

CLF-RL: Control Lyapunov Function Guided Reinforcement Learning

Kejun Li^{*,1}, Zachary Olkin^{*,2}, Yisong Yue³, Aaron D. Ames²

Abstract—Reinforcement learning (RL) has shown promise in generating robust locomotion policies for bipedal robots, but often suffers from tedious reward design and sensitivity to poorly shaped objectives. In this work, we propose a structured reward shaping framework that leverages model-based trajectory generation and control Lyapunov functions (CLFs) to guide policy learning. We explore two model-based planners for generating reference trajectories: a reduced-order linear inverted pendulum (LIP) model for velocity-conditioned motion planning, and a precomputed gait library based on hybrid zero dynamics (HZD) using full-order dynamics. These planners define desired end-effector and joint trajectories, which are used to construct CLF-based rewards that penalize tracking error and encourage rapid convergence. This formulation provides meaningful intermediate rewards, and is straightforward to implement once a reference is available. Both the reference trajectories and CLF shaping are used only during training, resulting in a lightweight policy at deployment. We validate our method both in simulation and through extensive real-world experiments on a Unitree G1 robot. CLF-RL demonstrates significantly improved robustness relative to the baseline RL policy and better performance than a classic tracking reward RL formulation.

Index Terms—Humanoid and Bipedal Locomotion; Reinforcement Learning

I. INTRODUCTION

ACHIEVING stable and agile bipedal locomotion remains a central challenge in robotics due to the high-dimensional, underactuated, and hybrid nature of legged dynamics, which combine continuous motion and discrete events such as foot-ground impacts. This hybrid behavior poses significant challenges for designing controllers that combine precise coordination with robustness for real-world deployment.

Traditional model-based control frameworks address these challenges through principled formulations such as trajectory optimization. Methods based on hybrid zero dynamics (HZD)

Manuscript received: Aug. 12, 2025; Revised Nov. 13, 2025; Accepted Dec. 17, 2025.

This paper was recommended for publication by Editor Jens Kober upon evaluation of the Associate Editor and Reviewers' comments. This research is supported by Technology Innovation Institute (TII) and the National Science Foundation Graduate Research Fellowship. Supplementary video available: <https://youtu.be/KxSwKXjtko>

* denotes equal contribution. ¹Kejun Li is with the Department of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA kli5@caltech.edu

²Zachary Olkin and Aaron D. Ames are with the Department of Control and Dynamical Systems, California Institute of Technology, CA, USA [{zolkin, ames}@caltech.edu">{zolkin, ames}@caltech.edu](mailto)

³Yisong Yue is with the Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA yue@caltech.edu

Digital Object Identifier (DOI): see top of this page

©2026 IEEE

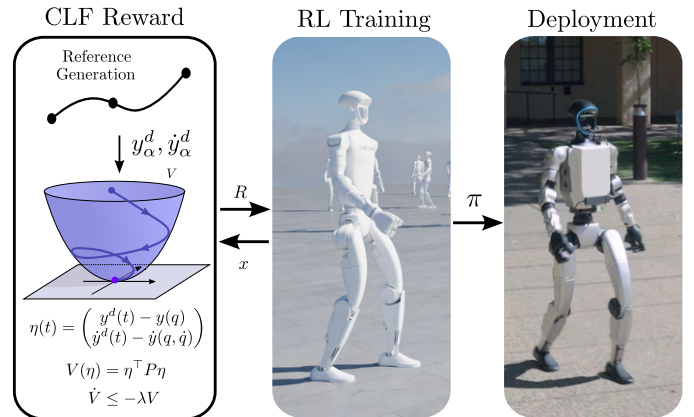


Fig. 1: Overview of our approach. A reference generator produces target trajectories, which are used to construct a CLF-based reward. An RL policy is trained in simulation with this reward and deployed on a real humanoid robot.

[1] and feedback-linearization have enabled offline generation of stable walking gaits under physical and task constraints [2]. These techniques provide strong guarantees on stability and constraint satisfaction but require solving large-scale nonlinear programs that are computationally demanding and sensitive to model mismatch, limiting their practicality for deployment. To enable faster and more reactive planning, many approaches leverage reduced-order models such as the linear inverted pendulum (LIP) and its variants [3], [4], which simplify the dynamics by focusing on the center of mass and foot placement. While computationally efficient and suitable for stabilization and mid-level planning, these models rely on strong assumptions such as massless legs and constant center of mass height, which limit their ability to generate highly dynamic and agile whole body motion.

Recent advances in computation and GPU-parallelizable physics simulators have made reinforcement learning (RL) a promising alternative for locomotion control. By offloading the heavy computation to offline training in massively parallel simulations, RL enables lightweight policy inference during online deployment to hardware [5], [6], [7], [8], [9]. However, applying RL to bipedal locomotion remains challenging due to its reliance on heuristic reward design, which is tedious to construct, sensitive to tuning, and, if poorly shaped, can lead to unstable gaits, prolonged training, and poor sim-to-real transfer. A strategy to alleviate the burden of this process is to embed model-based reference trajectories directly into the RL reward. This idea has been explored using reduced-order models [10], [11], full-order models [12], and with the HZD framework [13]. [14] proposes a different way to incorporate a

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

LIP controller into RL by generating desired footstep locations during training and using it to provide feedback control for rough terrain.

While prior model-guided approaches improve structure during training, they typically reward only the instantaneous tracking error, treating two states with the same error magnitude equally—even if one is actively converging toward the reference and the other is diverging. This can cause policies to undervalue transient corrections. To address these limitations, we propose a reward shaping framework built on control Lyapunov functions (CLFs), a fundamental tool in nonlinear control theory for generating certifiably stable controllers [15], [16], [17]. CLFs have previously been used in bipedal locomotion, often in conjunction with HZD, to formally guarantee the stability of periodic gaits [18] and have also been integrated into learning-based methods in other domains [19], [20]. For instance, [21] introduces a reward-resaping method that incorporates a candidate CLF for fine-tuning policies using minimal hardware data.

Our approach embeds both the CLF and its associated decrease condition directly into the RL reward, where the decrease condition encodes a desired minimal convergence rate, providing meaningful intermediate rewards. Although the idea may seem similar to potential shaping [22], there are fundamental differences. Potential shaping replaces the original reward, while we propose an additional reward, and this addition is designed directly according to the structure of a valid CLF. Further, our proposed framework reduces all the tracking to a single Lyapunov function, which as we discuss later, experimentally, does not work well with potential shaping. Overall, the proposed reward formulation is principled, easy to integrate, and less sensitive to manual reward balancing than conventional tracking formulations.

Building on these properties, we propose a structured reward shaping approach that combines model-based reference planning with CLF-inspired objectives to guide policy learning toward stable and robust behaviors (Fig. 1). We use two complementary planners for generating velocity-conditioned reference trajectories: (1) a reduced-order linear inverted pendulum (LIP) model for online generation, and (2) a precomputed hybrid zero dynamics (HZD) gait library from full-body offline trajectory optimization. Both produce periodic orbits but differ in fidelity and real-time suitability. From these trajectories, we construct a CLF $V = \eta^\top P \eta$, where η denotes the output tracking error. Once a reference trajectory is available, implementing the CLF decrease condition requires minimal effort: the tuning process is straightforward, involving far fewer reward terms than conventional training pipelines and resembling gain tuning in a tracking controller. The CLF and its associated stability condition is embedded into the RL reward to shape the policy to promote stable behaviors during training. We validate our approach through both simulation and hardware experiments on the Unitree G1 humanoid robot. Results demonstrate that CLF-based reward shaping improves tracking performance, reduces variance under randomized model perturbations, and enhances robustness compared to standard RL baselines—offering a promising path toward more robust and theoretically grounded locomotion learning.

II. PRELIMINARIES

Consider a robotic system with generalized coordinates $q = [q_b^\top, q_a^\top]^\top \in \mathcal{Q} \subset \mathbb{R}^n$, where $q_b \in SE(3)$ denotes the floating-base coordinates and $q_a \in \mathbb{R}^m$ the actuated degrees of freedom (DoFs). The control input is $u \in \mathbb{R}^m$, and the full-order state is given by $x = [q^\top, \dot{q}^\top]^\top$. Legged locomotion is often modeled as a hybrid system with continuous dynamics and discrete transitions such as impact events across domains [1]. Let \mathcal{D} denote the domain and \mathcal{S} the guard of the system. The hybrid system \mathcal{H} can be described as:

$$\mathcal{H} : \begin{cases} \dot{x} = f(x) + g(x)u & x \in \mathcal{D} \setminus \mathcal{S}, \\ x^+ = \Delta(x^-) & x^- \in \mathcal{S}, \end{cases} \quad (1)$$

where (1) represents the continuous full-order Lagrangian dynamics, while $\Delta : \mathcal{S} \rightarrow \mathcal{D}$ represents the discrete event when contact mode changes, and the superscripts “ $-$ ” and “ $+$ ” indicates the moments immediately before and after the discrete impact event, respectively.

Within each domain, the system dynamics can be derived from the Euler-Lagrange equation:

$$D(q)\ddot{q} + H(q, \dot{q}) = Bu + J_h(q)^T F \quad (3)$$

$$J_h(q)\ddot{q} + \dot{J}_h(q, \dot{q})\dot{q} = 0 \quad (4)$$

where $D(q) : \mathcal{Q} \rightarrow \mathbb{R}^{n \times n}$ is the mass-inertia matrix, $H : \mathcal{Q} \times T\mathcal{Q} \rightarrow \mathbb{R}^n$ contains the Coriolis and gravity terms, and $B \in \mathbb{R}^{n \times m}$ is the actuation matrix. The Jacobian of the holonomic contact constraint is $J_h(q) \in \mathbb{R}^{h \times n}$ and the associated constraint wrench is $F \in \mathbb{R}^h$.

A. Periodic Gait Generation

Legged locomotion is fundamentally periodic, alternating between stance and swing phases. This naturally motivates a step-to-step (S2S) analysis via the discrete-time Poincaré return map $P : \mathcal{S} \rightarrow \mathcal{S}$, with $x_{k+1} = P(x_k)$ giving the state from one impact to the next. A periodic gait is a fixed point x^* such that $P(x^*) = x^*$ and is exponentially stable when this fixed point is locally attracting [23]. Periodic motions can thus be synthesized by shaping the return map to yield a stable fixed point. Here, we focus on two complementary approaches: hybrid zero dynamics (HZD) and H-LIP models.

Hybrid Zero Dynamics: The Hybrid Zero Dynamics (HZD) approach synthesizes stable bipedal gaits by enforcing a set of phase-based virtual constraints, typically specified using Bézier polynomials with coefficients α . These constraints define desired trajectories for joint or end-effector motion as functions of a monotonically increasing phase variable. Driving the output of the virtual constraints to zero renders the zero dynamics manifold, denoted \mathcal{Z}^α , invariant and attractive. To achieve periodic walking under hybrid dynamics, the manifold must also be impact invariant, meaning that the system returns to \mathcal{Z}^α after a discrete event (e.g., foot impact):

$$\Delta(\mathcal{Z}^\alpha \cap \mathcal{S}) \subset \mathcal{Z}^\alpha \quad (5)$$

where Δ denotes the reset map and \mathcal{S} is the switching surface. When this condition holds, the system implicitly stabilizes a fixed point in the step-to-step dynamics.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

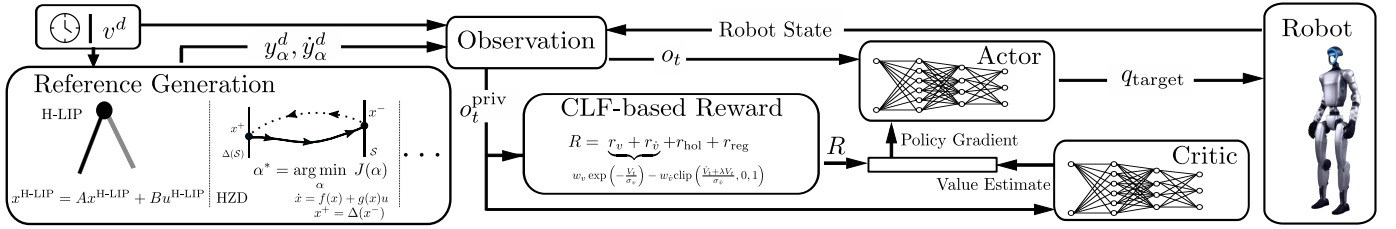


Fig. 2: Overview of CLF-guided reinforcement learning framework. A desired velocity v^d is passed to a reference generator (e.g., H-LIP or HZD) to produce targets $y_\alpha^d, \dot{y}_\alpha^d$. These, along with the robot state and privileged variables o_t^{priv} , are used to compute a CLF-based reward. The actor-critic policy is trained with this reward and outputs joint targets q_{target} for the robot.

Hybrid Linear Inverted Pendulum: Assuming constant (centroital) angular momentum and fixed p_{com}^z yields the well-known Linear Inverted Pendulum (LIP) for humanoid walking [3], [24]. To capture the hybrid nature of the gait, [25] introduced the Hybrid LIP (H-LIP), which, like LIP, assumes constant CoM height and step timing, but explicitly models single (SSP) and double-support (DSP) domains with a reset map. The mismatch to the full-order dynamics is treated as a bounded disturbance, enabling state-feedback foot-placement laws that render the horizontal CoM states bounded and invariant. Assuming no velocity jump, the two domains with step size $u^{\text{H-LIP}}$ compose into the equivalent reset:

$$\Delta_{\text{SSP}^- \rightarrow \text{SSP}^+} : \begin{cases} v_{\text{com}}^+ = v_{\text{com}}^- \\ p_{\text{com}}^+ = p_{\text{com}}^- + v_{\text{com}}^- T_{\text{DSP}} - u^{\text{H-LIP}}. \end{cases}$$

The S2S dynamics then take the linear form $x_{k+1}^{\text{H-LIP}} = Ax_k^{\text{H-LIP}} + Bu_k^{\text{H-LIP}}$, with $x^{\text{H-LIP}} = [p_{\text{com}}, v_{\text{com}}]^\top$, which serves as a reduced-order approximation of the Poincaré map of the full-order system; see [25] for derivation details.

Remarks: Although H-LIP enables online, feedback-based foot placement, we use steady-state trajectories derived by fixed-point identification and analytic rollout. Regardless of model order (full-order HZD vs. reduced-order H-LIP), both methods stabilize periodic gaits by regulating the Poincaré map, motivating our reward-shaping framework.

B. Control Lyapunov Functions (CLFs)

For a nonlinear control-affine system of the form $\dot{x} = f(x) + g(x)u$, a continuously differentiable, positive definite function $V(x) : \mathbb{R}^{2n} \rightarrow \mathbb{R}$ is called an exponentially stabilizing control Lyapunov function (CLF) if there exists a control input u such that, for some scalar $\lambda > 0$,

$$\inf_u \nabla_x V(x)(f(x) + g(x)u) < -\lambda V(x). \quad (6)$$

This inequality guarantees that $V(x)$ decreases along system trajectories, implying exponential convergence of the state x to the origin (or a desired manifold). Importantly, the CLF condition (6) defines a set of admissible control inputs rather than a unique control law. Any control input u that satisfies the inequality contributes to stability. This property makes CLFs especially appealing for reinforcement learning: rather than enforcing a fixed feedback controller, we can embed the CLF condition directly into the reward function, encouraging the policy to learn Lyapunov-stable behavior, without constraining it to a specific control action. For hybrid systems such as bipedal locomotion with discrete impact events, CLFs can be

applied only within each continuous domain of the dynamics. Prior work has shown that if the rate of convergence in the continuous phase is sufficiently fast, it can offset the destabilizing effects of impacts [18], providing certifiable stability.

III. REFERENCE-GUIDED REWARD SHAPING

We consider the reinforcement learning (RL) problem of learning a locomotion policy $\pi_\theta(a_t|o_t)$ that maps proprioceptive observations o_t to actions a_t using Proximal Policy Optimization (PPO) [26]. Instead of relying solely on handcrafted rewards, we embed model-based prior knowledge into the reward through a control Lyapunov function (CLF), promoting stability with respect to a reference trajectory. The full framework, including reference generation and CLF-based reward shaping, is illustrated in Fig. 2.

A. Reference Trajectory Generation

Our approach assumes a nominal reference trajectory that encodes a periodic or quasi-periodic walking motion given a desired velocity. We focus on two sources of reference trajectories: the H-LIP model and full-order HZD optimization, but the framework is agnostic to the origin of the trajectory.

H-LIP-Based Reference: Given a target walking velocity v^d , we first solve for the H-LIP fixed point $x^* = Ax^* + Bu^*$, where the step size $u^* = v^d T$ corresponds to the desired walking speed and nominal step time. This yields a periodic center-of-mass (CoM) state $x^* = [p_{\text{com}}, v_{\text{com}}]^\top$ consistent with the H-LIP S2S dynamics. Using this fixed point, we analytically generate the CoM trajectory over one step. The swing foot trajectory is constructed using a time-parameterized 5th-order Bézier curve to ensure foot clearance and smooth touchdown. For other end-effector or joint trajectories, we use simple sinusoidal patterns that are synchronized with the gait phase to produce time-varying desired trajectories.

HZD-Based Reference: To obtain HZD trajectories, we solve an offline trajectory optimization problem to determine the Bézier coefficients α that define the virtual constraints:

$$\begin{aligned} \{\alpha^*, X^*\} &= \underset{\alpha, X}{\operatorname{argmin}} \Phi(X) \\ \text{s.t. } \dot{x} &= f(x) + g(x)u && \text{(Dynamics)} \\ \Delta(\mathcal{S} \cap \mathcal{Z}_\alpha) &\subset \mathcal{Z}_\alpha && \text{(HZD Condition)} \\ X_{\min} &\leq X \leq X_{\max} && \text{(Decision Variables)} \\ c_{\min} &\leq c(X) \leq c_{\max}, && \text{(Physical Constraints)} \end{aligned}$$

where $X = (x_0, u_0, \dots, x_N, u_N) \in \mathbb{R}^{n_d}$ denotes the collection of decision variables, with $n_d \in \mathbb{N}$. Each x_i represents the

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

state at node i . The cost function is given $\Phi : \mathbb{R}^{n_d} \rightarrow \mathbb{R}_{\geq 0}$, while the constraints are represented by $c : \mathbb{R}^{n_d} \rightarrow \mathbb{R}^{n_p}$ with $n_p \in \mathbb{N}$. These constraints encode the physical laws governing locomotion, such as friction cone conditions, workspace limits, and actuator capacity [2]. Solving this optimization yields a stable periodic solution to the walking dynamics for a single step, parameterized by a fixed set of Bézier coefficients α^* . This solution is symmetrically remapped to produce a full gait cycle with alternating stance legs, yielding reference trajectories for the complete walking motion.

B. CLF-Based Reward Design

Let $\eta(t)$ denote the output tracking error between the current state and the reference trajectory:

$$\eta(t) = \begin{pmatrix} y^d(t) - y(q) \\ \dot{y}^d(t) - \dot{y}(q, \dot{q}) \end{pmatrix}, \quad (7)$$

and the corresponding control Lyapunov function is

$$V(\eta) = \eta^\top P \eta. \quad (8)$$

The matrix $P \succ 0$ is the unique solution to the Continuous-Time Algebraic Riccati Equation (CARE) for the feedback linearized model (see below). Prop. 1 proves that this is a valid CLF for the nonlinear system in single support flat-footed walking.

Proposition 1 (CLF Existence). *Consider the fully-actuated humanoid system given by Eq. (3) and satisfying the constraint given by Eq. (4) and friction cone constraints. If $Q \succ 0$ in the CARE, and the jacobian of the virtual constraints $\frac{\partial y}{\partial q}$ are invertible, then the function $V(\eta) = \eta^\top P \eta$ is a valid CLF for the full nonlinear dynamics during the single support phase.*

Proof. If the holonomic constraints, Eq. (4), are satisfied, and the friction cone is respected, then we can write the system dynamics without the constraint Jacobian [27]: $D(q)\dot{q} + H(q, \dot{q}) = Bu$. Define the virtual constraints error as $\bar{y} = y^d(t) - y(q)$ which we want to drive to 0. Differentiating twice yields

$$\ddot{\bar{y}} = \underbrace{\frac{\partial L_f \bar{y}}{\partial q} \dot{q} + \frac{\partial \bar{y}}{\partial q} (D(q)^{-1}(-H(q, \dot{q})))}_{L_f^2 \bar{y}(q, \dot{q})} + \underbrace{\frac{\partial \bar{y}}{\partial q} (D(q)^{-1}Bu)}_{L_g L_f \bar{y}(q, \dot{q})u}.$$

From the fully actuated assumption, $\dim q = \dim u = \text{rank } B$ [27]. Therefore $L_g L_f \bar{y}(q, \dot{q})$ is invertible under the assumptions of the proposition. We can now apply the feedback linearizing controller: $u = K(x)$, where $K(x)$ is defined as $K(x) := L_g L_f \bar{y}(q, \dot{q})^{-1} (-L_f^2 \bar{y}(q, \dot{q}) + v)$ where v is a new, auxiliary, input. This leads to the closed loop system $\dot{\eta} = A_\eta \eta + B_\eta v$ with A_η and B_η defined as $A_\eta = \begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix}$, $B_\eta = \begin{bmatrix} 0 \\ I \end{bmatrix}$. This is a double integrator system. We solve the CARE for this system and use the resulting P to synthesize an LQR gain: $K_\eta = R_\eta^{-1} B_\eta^\top P$. We can then define $v = K_\eta \eta$ as an optimal stabilizing auxiliary controller. Denote the closed loop η system as $A_\eta^{\text{cl}} = A_\eta + B_\eta K_\eta$. Consider the

function $V(\eta) = \eta^\top P \eta$. Using the feedback controller and differentiating yields

$$\dot{V}(\eta) = \frac{\partial V}{\partial \eta} \dot{\eta} = 2\eta^\top P A_\eta^{\text{cl}} \eta = \eta^\top (A_\eta^{\text{cl}\top} P + P A_\eta^{\text{cl}}) \eta. \quad (9)$$

Then, using the CARE, we get that $A_\eta^{\text{cl}\top} P + P A_\eta^{\text{cl}} = -Q - P B_\eta R_\eta^{-1} B_\eta^\top P := \bar{Q}$. By assumption Q is positive definite and note that the other term is positive semi-definite. Therefore

$$\dot{V}(\eta) \leq -\lambda_{\min}(\bar{Q}) \|\eta\|^2 \leq -\frac{\lambda_{\min}(\bar{Q})}{\lambda_{\min}(P)} V(\eta) \quad (10)$$

where λ_{\min} denotes the minimum eigenvalue. Using $V(\eta)$ as a Lyapunov function, the set $\{u \mid \dot{V} \leq -\lambda V\}$ is non-empty and thus $V(\eta)$ is a CLF for the nonlinear system. ■

Let $\sigma_v = \mu_{\max}(P) \eta_{\max}^2$, where $\mu_{\max}(P)$ is the maximum eigenvalue of the CLF matrix P and η_{\max} is an empirical bound on the tracking error. With $V_t = V(\eta(t))$, the CLF tracking reward is then

$$r_v = w_v \exp\left(-\frac{V_t}{\sigma_v}\right). \quad (\text{CLF Tracking})$$

We formulate the CLF decay reward using a clip function:

$$r_v^{\text{clip}} = -w_v \text{clip}\left(\frac{\dot{V}_t + \lambda V_t}{\sigma_v}, 0, 1\right), \quad (\text{CLF Decay})$$

where $\text{clip}(x, a, b) := \min(\max(x, a), b)$ and $\sigma_v = 2\|P\|\eta_{\max}^2 + \lambda\mu_{\max}^2$ and $\lambda > 0$ specifying the desired CLF decay rate. Specifically, this reward is providing a penalty if the CLF is not decaying sufficiently, otherwise there is no reward. Note that during RL training, to avoid explicitly computing $\dot{\eta}$ and \dot{V} , we approximate the Lyapunov derivative using finite differences: $\dot{V}_t \approx \frac{V_{t+1} - V_t}{\Delta t}$.

C. Auxiliary Rewards and Final Reward Composition

In addition to the CLF terms, we incorporate several auxiliary rewards commonly used in RL for legged systems. To enforce stance-foot holonomic constraints, we define a combined reward:

$$r_{\text{hol}} = w_{\text{hpos}} \exp\left(-\frac{\|p_{\text{st}} - p_{\text{st}}^0\|}{\sigma_p}\right) + w_{\text{hvel}} \exp\left(-\frac{\|v_{\text{st}}\|}{\sigma_v}\right),$$

where p_{st} and p_{st}^0 are the current and the initial stance foot positions when it enters the domain, v_{st} is the stance foot velocity, and w_{hpos} and w_{hvel} are weights.

To discourage excessive control effort, abrupt actions, and joint limit violations, we include a regularization reward:

$$r_{\text{reg}} = -w_u \|u_t\|^2 - w_{\Delta a} \|a_t - a_{t-1}\|^2 - w_{q_{\text{limit}}} \|\max(0, q_{\text{min}} - q_t) + \max(0, q_t - q_{\text{max}})\|_1,$$

where each term is weighted by its corresponding coefficient to balance control effort, motion smoothness, and adherence to joint limits.

The final reward used for policy training is the weighted sum of all components: $R = r_v + r_v^{\text{clip}} + r_{\text{hol}} + r_{\text{reg}}$. This reward structure encodes both task performance and physical feasibility, allowing the policy to balance tracking, stability, and regularization objectives. By combining model-based references (e.g.,

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE I: Reward weights used for HZD-CLF and LIP-CLF.

Reward Term	Weight
Torque penalty w_τ	1×10^{-5}
Action-rate penalty $w_{\Delta a}$	1×10^{-3}
Joint limit penalty $w_{q_{\text{limit}}}$	1.0
CLF tracking reward w_v	10.0
CLF decay penalty $w_{\dot{v}}$	2.0
Holonomic position reward w_{hpos}	4.0
Holonomic velocity reward w_{hvel}	2.0

from reduced-order planning or offline trajectory optimization) with CLF-inspired reward shaping and constraint-aware regularization, our framework embeds control-theoretic structure into the reinforcement learning process. This guides training toward stable, robust, and physically plausible locomotion, while preserving the flexibility and adaptability of model-free RL.

D. Implementation Details

We demonstrate our framework on the 29-DoF Unitree G1 humanoid robot, planning over 21 actuated DoF. The hand and wrist joints are fixed during training and controlled with zero desired position and velocity during deployment.

Reference Generation: We use the same set of end-effector and joint positions for both HZD and H-LIP-based trajectories. The full reference output $y_\alpha^d(t) \in \mathbb{R}^{n_y}$, with $n_y = 21$ is structured as follows: $y_\alpha^d = (p_{\text{com}}^d \ \phi_{\text{pelvis}}^d \ p_{\text{sw}}^d \ \theta_{\text{sw}}^d \ q_{\text{shoulder}}^d \ q_{\text{elbow}}^d)^\top$ where the terms represent the desired CoM position, pelvis orientation, swing foot pose, and arm joint angles. One can verify that this set of outputs lead to an invertible $\frac{\partial y}{\partial q}$ that satisfy Prop 1. To account for turning during training, we heuristically adjust the reference trajectory based on the commanded angular velocity. Specifically, we modify the yaw orientation of all end-effector frames to align with the integrated yaw derived from the commanded angular velocity over time. The HZD optimization is computed offline using IPOPT [28] and Casadi [29]. Hard constraints are used to enforce dynamics, periodicity, virtual constraints, and step length among others. The optimization problem is formulated as a multiple-shooting problem over a single swing phase.

Policy Structure: The policy receives a combination of proprioceptive and task-relevant inputs, including angular velocity, projected gravity, commanded linear and angular velocities, relative joint positions and velocities, the previous action, and a phase-based time encoding using $\sin(2\pi t/t_{\text{period}})$ and $\cos(2\pi t/t_{\text{period}})$. The stepping period t_{period} is set to 0.8 s, corresponding to a full gait cycle (i.e., two steps). The policy outputs joint position commands relative to a default symmetric standing pose at 50 Hz. This default pose is fixed across all policies we trained. Both the actor and critic networks use a fully connected feedforward architecture with hidden layer dimensions of [512, 256, 128] and Exponential Linear Unit activation function. The critic receives additional privileged information o_t^{priv} , such as stance and swing foot linear and angular velocities, reference trajectory positions and velocities, and binary contact state indicators.

Training Procedure: We use *IsaacLab*, a GPU-accelerated simulation framework built on top of NVIDIA Isaac Sim, with

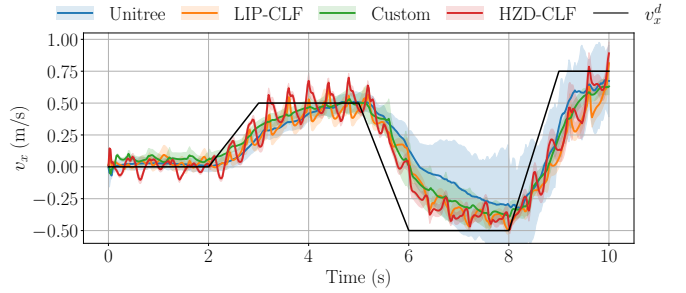


Fig. 3: Tracking performance with torso mass randomly displaced within a box of size $\pm[0.05(x), 0.05(y), 0.01(z)]\text{m}$ around the nominal location. Fifty displacements are uniformly sampled, and the resulting mean and standard deviation of performance for each policy are plotted.

its development initiated from the *Orbit* framework [30]. For policy training, we use the PPO implementation from Robotics System Lab RL library [31].

All training, including baseline and CLF-based results, use the same set of terrain settings, robot models, and domain randomization parameters. To improve policy robustness and facilitate transfer to hardware, we apply domain randomization to physical properties such as link masses, static and dynamic friction coefficients, and the center of mass position. Additionally, external perturbations are introduced by applying randomized velocity impulses in the x and y directions to the base link at fixed time intervals. Training is conducted over a range of commanded velocities, with linear velocity $v_x \in [-0.75, 0.75]$ and yaw rate $\omega_z \in [-0.5, 0.5]$. Given a commanded velocity, v^d , the reference trajectory is either retrieved from the precomputed gait library by selecting the closest matching forward velocity in the case of HZD, or generated online using the analytic form of the H-LIP gait library, which provides closed-form solutions across the velocity space. For both versions, we used the same reward listed under Table I.

The custom baseline policy is trained using a manually designed reward function composed of multiple heuristic terms. These include tracking linear velocities in the x and y directions as well as angular yaw velocity, while penalizing vertical velocity, angular velocities about the x and y axes, joint accelerations, joint velocities, joint torques, and abrupt changes in actions. The reward also discourages deviations from an upright base orientation and enforces compliance with joint limits. Additional terms promote behaviors such as maintaining an upright posture, preventing foot slip during contact, preserving nominal hip and torso positions, ensuring sufficient foot clearance, and regulating contact timing. Finally, the policy is penalized for deviations of the arm and torso joints from predefined reference configurations.

IV. RESULTS

We evaluate four policy variants: a baseline with rewards from the Unitree open-source implementation¹ (Unitree), a

¹<https://tinyurl.com/e5nrjdb9>

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE II: Tracking Velocity Performance Comparison.

Coord.	Method	Vel. Err. Mean ↓	Vel. Err. Std ↓
Center of Mass (cm/s)			
x	CLF (ours)	7.3 (-24.4%)	4.6 (-20.1%)
	Tracking only	9.6	5.8
y	CLF (ours)	5.9 (-4.6%)	3.7 (-7.0%)
	Tracking only	6.2	4.0
z	CLF (ours)	7.2 (-7.3%)	4.4 (-4.7%)
	Tracking only	7.7	4.7
Pelvis Orientation (rad/s)			
Roll	CLF (ours)	0.141 (-4.7%)	0.084 (-9.4%)
	Tracking only	0.148	0.093
Pitch	CLF (ours)	0.312 (-6.8%)	0.125 (-2.6%)
	Tracking only	0.335	0.128
Yaw	CLF (ours)	0.319 (-8.5%)	0.165 (-4.8%)
	Tracking only	0.348	0.173
Swing Ankle (cm/s)			
x	CLF (ours)	17.2 (-18.4%)	11.3 (-18.6%)
	Tracking only	21.1	13.9
y	CLF (ours)	9.7 (-13.3%)	7.4 (-10.4%)
	Tracking only	11.2	8.3
z	CLF (ours)	21.2 (+3.3%)	7.4 (-0.9%)
	Tracking only	20.5	7.4
Swing Ankle Orientation (rad/s)			
Roll	CLF (ours)	0.346 (-7.6%)	0.272 (+13.2%)
	Tracking only	0.374	0.240
Pitch	CLF (ours)	0.641 (+8.9%)	0.445 (+4.6%)
	Tracking only	0.589	0.425
Yaw	CLF (ours)	0.228 (-5.2%)	0.176 (-7.5%)
	Tracking only	0.241	0.190

baseline trained with hand-tuned custom rewards (Custom), a H-LIP-based CLF shaped policy (LIP-CLF), and an HZD-based CLF-shaped policy (HZD-CLF)². We also attempted to use potential shaping [22] to replace the tracking reward, but the policy never learned to walk. We hypothesize that potential shaping works better when some key rewards are left in their nominal form, but in our proposed framework there is only one tracking reward. The CLF-based policies were trained with many random seeds without any noticeable difference in performance, demonstrating that the proposed RL training pipeline is robust. All policies are first validated in sim-to-sim transfer using MuJoCo [32]. Because the Unitree-baseline underperforms compared to the hand-tuned baseline (Fig. 3), only three policies are tested on hardware (Fig. 5). The policies run onboard on an additional laptop mounted to the front of the torso during deployment, which also added additional mass (0.616 kg) to the robot.

CLF Decay Condition: To assess whether reference tracking alone is sufficient or if enforcing the CLF decay condition offers additional benefits, we conduct an ablation study. Table II compares two policies: one trained with only the tracking reward (r_v) and another trained with both the tracking and CLF decay rewards (r_v and $r_{\dot{v}}$). We simulate 200 instances of the robot under out-of-distribution domain randomization to evaluate robustness. The mean and standard deviation of the tracking errors are reported at steady state while walking at 0.75 m/s. The results show that including the CLF condition generally reduces both the mean tracking error and its variability, averaging about a 10% improvement in velocity tracking, and up to about 25%.

Robustness evaluation: To assess robustness to modeling

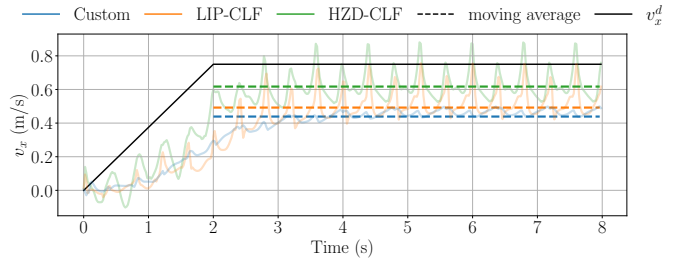


Fig. 4: Robustness evaluation in simulation: HZD-CLF, LIP-CLF and the baseline policy are tested with an extra 8kg on the torso. A 2-second ramp commanded each controller up to the maximum training velocity. The steady-state mean velocities are indicated by dashed lines.

mismatches, we introduce random perturbations to the torso mass location to simulate sim-to-real discrepancies in inertial properties. Specifically, the torso center of mass is uniformly displaced within a box of size $\pm[0.05 \text{ m } (x), 0.05 \text{ m } (y), 0.01 \text{ m } (z)]$. For each policy, we sample 50 randomized configurations and evaluate the velocity tracking performance across episodes, reporting the mean and standard deviation. As shown in Fig. 3, both CLF-RL policies exhibit significantly lower performance variance compared to the hand-tuned baseline, with the unitree-baseline having the largest variance. This indicates improved robustness and consistency under structural uncertainties. To further evaluate robustness, we introduce an 8kg payload to the torso and compare tracking performance across the three policy variants under a velocity ramp command. As shown in Fig. 4, the CLF-based RL policies maintain more accurate velocity tracking, demonstrating improved resilience to payload-induced dynamics changes.

Hardware Experiments: We deploy all the policies, except the Unitree baseline, on the Unitree G1 robot in a controlled indoor environment. Fig. 5 shows all three policies operating successfully on hardware. To quantify velocity tracking performance and robustness, we use a motion capture system to record global position and orientation data. This data is used solely for evaluation and is not fed into the controller. To remain consistent with simulation analysis, we track the pelvis frame and compute local-frame velocity by finite-differencing the position data (recorded at 240 Hz) and aligning it with the robot’s yaw orientation. Fig. 6 demonstrates the improved robustness of the HZD-CLF policy relative to the baseline. To test this robustness, we attach a backpack to the robot and load it with either 1.765 kg or 3.55 kg of additional mass. The HZD-CLF policy maintains consistent tracking performance across all conditions, whereas the baseline policy exhibits significant degradation. In the heavier case, the baseline policy drifts off the walking course and fails to complete the test. These hardware results corroborate the simulation findings in Fig. 3, which show that CLF-based policies exhibit lower performance variance under structural perturbations.

To evaluate the HZD-CLF policy under real-world conditions, we deploy the robot outdoors across a variety of environments. As shown in Fig. 7, the test route includes diverse flat-ground surfaces such as concrete and tiled walkways, as well

²Open source implementation is available at https://github.com/Zolkin1/robot_rl

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

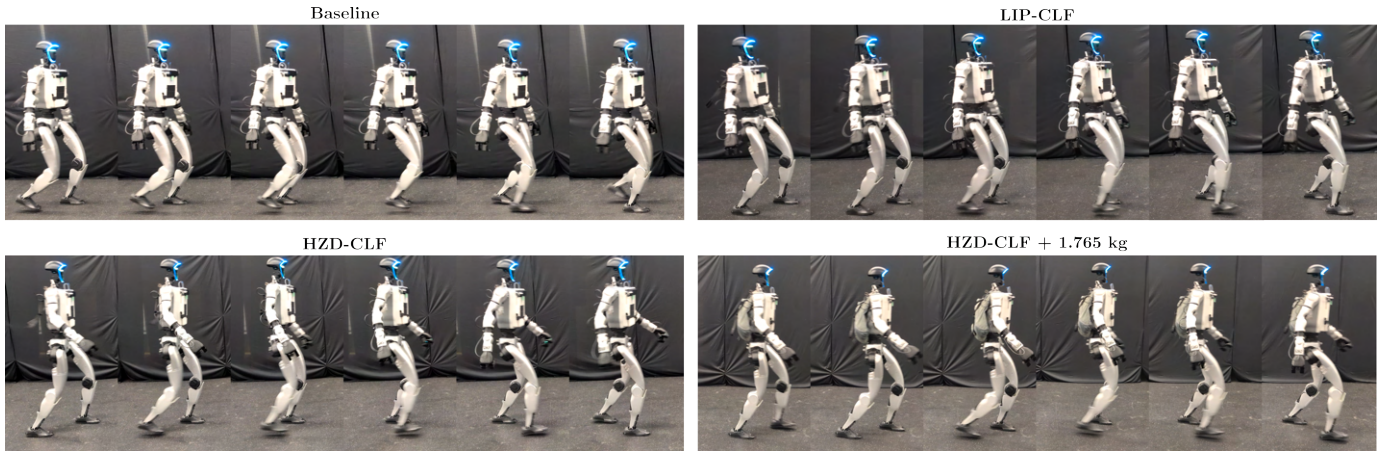


Fig. 5: Snapshots of the three policies throughout a stride on the Unitree G1 robot. These images depict the walking motion in steady state walking with a commanded velocity of $v_x^d = 0.75$ m/s.

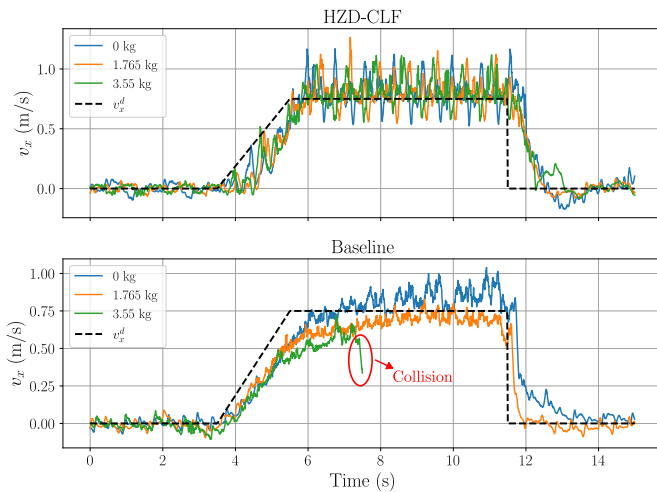


Fig. 6: Hardware testing with added backpack mass. HZD-CLF maintains good tracking performance, while the baseline drifts off the walking course and collides with a table.

as mild slopes like ADA-compliant ramps. The robot traverses all terrain types within a single continuous 0.25-mile walking trial without any failures. The distance was determined by experimental design considerations, rather than reflecting any limitation of the policy.

V. CONCLUSION

We presented CLF-RL, a structured reward shaping framework that integrates reference trajectory tracking with control Lyapunov functions (CLFs). By embedding CLF-based objectives directly into the reinforcement learning reward, our approach replaces heuristic reward design with a theoretically grounded stability metric. We showed that incorporating the CLF decay condition improves tracking performance over baseline tracking-only rewards. CLF-based policies exhibit superior robustness across a range of perturbations in both simulation and hardware, demonstrating increased reliability.

Our framework is modular and extensible. We demonstrated its compatibility with both reduced-order (H-LIP) and full-order (HZD) reference generators, though other trajectory planners could be incorporated. While our training only used

steady-state reference motions, the learned policies generalize well to transient motions—highlighting the framework’s flexibility. We validated the real-world robustness of the approach through extended outdoor trials, showing consistent and stable walking performance across varied flat-ground terrains. Finally, while our rewards are derived from CLFs that are theoretically sound for the full-order system, they are applied as guidance rather than hard constraints, leaving formal stability guarantees as a direction for future work.

REFERENCES

- [1] E. R. Westervelt, J. W. Grizzle, and D. E. Koditschek, “Hybrid zero dynamics of planar biped walkers,” *IEEE transactions on automatic control*, vol. 48, no. 1, pp. 42–56, 2003.
- [2] J. P. Reher, A. Hereid, S. Kolathaya, C. M. Hubicki, and A. D. Ames, “Algorithmic foundations of realizing multi-contact locomotion on the humanoid robot durus,” in *Algorithmic Foundations of Robotics XII: Proceedings of the Twelfth Workshop on the Algorithmic Foundations of Robotics*. Springer, 2020, pp. 400–415.
- [3] S. Kajita, F. Kanehiro, K. Kaneko, K. Fujiwara, K. Harada, K. Yokoi, and H. Hirukawa, “Biped walking pattern generation by using preview control of zero-moment point,” in *2003 IEEE international conference on robotics and automation (Cat. No. 03CH37422)*, vol. 2. IEEE, 2003, pp. 1620–1626.
- [4] S. Kajita, F. Kanehiro, K. Kaneko, K. Yokoi, and H. Hirukawa, “The 3D linear inverted pendulum mode: a simple modeling for a biped walking pattern generation,” in *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems.*, vol. 1. Maui, HI, USA: IEEE, 2001, pp. 239–246. [Online]. Available: <http://ieeexplore.ieee.org/document/973365/>
- [5] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, “Learning quadrupedal locomotion over challenging terrain,” *Science robotics*, vol. 5, no. 47, p. eabc5986, 2020.
- [6] Z. Xie, P. Clary, J. Dao, P. Morais, J. Hurst, and M. Panne, “Learning locomotion skills for cassie: Iterative design and sim-to-real,” in *Conference on Robot Learning*. PMLR, 2020, pp. 317–329.
- [7] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, “Learning to Walk in Minutes Using Massively Parallel Deep Reinforcement Learning,” Aug. 2022, arXiv:2109.11978 [cs]. [Online]. Available: <http://arxiv.org/abs/2109.11978>
- [8] Z. Zhuang, S. Yao, and H. Zhao, “Humanoid Parkour Learning,” Sep. 2024, arXiv:2406.10759 [cs]. [Online]. Available: <http://arxiv.org/abs/2406.10759>
- [9] Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, “Reinforcement Learning for Versatile, Dynamic, and Robust Bipedal Locomotion Control,” Jan. 2024, arXiv:2401.16889 [cs, eess]. [Online]. Available: <http://arxiv.org/abs/2401.16889>

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

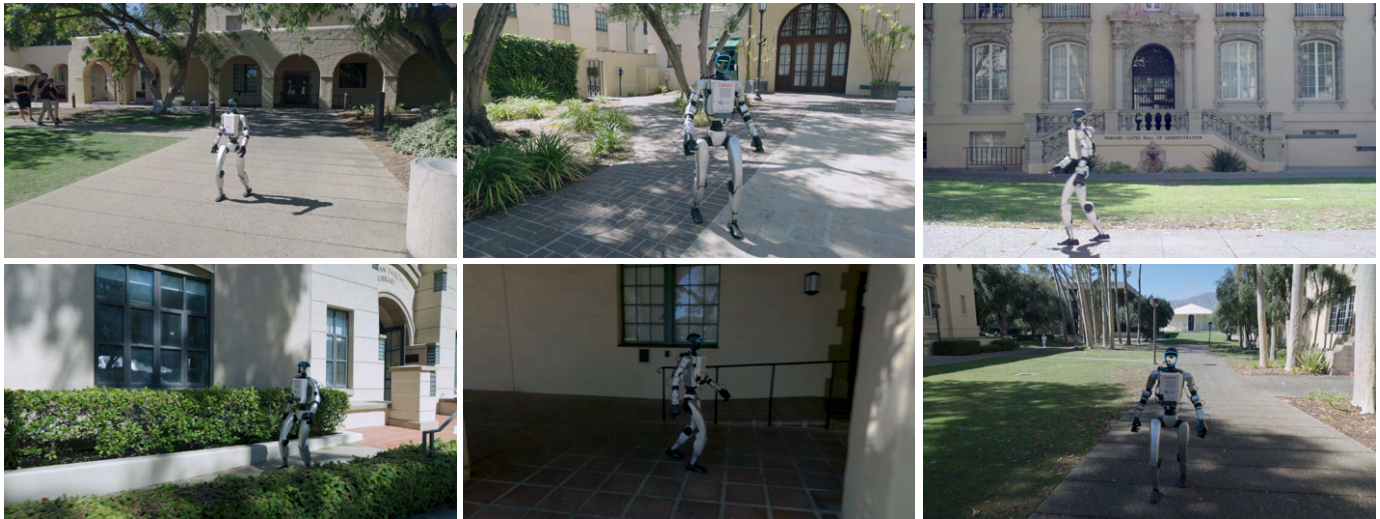


Fig. 7: Demonstration of extensive outdoor testing of the HZD-CLF policy shows its ability to handle diverse flat-ground surfaces, including various tile and concrete types, as well as mild uphill and downhill slopes.

- [10] R. Batke, F. Yu, J. Dao, J. Hurst, R. L. Hatton, A. Fern, and K. Green, "Optimizing Bipedal Maneuvers of Single Rigid-Body Models for Reinforcement Learning," Jul. 2022, arXiv:2207.04163 [cs]. [Online]. Available: <http://arxiv.org/abs/2207.04163>
- [11] K. Green, Y. Godse, J. Dao, R. L. Hatton, A. Fern, and J. Hurst, "Learning Spring Mass Locomotion: Guiding Policies with a Reduced-Order Model," Mar. 2021, arXiv:2010.11234 [cs]. [Online]. Available: <http://arxiv.org/abs/2010.11234>
- [12] F. Liu, Z. Gu, Y. Cai, Z. Zhou, H. Jung, J. Jang, S. Zhao, S. Ha, Y. Chen, D. Xu, and Y. Zhao, "Opt2skill: Imitating dynamically-feasible whole-body trajectories for versatile humanoid loco-manipulation," *IEEE Robotics and Automation Letters*, vol. 10, no. 11, pp. 12261–12268, 2025.
- [13] Z. Li, X. Cheng, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Reinforcement Learning for Robust Parameterized Locomotion Control of Bipedal Robots," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. Xi'an, China: IEEE, May 2021, pp. 2811–2817. [Online]. Available: <https://ieeexplore.ieee.org/document/9560769/>
- [14] 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)*. Abu Dhabi, United Arab Emirates: IEEE, Oct. 2024, pp. 11248–11255. [Online]. Available: <https://ieeexplore.ieee.org/document/10801468/>
- [15] E. D. Sontag, "A Lyapunov-Like Characterization of Asymptotic Controllability," *SIAM Journal on Control and Optimization*, vol. 21, no. 3, pp. 462–471, May 1983. [Online]. Available: <http://pubs.siam.org/doi/10.1137/0321028>
- [16] E. Sontag, "A 'universal' construction of Artstein's theorem on nonlinear stabilization."
- [17] Z. Artstein, "Stabilization with relaxed controls," *Nonlinear Analysis: Theory, Methods & Applications*, vol. 7, no. 11, pp. 1163–1173, Jan. 1983. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/0362546X83900494>
- [18] A. D. Ames, K. Galloway, K. Sreenath, and J. W. Grizzle, "Rapidly Exponentially Stabilizing Control Lyapunov Functions and Hybrid Zero Dynamics," *IEEE Transactions on Automatic Control*, vol. 59, no. 4, pp. 876–891, Apr. 2014. [Online]. Available: <http://ieeexplore.ieee.org/document/6709752/>
- [19] T. Westenbroek, A. Agrawal, F. Castañeda, S. S. Sastry, and K. Sreenath, "Combining Model-Based Design and Model-Free Policy Optimization to Learn Safe, Stabilizing Controllers," *IFAC-PapersOnLine*, vol. 54, no. 5, pp. 19–24, 2021. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S240589632101243X>
- [20] J. Choi, F. Castañeda, C. J. Tomlin, and K. Sreenath, "Reinforcement Learning for Safety-Critical Control under Model Uncertainty, using Control Lyapunov Functions and Control Barrier Functions," Jun. 2020, arXiv:2004.07584 [eess]. [Online]. Available: <http://arxiv.org/abs/2004.07584>
- [21] T. Westenbroek, F. Castañeda, A. Agrawal, S. Sastry, and K. Sreenath, "Lyapunov Design for Robust and Efficient Robotic Reinforcement Learning," Nov. 2022, arXiv:2208.06721 [cs]. [Online]. Available: <http://arxiv.org/abs/2208.06721>
- [22] S. H. Jeon, S. Heim, C. Khazoom, and S. Kim, "Benchmarking potential based rewards for learning humanoid locomotion," *arXiv preprint arXiv:2307.10142*, 2023.
- [23] B. Morris and J. W. Grizzle, "A restricted poincaré map for determining exponentially stable periodic orbits in systems with impulse effects: Application to bipedal robots," in *Proceedings of the 44th IEEE Conference on Decision and Control*. IEEE, 2005, pp. 4199–4206.
- [24] S. Kajita, F. Kanehiro, K. Kaneko, K. Fujiwara, K. Yokoi, and H. Hirukawa, "A realtime pattern generator for biped walking," in *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292)*, vol. 1. IEEE, 2002, pp. 31–37.
- [25] 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, 2022.
- [26] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
- [27] J. Grizzle, C. Chevallereau, A. D. Ames, and R. W. Sinnet, "3D Bipedal Robotic Walking: Models, Feedback Control, and Open Problems," *IFAC Proceedings Volumes*, vol. 43, no. 14, pp. 505–532, Sep. 2010. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1474667015370178>
- [28] A. Wächter and L. T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming," *Mathematical Programming*, vol. 106, no. 1, pp. 25–57, Mar. 2006. [Online]. Available: <http://link.springer.com/10.1007/s10107-004-0559-y>
- [29] J. Andersson, J. Gillis, G. Horn, J. Rawlings, and M. Diehl, "CasADi - A software framework for nonlinear optimization and optimal control," *Mathematical Programming Computation*, vol. 11, pp. 1–36, 2019.
- [30] M. Mittal, C. Yu, Q. Yu, J. Liu, N. Rudin, D. Hoeller, J. L. Yuan, R. Singh, Y. Guo, H. Mazhar, A. Mandlekar, B. Babich, G. State, M. Hutter, and A. Garg, "Orbit: A unified simulation framework for interactive robot learning environments," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3740–3747, 2023.
- [31] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," in *Proceedings of the 5th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 164. PMLR, 2022, pp. 91–100. [Online]. Available: <https://proceedings.mlr.press/v164/rudin22a.html>
- [32] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 5026–5033.