or solved through sampling-based methods for safe and efficient foot-step generation [5]. Instead of optimizing over the footstep locations, offline gait libraries generated with full order models [6] can be used with control barrier functions (CBFs) to modulate foot placement during execution [7], [8]. More recently, researchers have begun to unify footstep planning and locomotion control within a single optimization framework, enabling simultaneous reasoning over control actions and footstep locations [9], [10], [11]. While promising, these non-convex formulations impose substantial computational demands and remain challenging on high-DoF humanoid robots over discrete terrains. Consequently, many valid approaches are restricted to quasi-static motions [12].

Furthermore, a primary bottleneck for model-based methods lies in the integration of perception. The terrain dependency necessitates a rigorous terrain segmentation and state estimation pipeline [13]. In practice, purely model-based locomotion pipelines are already fragile, as they rely on highly accurate modeling and state feedback. Extracting precise mathematical representations for terrain geometry and foothold representations from noisy point clouds adds additional latency, uncertainty, complexity, and conservatism, as even minor

¹Authors are with the Department of Mechanical and Civil Engineering, California Institute of Technology, Pasadena, CA 91125.

²Authors are with the Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125.

This research is supported by Technology Innovation Institute (TII).

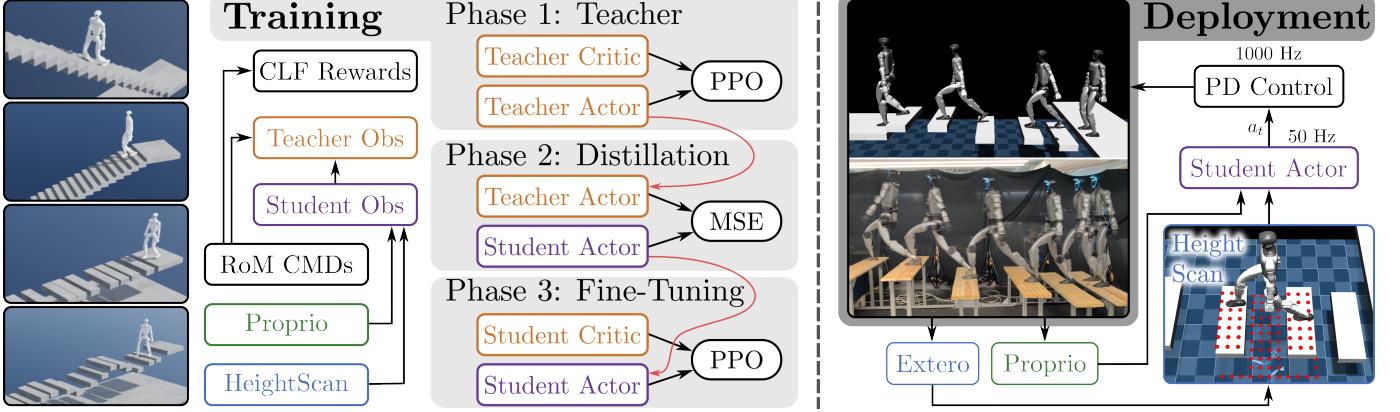


Fig. 2. A visual depiction of the model-guided RL architecture used to achieve stepping stones. The left column shows the four terrains trained in simulation, (top to bottom) stairs up, stairs down, flat stones, and height varying stones. Training is carried out in a teach-student fashion, with CLF rewards informed by the RoM. The student policy is successfully deployed in sim-to-sim transfer and on hardware.

segmentation or odometry errors can cause the planner to impose invalid constraints [14]. This dependence on clean, structured sensor data places unreasonably heavy demands on the perception module and makes it difficult to deploy pure model-based approaches in unstructured environments.

Conversely, deep reinforcement learning (RL) has emerged as a powerful alternative, demonstrating the ability to generate robust locomotion directly from high-dimensional sensor data, including height maps. RL policies naturally integrate proprioception and exteroception, learning robust behaviors that generalize to hardware deployment. However, standard RL pipelines often prioritize survival over precision. While successful on continuous rough terrain [15], [16], these methods struggle with the sparse foothold task, where precise end-effector control is paramount.

Early attempts to bridge this gap incorporated ground-truth terrain information directly into the RL observation space to guide foot placement [17], thus limited to simulation with perfect step knowledge. [18] focused on structured environments like stairs, but demonstrated limited foot placement accuracy and lacked real-world validation. More recently, vision-based RL has achieved hardware deployment [19], [20], [21], but these policies frequently exhibit low success rates in sparse-contact scenarios. Furthermore, unconstrained end-to-end approaches [22], [23] often exploit reward structures to maximize stability at the cost of motion quality, resulting in unnatural or period-two gaits. These methods can also be computationally prohibitive, with architectures like [23] requiring days of training.

The dichotomy is clear: model-based control excels at planning feasible motion under strict constraints (timing and placement) but is brittle to uncertainty; RL excels at robust execution under uncertainty but struggles to discover precise, constrained motions from scratch. This suggests that relying solely on reward engineering to induce precise stepping is inefficient. There is a growing consensus toward guided RL, where structured human trajectory references [24], [25] or model-based references [26], [27], [28], [29] constrain the learning manifold. By shaping the learning process around feasible reference motions, RL can focus on refining contact dynamics and disturbance rejection rather than rediscovering

the fundamentals of walking.

The core contributions of this work are the following,

- We propose a physics-guided reinforcement learning framework that tightly integrates a reduced-order, terrain-aware bipedal stepping-stone planner with CLF-based rewards design into the RL training loop, outlined in Fig. 2. The framework generates terrain-consistent references for constrained foot placement, CoM trajectory, and step timing for RL training with minimal computation overhead, enabling agile and safe locomotion over discrete aperiodic terrains.
- We remove the dependency and extended efforts on human motion dataset collections and retargeting pipelines in a multi-environment generalist locomotion policy. The model guidance provides a physically grounded, controllable, and transferable prior for RL in severely constrained environments.
- Our approach is both numerically and experimentally validated on the Unitree G1 humanoid robot over randomized stepping-stone terrains with gap and height variances, demonstrating improved stability, precise foot placement, generalization to unseen terrain, and consistent performance gains over model-free baselines.

II. PRELIMINARIES

A. Bipedal Robot Hybrid Dynamics

A bipedal robot can be modeled as a hybrid dynamical system $\Sigma = (\mathcal{D}, \mathcal{U}, \mathcal{G}, \Delta, f, g)$, where \mathcal{D} denotes the set of continuous domains, $\mathcal{U} \subset \mathbb{R}^{n_u}$ is the control input space, $\mathcal{G} \subset \mathcal{D}$ represents the guards corresponding to impact events, $\Delta : \mathcal{G} \rightarrow \mathcal{D}$ defines the discrete reset map, and f, g define the continuous dynamics. Let $\mathbf{x} = [\mathbf{q}^\top, \dot{\mathbf{q}}^\top]^\top \in \mathcal{T}\mathcal{Q}$ where \mathbf{q} is the generalized coordinate, the hybrid system's dynamics is given by:

$$\begin{cases} \dot{\mathbf{x}} &= f(\mathbf{x}) + g(\mathbf{x})\boldsymbol{\tau} & \mathbf{x} \in \mathcal{D} \setminus \mathcal{G}, \\ \mathbf{x}^+ &= \Delta(\mathbf{x}^-) & \mathbf{x}^- \in \mathcal{G}. \end{cases} \quad (1)$$

B. Reduced Order Models for Underactuated Dynamics

Bipedal walking is inherently underactuated, even with ankle actuation. The horizontal CoM cannot be arbitrarily

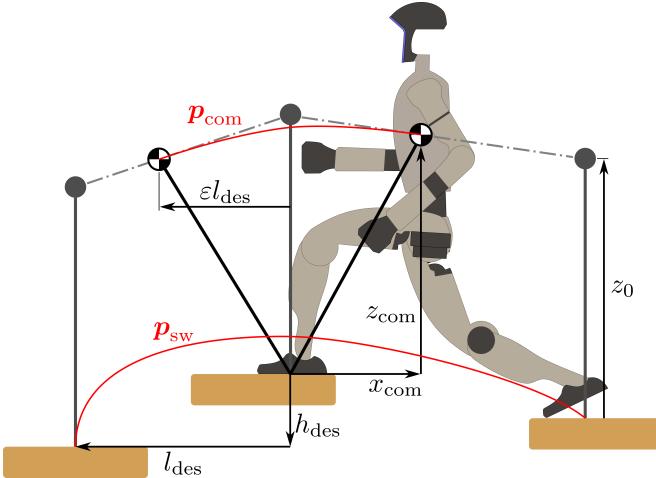


Fig. 3. A visual depiction of the LIP model used for gait synthesis, including the desired trajectories for the swing foot p_{sw} and CoM p_{com} .

regulated, especially at moderate or high speeds. For example, when a human leans forward, it does not take long before the center of pressure reaches the edge of the foot, prompting a step to prevent falling. To capture this dynamic tendency, the passive-ankle LIP model is commonly used. As shown in Fig. 3, we model the varying height stepping stones as virtual slopes by connecting the centers of the consecutive stones.

The LIP assumes the robot's CoM moves at constant height relative to the virtual slope along a massless leg with no ankle torque. Let $\mathbf{x} = [p \ L]^T$ represent the state vector, where p is the CoM position relative to the stance foot and L is the mass-normalized angular momentum about the stance foot. The sagittal-plane LIP dynamics can be expressed as:

$$\dot{\mathbf{x}} = \frac{d}{dt} \begin{bmatrix} p \\ L \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{z_0} \\ g & 0 \end{bmatrix} \begin{bmatrix} p \\ L \end{bmatrix}. \quad (2)$$

Here, z_0 is the constant vertical height of the CoM, and g is the gravitational acceleration. This linear system captures the underactuated evolution of the CoM state during single support, where momentum is generated solely by gravity-induced torques about the stance foot.

Orbital Energy and Step Viability: A key property of the LIP dynamics in (2) is the conservation of orbital energy, defined as:

$$E = \left(\frac{L}{z_0} \right)^2 - \lambda^2 p^2, \quad \text{with } \lambda = \sqrt{\frac{g}{z_0}}. \quad (3)$$

The orbital energy E partitions the state space into regions of viable motion: $E > 0$ corresponds to states with sufficient momentum to carry the CoM past midstance (forward walking), while $E < 0$ indicates bounded oscillations. Each orbital energy level defines a manifold of valid trajectories with varying boundary conditions. We leverage this property for stepping stones by selecting a target E^* that ensures forward progression.

Step Timing via Time-to-Impact: Given the linear structure of (2), the CoM evolution admits a closed-form solution:

$$p(t) = \cosh(\lambda t) p_0 + \frac{1}{\lambda z_0} \sinh(\lambda t) L_0, \quad (4)$$

where (p_0, L_0) is the initial state at the beginning of the step. This closed-form trajectory allows us to analytically compute the step duration T required for the CoM to reach a target switching state $p(T) = el_{\text{des}}$, where l_{des} is the distance to the next stepping stone from the current stance foot, as depicted in Fig. 3. This calculation ensures precise timing regulation over non-uniform footholds, which has proven crucial in the literature [30], [31].

Impact and Angular Momentum Reset Map: The robot's hybrid dynamics involve discrete changes at foot impact. At impact, the swing foot makes contact and becomes the new stance foot. Assuming instantaneous, inelastic impact and no slipping, the angular momentum about the new contact point is reset as:

$$L^+ = L^- + x_{\text{sw}} \dot{z}_{\text{com}}^- - z_{\text{sw}} \dot{p}^-, \quad (5)$$

where $(x_{\text{sw}}, z_{\text{sw}})$ is the position of the swing foot before impact, and $\dot{p}^-, \dot{z}_{\text{com}}^-$ are the CoM velocities just prior to impact. Critically, this map reveals that for a fixed foothold p_{sw} , the post-impact momentum L^+ can be actively regulated by modulating the pre-impact vertical velocity \dot{z}_{com}^- . We leverage this relationship to ensure the post-impact state always lies on the viable manifold defined by E^* .

C. Nonlinear Output Dynamics and Control

Let $\mathbf{y} = h(\mathbf{q}) \in \mathbb{R}^{n_y}$ denote the set of controllable outputs (e.g., CoM position, torso orientation, swing foot pose). For relative degree two outputs, the output dynamics satisfy

$$\ddot{\mathbf{y}} = L_f^2 h(\mathbf{q}, \dot{\mathbf{q}}) + L_g L_f h(\mathbf{q}, \dot{\mathbf{q}}) \boldsymbol{\tau}, \quad (6)$$

To ensure exponential convergence of outputs, we define a Control Lyapunov Function (CLF)

$$V(\boldsymbol{\eta}) = \frac{1}{2} \boldsymbol{\eta}^\top P \boldsymbol{\eta}, \quad \boldsymbol{\eta} = \begin{bmatrix} \mathbf{y}^a - \mathbf{y}^d \\ \dot{\mathbf{y}}^a - \dot{\mathbf{y}}^d \end{bmatrix}, \quad (7)$$

with $P > 0$ obtained from a continuous-time algebraic Riccati equation (CARE). The CLF decreasing condition

$$\dot{V}(\boldsymbol{\eta}) \leq -c V(\boldsymbol{\eta}) \quad (8)$$

is typically enforced through a quadratic program (QP) to compute the control input $\boldsymbol{\tau}$ that provides an exponential tracking condition on the output dynamics while satisfying the constrained dynamics and respecting the torque limit and contact constraints. In this work, CLF is directly embedded in the RL training as rewards as in [29].

D. Reinforcement Learning

We formulate the task of sparse foothold traversal as a partially observable Markov decision process (POMDP), defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{Z}, \mathcal{R}, \gamma)$, where \mathcal{S} is the set of states, \mathcal{A} is the set of actions, \mathcal{O} is the set of observations from proprioception and perception, $\mathcal{T}(s'|s, a)$ is the state transition probability, $\mathcal{Z}(o|s)$ is the observation likelihood, $\mathcal{R}(s, a)$ is the instantaneous reward function and $\gamma \in [0, 1]$ is the discount factor. The agent seeks to maximize the expected return

$$J(\pi_\theta) = \mathbb{E}_{\pi \sim \pi_\theta} \left[\sum_{t=0}^T \gamma^t r_t \right], \quad (9)$$

where $\pi_\theta(a_t|s_t)$ is the stochastic policy parameterized by θ . **Proximal Policy Optimization (PPO):** We use PPO [32] to optimize both the teacher and student policies. PPO improves training stability by constraining the policy update through a clipped surrogate objective:

$$\mathcal{L}_{\text{PPO}} = \mathbb{E}_t \left[\min \left(\rho_t(\theta) \hat{A}_t, \text{clip}(\rho_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right], \quad (10)$$

where $\rho_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is the importance sampling ratio and \hat{A}_t is the advantage estimate.

Teacher-Student Distillation: To improve learning under partial observability, we adopt a teacher-student framework. The teacher’s policy $\pi_T(a_t|\mathbf{o}_t, \mathbf{o}_t^{\text{priv}})$ has access to privileged information such as ground-truth terrain and contact phase. During each environment step, both the teacher policy and the student policy $\pi_s(a_t|\mathbf{o}_t)$ compute their respective actions. The student is then trained to match the teacher’s output using a per-step supervised behavior cloning loss:

$$\mathcal{L}_{\text{distill}} = \mathbb{E}_t \left[\left\| \pi_s(\mathbf{o}_t) - \pi_T(\mathbf{o}_t, \mathbf{o}_t^{\text{priv}}) \right\|^2 \right]. \quad (11)$$

During distillation, the student policy minimizes this loss over on-policy rollouts. After distillation, the student is further fine-tuned using PPO to adapt to the partially observed setting and improve robustness under noise. This two-stage training procedure allows us to deploy a policy that retains the teacher’s precision without relying on privileged inputs, and is especially useful when standard asymmetric actor-critic approaches fail to converge due to high information asymmetry.

III. MODEL-BASED REFERENCE GENERATION

As shown in Fig. 2, our control approach integrates a model-based locomotion planner with a teacher-student learning framework. First, a reduced-order walking model is used to plan dynamically and environmentally consistent reference trajectories for each time step (foot placement, center-of-mass trajectory, and step timing). Second, a multi-stage teacher-student RL pipeline leverages these references. This sections detail the model-based planner and how it guides the RL training process.

Step Time: Our planner builds upon the underactuated LIP model as in Sec. II. The step duration, T_s , is determined analytically based on the current post-impact state \mathbf{x}^+ and the required forward displacement l_{des} to the upcoming stepping stone as shown in Fig. 3. T_s is solved from the closed-form solution of the LIP dynamics, which specifies the time required for the horizontal CoM position to reach a target pre-impact location [9], $x_{\text{com}}^d = \epsilon l_{\text{des}}$, where $\epsilon \in (0, 1)$. In this work, we used $\epsilon = 0.6$. This process dynamically updates step timing, rather than relying on interpolation among gait libraries or fixed timing assumptions.

Orbital Energy Regulation using Vertical CoM Reference: We regulate the post-impact orbital energy in the sagittal plane E^+ to a desired $E^* = 0.6$ through regulating the momentum L_y at impact. In particular, we regulate the desired pre-impact vertical CoM velocity, $u_{\text{des}} = \dot{z}_{\text{com}}^-$ as in [9]. This target velocity acts as a discrete input to the hybrid system,

calculated using the momentum transfer law derived from the underactuated robot dynamics as in Eq. (5). The desired vertical velocity u_{des} is calculated based on the required change in angular momentum and the stone configuration:

$$u_{\text{des}} = \frac{1}{l_{\text{des}}} (\hat{L}_{\text{des}} - \hat{L}_y^-(\mathbf{x}) + h_{\text{des}} \hat{x}_{\text{com}}^-), \quad (12)$$

where $\hat{\cdot}$ is used to denote the estimation of a parameter using LIP dynamics and \hat{L}_{des} corresponds to the desired orbital energy E^* at post-impact. The estimation of pre-impact underactuated states \hat{x}_{com}^- and \hat{L}_y^- is performed using LIP solution given the time-to-impact. Other pre-impact states, including z_{com}^- and \dot{x}_{com}^- are approximated by weighted averages of the desired pre-impact ones and the current ones.

CoM Translation Reference: For CoM motion, we replace the MPC-style planner used in prior model-based work [9] with closed-form cubic splines that provides smooth references inside the RL training loop. The splines are parameterized by the phasing variable $s = \frac{t}{T_s} \in [0, 1]$, ensuring time-normalized trajectories across different step durations. The coefficients are obtained by enforcing boundary conditions on position and velocity at the beginning and end of each step. The vertical spline boundaries are set to satisfy momentum regulation, while the sagittal spline connects the current state to a target end-of-step state derived analytically from the desired orbital energy E^* . To ensure lateral balance, the reference follows the analytical solution of the HLIP model [33], directly mapping the planner’s desired velocity and position to the training reference.

Swing Foot Translation Reference: The swing-foot trajectory is generated by interpolating between the lift-off state and a target foothold using Bézier polynomials for all three dimensions. The target x_{sw} and z_{sw} are strictly determined by the geometry of the upcoming stepping stone. The lateral target w_{des} is produced by the HLIP reference to regulate step width. The final trajectory evolves as:

$$x, y_{\text{sw}}^d(s) = (1 - b_h(s))x, y_{\text{sw}}(\mathbf{x}^+) + b_h(s)l, w_{\text{des}}, \quad (13)$$

$$z_{\text{sw}}^d(s) = b_v(s), \quad (14)$$

where b_h and b_v are sets of Bézier polynomials [6]. The coefficients of b_h are $[0, 0, 0, \mathbf{1}_3]$, where $\mathbf{1}_N$ indicates a row vector of size N with all elements being 1. The coefficients of b_v are $[z_{\text{sw}}(\mathbf{x}^+), z_{\text{sw}}^{\max} \mathbf{1}_3, h_{\text{des}}, h_{\text{des}}]$, where z_{sw}^{\max} determines the swing clearance and is computed from the maximum height of the previous, current, and upcoming stones with an added safety margin.

Other References: To ensure stable contact and natural posture, the reference orientation for both the pelvis and the swing foot is aligned with the local terrain heading. Simultaneously, the arm joints track a nominal, phase-synchronized reference trajectory to aid in angular momentum cancellation and generate naturalistic gait dynamics.

Collectively, this planner functions as a real-time reference generator, producing a full state trajectory at every time step. This includes the desired output \mathbf{y}_d , the actual output \mathbf{y}_a , and holonomic constraint information. The tracked outputs include CoM position and velocity, swing-foot position and velocity, pelvis and swing-foot orientation and angular velocity, and

upper-body joint states. The holonomic constraints contain stance-foot position, velocity, orientation, and angular velocity, ensuring that ground-contact conditions remain consistent throughout single support. In the subsequent training phase, these trajectories are used throughout training to define CLF-based rewards, enforce CLF decreasing conditions, and incorporate holonomic and holonomic-velocity constraints within the RL objective [29].

IV. LEARNING ENVIRONMENT SETUP

To effectively leverage the model-based references described in III, the training framework operates hierarchically. The reduced-order model functions as a high-level stepping-stone planner, dictating the step timing T_s and dynamic targets u_{des} . The RL policy then serves as a low-level, high-frequency controller that learns the residual full-body dynamics necessary to robustly track these targets on the high-DoF humanoid model. This fusion combines the strengths of both paradigms: the model-based planner generates references consistent with environmental constraints, while the RL policy handles unmodeled dynamics (e.g., friction, compliance) and minimizes tracking errors using the full robot actuation capabilities. This hybrid approach mitigates the brittleness of pure model-based control while significantly improving sample efficiency compared to model-free baselines as in [22], [23], which often struggle to discover precise stepping strategies from scratch.

A. Terrain Generation

We construct a family of procedurally generated stepping-stone terrains with randomized geometric properties. A terrain instance is drawn from four terrain types: upstairs, downstairs, flat stepping stones, and height-varying stepping stones. These terrains differ in their sagittal spacing, vertical offsets, and the periodicity of the stone sequence.

We adopt a difficulty-indexed terrain sampling that incrementally expands the geometric variability of the stones. In general, at low difficulty, the stones are placed with small gaps with less variation and less vertical variation (if any). As difficulty increases, the allowable sagittal gap widens toward the maximum stone spacing; the stone height offsets expand symmetrically to introduce positive and negative elevation changes. Each episode consists of a procedurally generated terrain with parameters sampled from:

- Gap distance: $[0.3, 0.3 + 0.4d]$ m
- Height variation: $[-0.2d, 0.2d]$ m
- Stone x: $[0.13, 0.3]$ m
- Stone z: $[0.75, 1.25]$ m
- Stair depth: $[0.2, 0.3]$ m
- Absolute stair height: $[0., 0.2d]$ m

where $d \in [0, 1]$ represents terrain difficulty and corresponds to 10 discrete terrain levels. These parameters govern the placement, shape, and elevation of each stone. In addition, the supporting platform under the stones is randomized within a bounded vertical range $[0, -1]$ m. This produces a unified progression across all types of terrains, forcing the policy to learn generalizable stepping behaviors that remain stable under variations in gap width, vertical spread, and platform height.

To enable stable learning on challenging foothold arrangements, we implement a curriculum that adjusts terrain difficulty automatically based on task performance. At the end of each episode, the environment evaluates whether the robot successfully traversed the terrain. A terrain level advances only after three consecutive successful traversals, preventing noisy or accidental successes from triggering progression. The curriculum is monotonic: once an environment moves to a higher difficulty level, it never regresses. Because higher levels sample from a wider parameter space that still includes easier configurations, the agent continues to experience a mix of difficulties while steadily shifting toward more challenging foothold patterns. This approach ensures smooth progression and encourages the emergence of robust, generalizable locomotion behaviors across the entire terrain family.

B. Training Framework

The CLF-based reward formulation is inherently phase-dependent, aligning with terrain structure and command sequences. However, this phase information is unavailable at deployment, preventing its direct use in the student's observation space. Furthermore, we observed that a standard asymmetric actor-critic setup—where only the critic has access to privileged information—failed to converge on the stepping stone task. The optimization landscape for such precise, sparse-foothold locomotion is too difficult for PPO to traverse from scratch; without a structured prior, the policy consistently collapses to a local optimum (e.g., standing still). To overcome this, we adopt a teacher-student distillation framework designed to provide the student policy with a robust initialization prior. The pipeline proceeds in three steps, outlined in Fig. 2:

- 1) Teacher Training: A privileged teacher policy is first optimized to master basic locomotion (standing and walking) on nominal terrain. This phase leverages the full CLF-based reward structure and ground-truth phase information to generate dynamically consistent motion.
- 2) Distillation (Warm Start): We distill this basic competence into the student policy using behavior cloning. This step removes the dependency on privileged information, effectively "warm-starting" the student with a stable, non-privileged walking prior.
- 3) Curriculum Fine-tuning: Finally, the initialized student is trained via PPO on the full stepping-stone curriculum. Because the distillation phase positions the student in a favorable region of the optimization landscape, it can successfully adapt to the challenging height-varying stones and injected noise.

The student policy operates with deployable observations:

- **Command Vector** $c_t \in \mathbb{R}^d$: encodes task-specific information such as desired velocity.
- **Proprioception** $o_t^{\text{proprio}} \in \mathbb{R}^n$: consists of base angular velocity ω_t , projected gravity vector g_t , joint positions, and joint velocities.
- **Exteroception** o_t^{extero} : a local 1m×1m heightmap, with grid size 0.1m, capturing terrain geometry.
- **Last Action** $a_{t-1} \in \mathbb{R}^{21}$

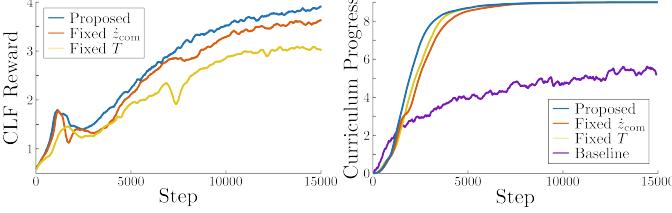


Fig. 4. (Left) The CLF reward curves during student policy training for the model-guided policies. The proposed method obtains better tracking than both the Fixed \dot{z}_{com} and Fixed T methods. (Right) A comparison between the model-guided policies and a naive policy; the naive policy is not able to handle the more difficult terrains.

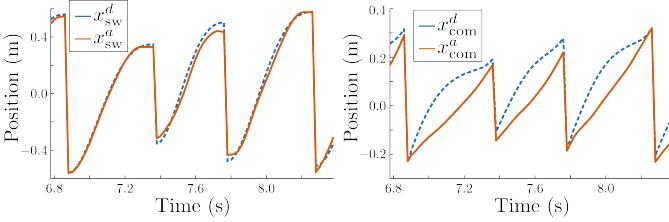


Fig. 5. The swing foot (left) and CoM (right) tracking in the x direction. Precise foot tracking is critical to hit constrained footholds, and CoM tracking is particularly good around contact, improving performance on future stones.

The actions $a_t \in \mathbb{R}^{21}$ encode desired angles for the 12 lower body joints, waist yaw joint, as well as 8 upper body joints. A PD controller is used for the low-level tracking with desired joint velocities trivially zero. Additionally, the teacher policy has access to privileged inputs, including stance foot and phase-related CLF terms and targeted swing foot position given terrain information and current stance foot location.

During teacher training and distillation, all observations are noise-free to ensure stable supervision. After distillation, the student policy is fine-tuned using reinforcement learning with deployable observations and injected noise to enhance generalization. The fine-tuning stage uses the same reward without access to teacher supervision or privileged inputs. This two-stage pipeline enables the student policy to first acquire structured behavior from a privileged teacher and then adapt to noisy and partial observations during fine-tuning for real-world deployment readiness.

V. RESULTS

All training is performed using a 21-DoF Unitree G1 humanoid in IsaacLab [34] with IsaacSim [35] physics engine. The robot is tasked with traversing randomized stepping-stone terrains and stair terrains. The policy outputs desired joint position targets at 50 Hz, which are tracked by a low-level PD controller.

Hyperparameters: We train using PPO with 4096 parallel environments using an MLP network with [512, 256, 128] hidden dimensions. Each policy is trained for 16k iterations from the initial flat ground walking policy. Details for training parameters are provided in Table II in the Appendix.

Domain Randomization: To facilitate sim-to-real transfer, we apply standard domain randomization to the robot's inertial properties, like CoM position and base mass. Contact dynamics are randomized with friction coefficients sampled uniformly between 0.4 and 1.0. Noise is injected into observations, including the elevation map. To enhance disturbance

TABLE I
SUCCESS RATE ACROSS TERRAIN CONDITIONS FOR BASELINE, ABLATIONS, AND THE PROPOSED METHOD.

Terrain	Baseline	Fixed \dot{z}_{com}	Fixed T	Proposed
Flat (mid difficulty)	0.92896	1	1	1
Flat (high difficulty)	0.55762	0.99854	1	1
Height-varying (mid)	0	0.99194	0.99902	1
Height-varying (high)	0	0.94116	0.87915	0.97217

rejection, we apply random velocity perturbations to the base during training, with linear velocities up to 0.5 m/s and angular velocities up to 0.4 rad/s.

Baseline Construction: As prior stepping-stone RL frameworks do not release implementation details, we implement our own direct end-to-end RL baseline. To ensure a fair comparison, the baseline uses the same student-level observation space, identical PPO hyperparameters, and the reward structure provided by the Unitree RL Lab [36]. When trained directly on the full stepping-stone environment, the policy immediately collapses to standing and makes no learning progress. To give the baseline the best possible chance, we adopt a two-stage curriculum: the policy is first pre-trained for 2k iterations on nominal flat-ground walking, and then fine-tuned exclusively on flat stepping stones. This modified curriculum allows the baseline to achieve non-zero performance on simple terrains for comparison.

Ablation Construction: In addition to the end-to-end RL baseline, we conduct two ablations to isolate the contribution of each component in the model-based reference: *fixed step timing* ($T = 0.4$ s) and *no momentum shaping* ($u_{\text{des}} = 0$).

A. Quantitative Results

Comparison with End-to-end Baseline: Across all experiments, the proposed model-guided RL policy and its two ablations demonstrate consistent and stable learning behavior, rapidly converging to high performance. In contrast, the end-to-end baseline improves only on flat stepping stones. While it learns to traverse flat stones, it saturates at a much lower curriculum level and ultimately converges to an asymmetric period-2 gait, consistent with findings in [22].

This contrast highlights a key limitation of pure RL: without model-based structure, the policy struggles to discover the precise, height-aware foot placements necessary for stepping-stone locomotion. Our model-guided approach leverages the RoM-based planner references to efficiently acquire feasible stepping strategies and refine them throughout training. Notably, as reported in Table I, it attains 100% success on challenging flat stepping stones, compared to only 56% for the end-to-end RL baseline. The success rate is evaluated over 4096 environments.

Ablation Analysis: To understand the contribution of each component in the model-guided RL pipeline, we conduct ablation studies that remove either (i) step timing adaptation or (ii) vertical COM momentum regulation. The CLF-reward traces (Fig. 4, left) show that all model-guided variants learn rapidly and stably, exhibiting a consistent upward trend throughout training. The full proposed method achieves the



Fig. 6. Snapshots of humanoid locomotion over a long stepping-stone course in MuJoCo simulation.

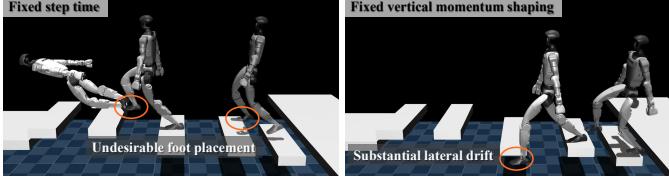


Fig. 7. Snapshots of sim-to-sim ablation trials.

highest CLF reward, reflecting better satisfaction of the model-based stability objective. Removing vertical COM regulation lowers performance, while fixing the step timing produces more pronounced degradation, with slower improvement and reduced steady-state reward. Interestingly, RL remains robust even under these ablations: despite the missing structure, the policies partially compensate by rediscovering surrogate strategies—either by implicitly regulating momentum through other reward channels or by learning to “survive” with less structured behavior.

Across evaluation terrains, all model-guided approaches achieve high success rates in simpler settings, including both difficulty levels of flat stepping stones and the mid-difficulty height-varying terrain. The advantage of the proposed method becomes particularly evident on the most challenging height-varying task, where it outperforms both ablations (Table I).

B. Simulation Evaluation

The high success rates on challenging terrain can be attributed to the policy’s ability to track the model-based references, as illustrated in Fig. 5. The proposed method exhibits precise tracking of the swing foot and CoM trajectories in the sagittal direction. To assess the generality and robustness of the learned locomotion strategies before transfer to hardware, we perform a sim-to-sim transfer study in which all final policies are deployed in a separate MuJoCo environment with independently parameterized dynamics, contact models, and actuator responses. As shown in Fig. 6, the proposed method successfully traverses an extended, mixed-terrain course that includes up-and-down stairs, height-invariant stepping stones, height-varying stones with large vertical offsets, sparse and narrow footholds, and an out-of-distribution mixed sequence designed to probe emergent robustness. These results highlight that the reduced-order structure embedded in the proposed approach not only accelerates learning but also significantly enhances transferability across simulation platforms.

Moreover, the ablated policies exhibit reduced sim-to-sim transfer reliability, as reflected in gait irregularities and inconsistent stepping patterns observed in the gait tiles (Fig. 7). This sensitivity underscores the stabilizing role of both step timing adaptation and vertical COM regulation in promoting

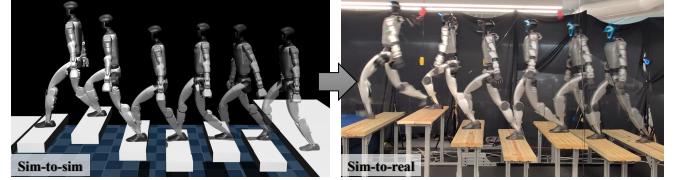


Fig. 8. Comparison of sim-to-sim and sim-to-real snapshots.

consistent, transferable locomotion strategies. For instance, the fixed T case exhibits inaccurate robot foot placements and undesirable flight phases; In the fixed \dot{z}_{com} case, the robot has significant y -direction drift.

Additionally, the resulting policy from the proposed method is robust to withstand external perturbations during stepping-stone traversal, such as unknown pushes to the robot body in the range of $\pm 100\text{Nm}$ for a 0.2s duration. We refer readers to the supplementary videos for more details.

C. Real-World Transfer

To evaluate our learned policy on hardware, we adopt a deployment pipeline similar to the GR-1 system described in [23]. The robot’s pose is tracked using a Qualisys motion capture system, and elevation maps are constructed by ray-casting onto a pre-designed terrain mesh that matches the physical testbed. Figure 1 shows gait tiles extracted from hardware rollouts of the G1 robot. The proposed controller produces a stable and repeatable stepping pattern with consistent swing trajectories and smooth torso motion, closely matching the behavior observed in simulation, as seen in Fig. 8. The tiles illustrate that the learned policy maintains foot placement accuracy and gait symmetry despite unmodeled dynamics, actuator latency, and real-world disturbances. These results confirm that the model-guided structure enables reliable hardware deployment without additional tuning.

Beyond standard evaluations, we challenged the policy with a mixed stepping-stone course featuring irregular stone depths, which is a configuration absent from the training curriculum. Despite this distribution shift, the policy generalized successfully (Fig. 1, top). This suggests that the model-guided references prevent the policy from overfitting to fixed geometric templates, enabling zero-shot generalization to unstructured terrain topologies.

VI. CONCLUSION

In this work, we presented a model-guided RL framework that bridges the gap between the precision of model-based planning and the robustness of data-driven control for bipedal locomotion on constrained terrains. By integrating a reduced-order stepping planner directly into the training loop, we generated dynamically consistent references to guide the policy

toward feasible motion manifolds without over-constraining the learning process. The successful zero-shot transfer to both a disparate physics simulator and physical hardware highlights the framework's effectiveness and robustness, offering a promising direction for deploying agile humanoids in complex, unstructured environments.

REFERENCES

- [1] R. J. Griffin, G. Wiedebach, S. McCrory, S. Bertrand, I. Lee, and J. Pratt, "Footstep Planning for Autonomous Walking Over Rough Terrain," in *2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*, Oct. 2019, pp. 9–16.
- [2] B. Ponton, M. Khadiv, A. Meduri, and L. Righetti, "Efficient Multicontact Pattern Generation With Sequential Convex Approximations of the Centroidal Dynamics," *IEEE Transactions on Robotics*, vol. 37, no. 5, pp. 1661–1679, Oct. 2021.
- [3] Y. Kojio, Y. Omori, K. Kojima, F. Sugai, Y. Kakiuchi, K. Okada, and M. Inaba, "Footstep Modification Including Step Time and Angular Momentum Under Disturbances on Sparse Footholds," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4907–4914, Jul. 2020.
- [4] R. Deits and R. Tedrake, "Footstep Planning on Uneven Terrain with Mixed-Integer Convex Optimization," in *2014 IEEE-RAS International Conference on Humanoid Robots*, 2014, pp. 279–286.
- [5] N. Perrin, O. Stasse, L. Baudouin, F. Lamiriaux, and E. Yoshida, "Fast Humanoid Robot Collision-Free Footstep Planning Using Swept Volume Approximations," *IEEE Transactions on Robotics*, vol. 28, no. 2, pp. 427–439, Apr. 2012.
- [6] J. W. Grizzle, C. Chevallereau, R. W. Sinnott, and A. D. Ames, "Models, feedback control, and open problems of 3D bipedal robotic walking," *Automatica*, vol. 50, no. 8, pp. 1955–1988, Aug. 2014.
- [7] Q. Nguyen, A. Hereid, J. W. Grizzle, A. D. Ames, and K. Sreenath, "3D dynamic walking on stepping stones with control barrier functions," in *2016 IEEE 55th Conference on Decision and Control (CDC)*. Las Vegas, NV, USA: IEEE, Dec. 2016, pp. 827–834.
- [8] Q. Nguyen, A. Agrawal, W. Martin, H. Geyer, and K. Sreenath, "Dynamic bipedal locomotion over stochastic discrete terrain," *The International Journal of Robotics Research*, vol. 37, no. 13–14, pp. 1537–1553, Dec. 2018.
- [9] M. Dai, X. Xiong, and A. Ames, "Bipedal Walking on Constrained Footholds: Momentum Regulation via Vertical COM Control," in *2022 International Conference on Robotics and Automation (ICRA)*, May 2022, pp. 10435–10441.
- [10] B. Acosta and M. Posa, "Bipedal Walking on Constrained Footholds with MPC Footstep Control," Sep. 2023, arXiv:2309.07993 [cs].
- [11] Z. Xiang, V. Paredes, and A. Hereid, "Adaptive Step Duration for Precise Foot Placement: Achieving Robust Bipedal Locomotion on Terrains with Restricted Footholds," Mar. 2024, arXiv:2403.17136 [cs, eess].
- [12] M. F. Fallon, P. Marion, R. Deits, T. Whelan, M. Antone, J. McDonald, and R. Tedrake, "Continuous humanoid locomotion over uneven terrain using stereo fusion," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, Nov. 2015, pp. 881–888.
- [13] B. Acosta and M. Posa, "Perceptive Mixed-Integer Footstep Control for Underactuated Bipedal Walking on Rough Terrain," May 2025, arXiv:2501.19391 [cs].
- [14] M. Fallon and M. Antone, "Plane Seg – Robustly and Efficiently Extracting Contact Regions from Depth Data," https://github.com/ori-drs/plane_seg, 2019.
- [15] W. Cui, S. Li, H. Huang, B. Qin, T. Zhang, hanjinchao, L. Zheng, Z. Tang, C. Hu, N. Yan, J. Chen, and Z. Jiang, "Adapting Humanoid Locomotion over Challenging Terrain via Two-Phase Training," Sep. 2024.
- [16] X. Gu, Y.-J. Wang, X. Zhu, C. Shi, Y. Guo, Y. Liu, and J. Chen, "Advancing Humanoid Locomotion: Mastering Challenging Terrains with Denoising World Model Learning," in *Robotics: Science and Systems XX*. Robotics: Science and Systems Foundation, Jul. 2024.
- [17] Z. Xie, H. Y. Ling, N. H. Kim, and M. v. d. Panne, "ALLSTEPS: Curriculum-driven Learning of Stepping Stone Skills," Aug. 2020, arXiv:2005.04323 [cs].
- [18] R. P. Singh, M. Benallegue, M. Morisawa, R. Cisneros, and F. Kanehiro, "Learning Bipedal Walking On Planned Footsteps For Humanoid Robots," in *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, Nov. 2022, pp. 686–693, iSSN: 2164-0580.
- [19] H. Duan, A. Malik, M. S. Gadde, J. Dao, A. Fern, and J. Hurst, "Learning Dynamic Bipedal Walking Across Stepping Stones," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2022, pp. 6746–6752, iSSN: 2153-0866.
- [20] H. Duan, A. Malik, J. Dao, A. Saxena, K. Green, J. Siekmann, A. Fern, and J. Hurst, "Sim-to-Real Learning of Footstep-Constrained Bipedal Dynamic Walking," 2022, p. 7.
- [21] M. Kim, B. Acosta, P. Chaudhari, and M. Posa, "Learning a Vision-Based Footstep Planner for Hierarchical Walking Control," in *2025 IEEE-RAS 24th International Conference on Humanoid Robots (Humanoids)*, Sep. 2025, pp. 1–8, iSSN: 2164-0580.
- [22] H. Wang, Z. Wang, J. Ren, Q. Ben, T. Huang, W. Zhang, and J. Pang, "BeamDojo: Learning Agile Humanoid Locomotion on Sparse Footholds," in *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, Jun. 2025.
- [23] 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, Aug. 2025.
- [24] Q. Liao, T. E. Truong, X. Huang, G. Tevet, K. Sreenath, and C. K. Liu, "BeyondMimic: From Motion Tracking to Versatile Humanoid Control via Guided Diffusion," Aug. 2025, arXiv:2508.08241 [cs].
- [25] F. Liu, Z. Gu, Y. Cai, Z. Zhou, H. Jung, J. Jang, S. Zhao, S. Ha, Y. Chen, D. Xu *et al.*, "Opt2skill: Imitating dynamically-feasible whole-body trajectories for versatile humanoid loco-manipulation," *IEEE Robotics and Automation Letters*, 2025.
- [26] W. Suliman, E. Davydenko, E. Chaikovskaya, and R. Gorbachev, "Reinforcement Learning-Based Footstep Control for Humanoid Robots on Complex Terrain," *IEEE Access*, vol. 13, pp. 178931–178944, 2025.
- [27] H. J. Lee, S. Hong, and S. Kim, "Integrating Model-Based Footstep Planning with Model-Free Reinforcement Learning for Dynamic Legged Locomotion," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2024, pp. 11248–11255, iSSN: 2153-0866.
- [28] F. Jenelten, J. He, F. Farshidian, and M. Hutter, "DTC: Deep Tracking Control," *Science Robotics*, vol. 9, no. 86, p. eadh5401, Jan. 2024.
- [29] K. Li, Z. Olkin, Y. Yue, and A. D. Ames, "CLF-RL: Control Lyapunov Function Guided Reinforcement Learning," Aug. 2025, arXiv:2508.09354 [cs].
- [30] M. Khadiv, A. Herzog, S. A. A. Moosavian, and L. Righetti, "Walking Control Based on Step Timing Adaptation," *IEEE Transactions on Robotics*, vol. 36, no. 3, pp. 629–643, Jun. 2020.
- [31] J. Li, Z. Duan, J. Ma, and Q. Nguyen, "Gait-Net-augmented Implicit Kino-dynamic MPC for Dynamic Variable-frequency Humanoid Locomotion over Discrete Terrains," in *Robotics: Science and Systems XXI*, vol. 21, Jun. 2025.
- [32] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," Aug. 2017, arXiv:1707.06347 [cs].
- [33] X. Xiong and A. Ames, "3-D Underactuated Bipedal Walking via H-LIP Based Gait Synthesis and Stepping Stabilization," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2405–2425, Aug. 2022.
- [34] M. Mittal *et al.*, "Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning," *arXiv preprint arXiv:2511.04831*, 2025.
- [35] NVIDIA, "Nvidia isaac sim," <https://developer.nvidia.com/isaac/sim>, 2024.
- [36] U. Robotics, "Unitree rl lab," https://github.com/unitreerobotics/unitree_rl_lab, 2024.

APPENDIX

TABLE II
SUMMARY OF TRAINING CONFIGURATIONS USED FOR PPO TRAINING, PPO FINE-TUNING, AND STUDENT–TEACHER DISTILLATION.

Parameter	PPO Train	PPO Finetune	Distillation
Steps per environment	24	24	120
Init noise std	1.0	1.0	0.1
Learning rate	1×10^{-3}	1×10^{-4}	5×10^{-4}
Entropy coefficient	0.008	0.003	—
Clip parameter	0.2	0.1	—
Learning epochs	5	5	10
Grad norm max	1.0	1.0	2.0
Loss type	clipped PPO	clipped PPO	Huber
Distillation gradient len	—	—	2