# Reachability-Guided Safe Policy Learning for Legged Locomotion

V Abhijith Narayan Rao, Shishir Kolathaya and Manan Tayal
Center for Cyber Physical Systems, Indian Institute of Science

***Abstract*—Hamilton–Jacobi (HJ) reachability offers principled safety signals, but integrating them into high-dimensional legged locomotion policies is difficult. Existing methods usually enforce safety at deployment time through shielding, controller switching, or scalar constrained objectives, which can introduce conservatism or gradient conflict with task learning. We instead integrate safety at training time in gradient space. Our method computes task and reachability-derived safety gradients, projects the safety gradient into the null space of the task gradient, and relaxes this projection near unsafe states with a reachability-aware damping schedule and bounded dual scaling. This yields a single end-to-end policy with no runtime safety filter or recovery controller. Evaluated on a quadrupedal robot in stair, climb, and corridor tasks, our method reduces joint-limit violations by $96.7\%$ and $94.6\%$ on climb and stair, and reduces collisions by $68.3\%$ and $41.3\%$ in corridor and scattered obstacle terrains while maintaining strong task performance. These results show that gradient-space safety integration is a practical alternative to deployment-time safety intervention for agile legged locomotion.**

## I. Introduction

Legged robots are rapidly shifting from flat floors to agile traversal of complex terrains, driven by parallel deep reinforcement learning (RL) [23, 11, 14, 17]. However, standard training pipelines lack tools to enforce physical safety constraints, such as joint limits or collision clearances, which are crucial to protect hardware. Typically, safety is handled via heuristic reward shaping or terminations, which couple task and safety performance and require tedious penalty weight tuning [12, 5, 21].

Hamilton–Jacobi (HJ) reachability analysis [18, 4] provides a principled alternative by computing a safety value function whose sign indicates recoverability. Recent neural reachability approximations [3, 10, 8, 28] have made this analysis tractable for high-dimensional systems, establishing reachability as an online, differentiable training signal.

However, existing HJ reachability and barrier methods operate at *deployment time*: action shielding overrides control outputs via online filters [6, 2, 27], and recovery switching invokes fallback controllers when safety drops [7, 9, 29]. For agile locomotion, these approaches introduce major limitations. First, dual-policy schemes like Agile But Safe (ABS) [9] operate on coarse 2D centroidal/body-level velocities, leaving the robot vulnerable to joint-space limit violations or 3D leg-environment collisions. Second, continuous switching triggers abrupt, destabilising control handoffs. Third, constrained-MDP methods treat safety as a scalar objective to be summed with task performance, forcing the optimizer to resolve competing gradients implicitly.

In this work, we leverage HJ reachability *at training time to synthesize a single, unified policy* that directly satisfies joint-level physical constraints while optimizing task rewards. Instead of a scalar combination, we project the safety gradient into the null space of the task gradient, maintaining first-order task neutrality. Near the boundary of the recoverable set, a reachability-aware damping schedule smoothly relaxes this projection, allowing safety to oppose the task to escape danger without hand-tuned relative weights.

By internalizing safety boundaries entirely during training, our framework, *Reachability-Guided Policy Optimisation*, requires only one additional backward pass per optimization step. Crucially, the resulting policy deploys as a standard policy architecture, achieving no additional overhead beyond the base policy network evaluation and eliminating the need for runtime safety filters or recovery controllers. To conclude the following are our contributions:

- We formulate safe legged locomotion as a dual-gradient PPO framework that optimizes physical safety compliance without compromising target locomotion performance, decoupling safety and performance signals at training time.
- We propose an *asymmetric null-space projection* that projects reachability-derived safety gradients strictly into the orthogonal null space of the task gradient, guaranteeing first-order task-neutral safety learning under nominal conditions.
- We develop a *reachability-aware damping schedule* and bounded dual scaling that dynamically relax orthogonality near critical boundaries, granting safety gradients active control priority to prevent hardware failures when in imminent danger.
- We validate our approach on a physical quadrupedal robot (Unitree Go1) over stair, climb, and corridor terrains, demonstrating a $94.6\%$ to $96.7\%$ reduction in joint-limit violations under sparse constraints and a strict Pareto improvement in depth-based collision avoidance.

## II. Related Work

Hamilton–Jacobi reachability analysis [18, 4] characterises the maximal controlled-invariant safe set, with recent learned approximations [3, 8, 10, 28] enabling high-dimensional tractability. Traditional integrations operate at deployment time: shielding overrides actions via filters [6, 2], recovery switching activates fallback controllers [7, 9, 29], and constrained-RL approaches employ scalar Lagrangians [1, 26]. Agile But Safe

(ABS) [9] is a representative switching approach, learning a reach-avoid value network in centroidal velocity space to trigger recovery policies. However, ABS focuses on planar collision avoidance and operates at the coarse body-velocity level, leaving joint-level policies vulnerable to physical violations. Our method instead projects reachability gradients directly into the joint-space actor at training time. In legged locomotion, model-free PPO [23, 14, 17] achieves high capability but relies on implicit heuristics; SoloParkour [5] and Kim et al. [12] treat constraints explicitly but combine them at the scalar loss level. Conversely, gradient-surgery methods (e.g., PCGrad [30], CAGrad [15]) project conflicting gradients symmetrically, which fails to capture the asymmetric priority of safety over tasks.

Rather than switching policies online or optimizing competing scalars, we inject safety *at training time* via an *asymmetric* null-space projection (where only the safety gradient is modified to avoid task interference). Its strictness is automatically modulated by the reachability value, yielding a single controller that handles heterogeneous constraints without online filters, recovery networks, or value-based switching. To our knowledge, this is the first framework combining reachability-derived policy gradients and null-space projections for safe locomotion.

## III. Background and Problem Setup

### A. Hamilton–Jacobi Reachability

Consider a dynamical system with state $s \in \mathcal{S}$ and $K$ safety constraints. We encode each constraint as a signed margin function $\ell_k : \mathcal{S} \to \mathbb{R}$, where $\ell_k(s) \geq 0$ denotes safety and $\ell_k(s) < 0$ a violation (e.g., clearance distance). The aggregate margin is $\ell(s) = \min_k \ell_k(s)$, and the safe set is $\mathcal{S}_{\text{safe}} = \{s : \ell(s) \geq 0\}$. Hamilton–Jacobi (HJ) reachability analysis [18, 4] computes the safety value function $V_{\text{safe}}$:

$$V_{\text{safe}}(s) \;=\; \sup_{\pi} \inf_{t \geq 0} \; \ell\big(s_t \mid s_0 = s,\, \pi\big), \tag{1}$$

representing the worst-case safety margin from state $s$ under the optimal safety policy. States with $V_{\text{safe}}(s) \geq 0$ form the controlled-invariant safe set. The gradient $\nabla_s V_{\text{safe}}$ points toward safer states. Classic HJ analysis solves an HJI partial differential equation (PDE) on grids, scaling exponentially with the state dimension and limiting applicability to low-dimensional systems [18].

To bridge this gap, Fisac et al. [8] cast $V_{\text{safe}}$ as a self-consistency Bellman equation:

$$V_{\text{safe}}(s) \;=\; \min\big(\ell(s),\, \max_{a} \; V_{\text{safe}}(s')\big), \tag{2}$$

where $s' \sim \mathcal{P}(\cdot \mid s, a)$. This recursion can be approximated by a learned network $V_{\text{safe}}^{\phi}$ using standard temporal-difference learning, enabling reachability on high-dimensional systems [10, 3]. We jointly train $V_{\text{safe}}^{\phi}$ to provide safety gradients for null-space projection.

### B. Problem Setup

We model safe locomotion as a constrained partially observable MDP $(\mathcal{S}, \mathcal{A}, \mathcal{O}, r, \ell, \mathcal{P}, \gamma)$, where $\mathcal{S}$ is the privileged simulator state, $\mathcal{A}$ is the joint action space, and $\mathcal{O}$ denotes the observation space.

At each time-step $t$, the actor receives a proprioceptive observation history $o_t \in \mathcal{O}$ consisting of a 5-step history of joint angles relative to their nominal positions, joint velocities, projected gravity vector, target linear and angular base velocity commands, and the previous joint action. To incorporate exteroceptive terrain information, the observation $o_t$ additionally includes visual features representing the geometry of the terrain immediately in front of the robot. To handle raw structural visual inputs without relying on computationally expensive convolutional decoders online, we train a kinodynamic internal world model following VIP-Loco [25], which maps raw sequential depth images and proprioceptive streams into compact 64-D exteroceptive latent features integrated directly into $o_t$. The privileged simulator state $s_t \in \mathcal{S}$ elements (only available during training to asymmetric critics) contains these proprioceptive and depth representations as well as privileged properties including the exact base linear velocity, ground reaction forces, contact indicators, terrain height scan heights directly under the robot feet, and local physical parameters (e.g., ground friction and robot mass).

The policy maps observations to joint action targets, $\pi_\theta : \mathcal{O} \to \Delta(\mathcal{A})$. In our implementation, actions $a_t \in \mathcal{A}$ are joint-angle position offsets (12-D for the quadruped morphology) relative to the default nominal stance. These actions are fed to a low-level PD controller operating at a higher frequency to compute corresponding motor torques:

$$\tau_k \;=\; K_p\,(q_{\text{cmd},k} - q_k) \;-\; K_d\,\dot{q}_k, \tag{3}$$

where $q_{\text{cmd},k} = q_{\text{nominal},k} + a_{t,k}$ represents the target angle command for joint $k$.

We use an asymmetric actor-critic workspace where task critic $V_\psi^\pi$ and safety critic $V_\phi^{\text{safe}}$ access state $\mathcal{S}$ during training while the actor evaluates only $o_t$ [20]. Critic network parameters ($\psi$ and $\phi$) are trained via distinct value regression objectives, while actor parameters $\theta$ are updated via policy surrogate gradients. The objective is:

$$\begin{aligned} &\max_{\theta} \; \mathbb{E}_{\pi_\theta}[\textstyle\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)] \\ &\text{s.t.} \;\; V_{\text{safe}}(s_t) \;\geq\; \delta_{safe}, \end{aligned} \tag{4}$$

where $\delta_{safe} \geq 0$ denotes the target safety clearance.

## IV. Methodology

We present a decoupled training-time geometric solution designed to maximize physical safety compliance without compromising locomotion performance. Standard formulations like reward shaping inherently balance safety and velocity, trading off tracking performance. In contrast, we separate task and safety optimization directions. We isolate independent gradients for task progress and safety boundaries (Sec. IV-A), project safety into the null space of the task gradient under nominal conditions (Sec. IV-B), and introduce a reachability-aware damping schedule with adaptive dual scaling to prioritize safety under critical danger (Sec. IV-C), as summarized in Algorithm 1.

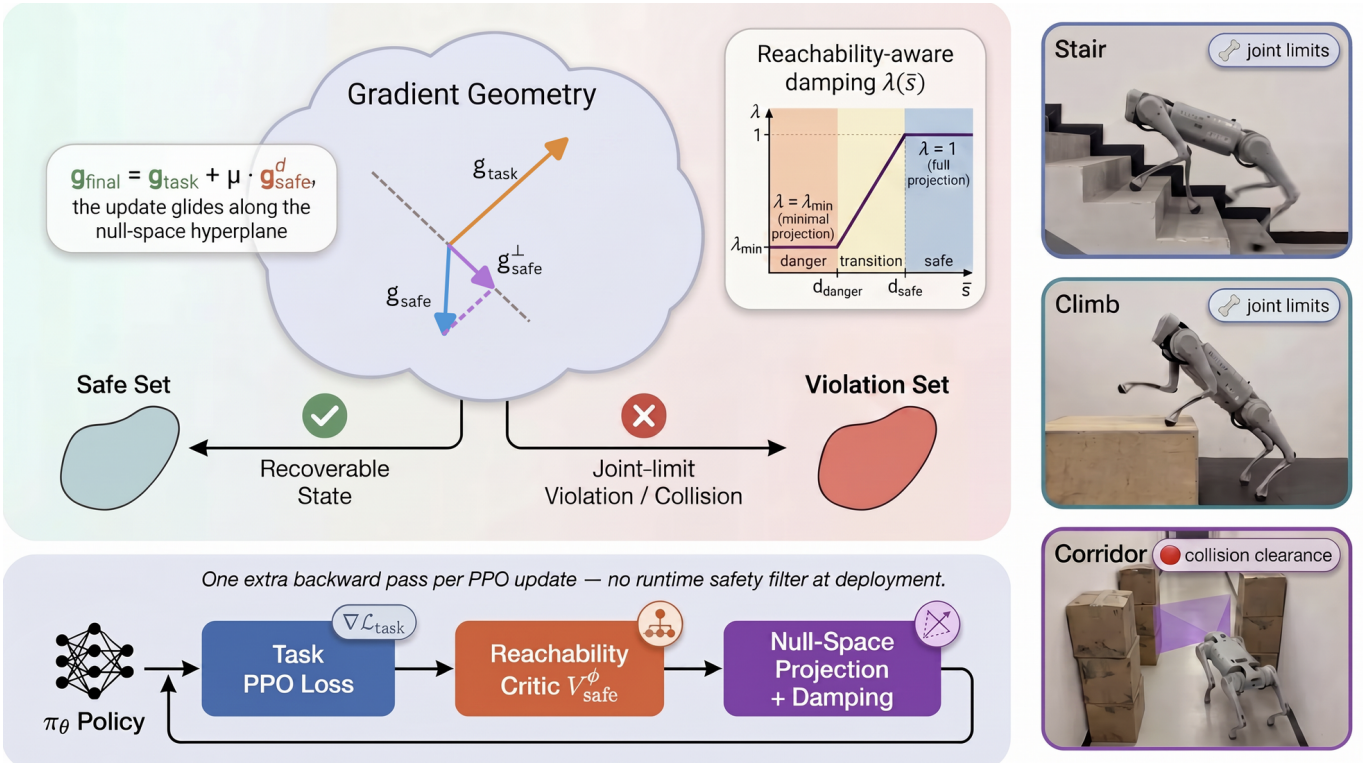


Fig. 1. **Reachability-guided safe policy optimisation.** At every PPO update we compute a task gradient and a reachability-derived safety gradient, then project the safety gradient into the null space of the task gradient with strictness modulated by a reachability-aware damping schedule. The resulting update preserves first-order task improvement while steering the policy toward more recoverable states. Right: representative Unitree Go1 trajectories on stair, climb, and corridor terrains.

### A. Dual-Gradient PPO Formulation

To optimize safety without compromising target performance, we isolate their respective optimization signals. Conflating safety with task progress in joint rewards leads to sub-optimal policies. We thus formulate independent training trajectories to generate distinct gradient fields.

We optimize policy parameters $\theta$ and value-critic parameters $(\psi, \phi)$ via PPO [24]. Rollouts yield task advantages $A_{\text{task}}(t)$ from task reward return $r(s_t, a_t)$ and safety advantages $A_{\text{safe}}(t)$ derived from the learned reachability critic $V^{\phi}_{\text{safe}}$. To prevent temporal compensation (using long stretches of safe trajectories to mask deep, localized violations), we propagate worst-case safety performance backward through a recursive temporal logic operator:

$$R_{\text{safe}}(t) \;=\; \alpha \cdot \min\big(\ell(s_t),\, R_{\text{safe}}(t{+}1)\big) \;+\; (1-\alpha)\cdot \ell(s_t), \quad (5)$$

with the advantages derived as $A_{\text{safe}}(t) = R_{\text{safe}}(t) - V^{\phi}_{\text{safe}}(o_t)$, where $\alpha \in (0, 1)$ weights worst-case propagation and prevents temporal discounting from diluting proximal failures. The task surrogate objective for the actor $\theta$ is:

$$L_{\text{task}}(\theta) \;=\; \mathbb{E}_t\big[\min\big(\rho_t A_{\text{task}}(t),\, \text{clip}(\rho_t, 1{-}\epsilon, 1{+}\epsilon)\, A_{\text{task}}(t)\big)\big] + c_e H[\pi_\theta], \quad (6)$$

while the separate task value critic $\mathcal{L}^{\text{task}}_V(\psi)$, reachability actor surrogate $L_{\text{safe}}(\theta)$, and reachability value critic $\mathcal{L}_{V_{\text{safe}}}(\phi)$ objectives are:

$$\mathcal{L}^{\text{task}}_V(\psi) \;=\; \mathbb{E}_t\big[(V_\psi(s_t) - G^{\text{task}}_t)^2\big], \quad (7)$$

$$L_{\text{safe}}(\theta) \;=\; \mathbb{E}_t\big[\min\big(\rho_t A_{\text{safe}}(t),\, \text{clip}(\rho_t, 1{-}\epsilon, 1{+}\epsilon)\, A_{\text{safe}}(t)\big)\big], \quad (8)$$

$$\mathcal{L}_{V_{\text{safe}}}(\phi) \;=\; \mathbb{E}_t\big[(V^{\phi}_{\text{safe}}(s_t) - R_{\text{safe}}(t))^2\big], \quad (9)$$

where $\rho_t$ represents the probability ratio of successive policy parameters.

From these separate loss formulations, we obtain independent gradient directions on the joint policy parameters $\theta$:

$$g_{\text{task}} \;=\; \nabla_\theta L_{\text{task}}(\theta) \quad \text{and} \quad g_{\text{safe}} \;=\; \nabla_\theta L_{\text{safe}}(\theta). \quad (10)$$

These isolated vector fields represent pure updates: $g_{\text{task}}$ acts to maximize forward velocity tracking, while $g_{\text{safe}}$ pulls parameters strictly toward greater reachability margins. Simply summing them $(g_{\text{task}} + g_{\text{safe}})$ forces direct competition and gradient cancellation, triggering severe training-time performance oscillations. This motivates our training-time decoupled geometric integration.

### B. Asymmetric Null-Space Projection

Prior formulations typically address gradient conflict symmetrically by projecting both gradients onto mutually agreeable directions [30] or by executing hard binary switches online [9].

However, online switching triggers destabilizing control handoffs on hardware, while symmetric training-time projections degrade the baseline locomotion trajectory defined by $g_{\text{task}}$.

To sidestep these issues, we project the safety gradient $g_{\text{safe}}$ strictly into the null space of the task gradient $g_{\text{task}}$, leaving the primary locomotion direction completely uncorrupted.

$$g_{\text{safe}}^{\perp} \;=\; g_{\text{safe}} \;-\; \frac{g_{\text{safe}}^{\top} g_{\text{task}}}{\|g_{\text{task}}\|^2 + \epsilon^{\perp}}\, g_{\text{task}}, \tag{11}$$

which guarantees $g_{\text{safe}}^{\perp} \cdot g_{\text{task}} = 0$ (ignoring numerical stabilizer $\epsilon^{\perp} > 0$). Under nominal safety conditions ($\lambda = 1$), the update direction is $g_{\text{final}} = g_{\text{task}} + \mu g_{\text{safe}}^{\perp}$. Since $g_{\text{safe}}^{\perp} \cdot g_{\text{task}} = 0$, the first-order update projection is $g_{\text{task}}^{\top} g_{\text{final}} = \|g_{\text{task}}\|^2$, which is identical to the pure task update $g_{\text{task}}$. Thus, the first-order variation has $L_{\text{task}}(\theta + \eta\, g_{\text{final}}) = L_{\text{task}}(\theta) + \eta \|g_{\text{task}}\|^2 + O(\eta^2)$, confirming that task improvement is mathematically decoupled from both the safety gradient and the dual scale $\mu$. This asymmetric design ensures that the policy internalizes safer locomotion dynamics without sacrificing baseline performance.

### C. Damped Optimization and Adaptive Scaling

Under nominal conditions, this strict orthogonal relation is ideal. However, as the robot approaches the boundaries of its recoverable set, safety can no longer remain task-neutral; it must be permitted to directly oppose and override task commands (such as braking or retreating) to prevent hardware failures.

To modulate this priority shift dynamically, we scale projection strictness via a reachability-aware damping schedule. Let $\bar{s} = \min_{t \in \text{batch}} V_{\text{safe}}^{\phi}(s_t)$ capture the worst-case reachability value within a rollout sequence. The damping extraction fraction $\lambda(\bar{s})$ is defined as:

$$\lambda(\bar{s}) \;=\; \lambda_{\min} \;+\; (1 - \lambda_{\min}) \cdot \text{clamp}\left( \frac{\bar{s} - d_{\text{danger}}}{d_{\text{safe}} - d_{\text{danger}}},\; 0,\; 1 \right), \tag{12}$$

which smoothly interpolates between nominal operation ($\lambda = 1$) for strict task protection and critical danger ($\lambda = \lambda_{\min}$), letting safety actively overpower task-oriented commands. The damped safety gradient is formulated as:

$$g_{\text{safe}}^{d} \;=\; g_{\text{safe}} \;-\; \lambda(\bar{s}) \cdot \frac{g_{\text{safe}}^{\top} g_{\text{task}}}{\|g_{\text{task}}\|^2 + \epsilon^{\perp}}\, g_{\text{task}}. \tag{13}$$

While damping handles geometry-based priority shifts, sparse constraints (like joint limits) produce weak gradients inside safe subsets, enabling drift near thresholds. We resolve this by incorporating an adaptive dual variable multiplier $\mu \in [\mu_{\min}, \mu_{\max}]$ to scale safety authority based on tracking criteria. Specifically, we track the exponential moving average (EMA) of episode joint-violation counts, denoted as $\bar{C}_{\text{viol}}$. The multiplier is updated at the end of each training iteration via projected dual ascent:

$$\mu \;\leftarrow\; \text{clip}\big(\, \mu \;+\; \eta_{\mu} \big( \bar{C}_{\text{viol}} - C_{\text{target}} \big),\; \mu_{\min},\; \mu_{\max} \big)\,, \tag{14}$$

where $\eta_{\mu} = 0.001$ is the dual step size, and $C_{\text{target}} = 5$ represents the target maximum joint violation count per episode. The final integrated update step is:

$$g_{\text{final}} \;=\; g_{\text{task}} \;+\; \mu\, g_{\text{safe}}^{d}, \tag{15}$$

which is clipped to a maximum norm of $g_{\max}$ and optimized via Adam. This formulation guarantees (i) first-order task neutrality within nominal safety boundaries (where $\lambda = 1$, rendering $g_{\text{task}}^{\top} g_{\text{final}} = \|g_{\text{task}}\|^2$); (ii) reachability-driven safety priority that scales naturally with physical risk $\bar{s}$; and (iii) training-time internalization, allowing end-to-end deployment as a single feedforward neural network.

## V. Experiments

We evaluate our reachability-guided null-space framework on a 12-DoF quadrupedal robot (Unitree Go1) to demonstrate its ability to maximize physical safety compliance on complex terrains under two distinct classes of physical safety specifications: sparse proprioceptive constraints and dense perceptual constraints.

To ground our mathematical framework empirically, we instantiate the signed safety margin $\ell(s_t)$ from Eq. (5) under these two different constraint regimes:

**Proprioceptive Joint-Limit Avoidance:** For each joint $j$ with positional boundaries $[q_j^{\min}, q_j^{\max}]$, the safety margin is:

$$\ell_j \;=\; \min\!\big(q_j - q_j^{\min},\; q_j^{\max} - q_j\big) \;-\; d_{\text{margin}}, \tag{16}$$

where $d_{\text{margin}} > 0$ is a clearance buffer. This represents a *sparse* boundary constraint, emitting nonzero gradients only in close proximity to physical extremities.

**Perceptual Collision Avoidance:** For 3D obstacle boundaries detected by depth sensors, the clearance margin incorporates both the spatial clearance to the obstacle $h = d_{\text{obs}} - d_{\text{safe}}^{\text{col}}$ and its temporal approach velocity $\dot{h}$ to preemptively steer in fast dynamic regimes:

$$\ell_{\text{col}}(h, \dot{h}) \;=\; h \;+\; \alpha_h \dot{h}, \tag{17}$$

where $d_{\text{safe}}^{\text{col}}$ is the safety radius clearance and $\alpha_h = 0.3$. This approach-rate formulation triggers lateral evasive adjustments proportional to the rate of approach.

We transition seamlessly between these settings simply by swapping the underlying sensor-driven margin function $\ell(s_t)$ within our reachability formulation. Our experiments are designed to address two focal questions: **Q1:** *Does training-time geometric projection effectively eliminate joint-limit violations under sparse signals without eroding agile climbing performance? (Section V-B)* and **Q2:** *Can geometric updates achieve a task–safety Pareto improvement under continuously active perceptual signals? (Section V-C)*

### A. Experimental Setup

We construct three challenging terrains in simulation: **stair** (regular staircase ascent), **climb** (irregular height steps), and **corridor** (cluttered tracks with horizontal obstacles). We evaluate performance against two primary baselines: (i) a standard **reward-shaping formulation** [21] that joins the safety margin $\ell(s_t)$ into the performance objective as a scalar penalty; and (ii) a **PPO Lagrangian** formulation [22, 26]

**Algorithm 1** Reachability-Guided Null-Space Policy Optimisation

**Input:** actor $\theta$, task critic $V^{\psi}_{\text{task}}$, reachability critic $V^{\phi}_{\text{safe}}$, dual $\mu \in [\mu_{\min}, \mu_{\max}]$, thresholds $d_{\text{safe}} > d_{\text{danger}}$, target limit $C_{\text{target}} = 5$
**for** iteration $= 1, 2, \ldots$ **do**
  Collect rollout buffer $\mathcal{D}$ with $\pi_\theta$
  Compute $\ell(s_t)$ and reachability return $R_{\text{safe}}(t)$ via Eq. (5)
  Compute task and safety advantages $A_{\text{task}}(t)$, $A_{\text{safe}}(t)$ via GAE
  **for** epoch $= 1, \ldots, K$ **do**
    $g_{\text{task}} \leftarrow \nabla_\theta L_{\text{task}}(\theta; \mathcal{D})$ ▷ task PPO loss, Eq. (6)
    Compute backward pass: $g_{\text{safe}} \leftarrow \nabla_\theta L_{\text{safe}}(\theta; \mathcal{D})$ ▷ Eq. (8)
    Compute $\bar{s} = \min_{t \in \mathcal{D}} V^{\phi}_{\text{safe}}(s_t)$ and $\lambda(\bar{s})$ via Eq. (12)
    $g^{d}_{\text{safe}} \leftarrow g_{\text{safe}} - \lambda(\bar{s}) \cdot \frac{g_{\text{safe}} g_{\text{task}}}{\|g_{\text{task}}\|^2 + \epsilon^{\perp}} g_{\text{task}}$ ▷ Eq. (13)
    $g_{\text{final}} \leftarrow \text{clip_norm}\big(g_{\text{task}} + \mu\, g^{d}_{\text{safe}},\, g_{\max}\big)$
    Update $\theta$ with Adam; update $\psi, \phi$ with their value-regression gradients
  **end for**
  Update violation EMA: $\bar{C}_{\text{viol}} \leftarrow (1-\beta)\bar{C}_{\text{viol}} + \beta C_{\text{viol}}$
  $\mu \leftarrow \text{clip}\big(\mu + \eta_\mu(\bar{C}_{\text{viol}} - C_{\text{target}}),\, \mu_{\min},\, \mu_{\max}\big)$
**end for**

that updates a scalar dual variable multiplier on the safety boundary. All controllers share identical network capacities and PPO hyperparameters. Proprioceptive configurations and domain randomizations are itemized in Appendix I, with all safety parameters cataloged in Appendix A.

### *B. Joint-Limit Avoidance and Task–Safety Tradeoff*

The reachability-guided null-space policy reduces joint violations per episode by $96.7\%$ on **climb** ($2.77 \pm 7.80$ vs. $84.13 \pm 141.85$) and $94.6\%$ on **stair** ($3.70 \pm 7.34$ vs. $68.07 \pm 76.03$), falling per-step to $0.0029$ and $0.0039$, respectively (Table I). Entirely violation-free episodes rise to $45.3\%$ (climb) and $38.3\%$ (stair). The mean safety margin $\ell$ remains positive in both regimes ($+0.086$ on climb; $+0.098$ on stair), steering the policy toward safer joint-space operating regions.

On **stair**, the null-space policy simultaneously attains the highest mean return ($35.01 \pm 2.06$) and the lowest velocity tracking error ($0.101$ m/s) among all methods, acting as a natural gait regularizer that stabilizes the locomotion cycle. On the irregular steps of **climb**, our method accepts a modest return reduction ($22.37 \pm 8.64$) in exchange for the large safety gain, reflecting active damping corrections triggered by chaotic foothold geometries. Detailed breakdown of joint safety margins is provided in Appendix C.

### *C. Depth-Based Collision Avoidance*

The collision-avoidance evaluation tests our framework using rich, dense perceptual signals on two complementary perceptual terrains: the **corridor**, whose sinusoidally winding walls form a spatially correlated, partially observable obstacle field, and the **scattered-obstacle** terrain (Appendix G), whose independently placed pillars form a spatially decorrelated, locally observable field. Quantitative values for both are summarized in Table II.

*a) Corridor (spatially correlated, partially observable).:* On the corridor, our null-space policy reduces the collision rate by $68.3\%$ relative to reward shaping ($13.9\% \rightarrow 4.4\%$) and by $63.6\%$ relative to PPO Lagrangian ($12.1\% \rightarrow 4.4\%$). Crucially, this is a strict Pareto improvement: our method achieves the highest mean return ($33.9 \pm 2.4$) and the lowest velocity tracking error ($0.085$ m/s), while dramatically compressing performance variance ($\sigma = 2.4$). Ours also yields a $73\%$ increase in mean lateral velocity ($0.085$ m/s vs. $0.049$ m/s for basic shaping), showing that the training-time geometric projection discovers active, lateral evasive behaviors naturally without any manual reward engineered for sidestepping. By contrast, the PPO Lagrangian baseline, which penalizes collision proximity on a scalar level, fails to find these task-compatible parameters and yields a $12.1\%$ collision rate. This isolates the source of our gain: it is the *geometric mode* of integration, rather than simply having constraint values in a scalar objective, that delivers the improvement.

*b) Scattered-obstacle (spatially decorrelated, locally observable):* On the scattered-obstacle terrain our method again attains the lowest collision rate ($2.7\%$), a $41.3\%$ reduction relative to Reward Shaping ($4.6\%$) and a $40.0\%$ reduction relative to PPO Lagrangian ($4.5\%$), while also achieving the highest mean and minimum safety values. As on the corridor, the safety gain is realised through active lateral evasion: mean lateral velocity roughly doubles ($0.078$ vs. $0.033$ m/s) and mean yaw rate increases ($0.371$ vs. $0.123$ rad/s), indicating that the policy actively steers around obstacles. Here the gain is accompanied by a modest reduction in mean return ($24.06 \pm 4.76$ vs. $27.13 \pm 3.29$ for Reward Shaping), so unlike the corridor the scattered setting is a favourable safety–return trade-off rather than a strict Pareto improvement. The three methods are markedly closer here than on the corridor because each scattered obstacle is spatially isolated and locally observable, so short-horizon reactive evasion already suffices. Trajectory kinematics and yaw-rate profiles are analyzed in Appendix D.

TABLE I

JOINT-LIMIT SAFETY RESULTS OVER $N = 1{,}000$ EPISODES (MEAN $\pm$ STD). $\downarrow$ LOWER IS BETTER; $\uparrow$ HIGHER IS BETTER. BEST IN **BOLD**.

| | **Climb** | | | **Stair** | | |
|---|---|---|---|---|---|---|
| Metric | Reward Shaping | PPO Lagrangian | Ours | Reward Shaping | PPO Lagrangian | Ours |
| Return $\uparrow$ | $25.88 \pm 7.19$ | $\mathbf{26.70 \pm 6.65}$ | $22.37 \pm 8.64$ | $25.40 \pm 4.20$ | $27.37 \pm 6.57$ | $\mathbf{35.01 \pm 2.06}$ |
| Viol./episode $\downarrow$ | $84.13 \pm 141.85$ | $109.32 \pm 197.58$ | $\mathbf{2.77 \pm 7.80}$ | $68.07 \pm 76.03$ | $38.91 \pm 46.34$ | $\mathbf{3.70 \pm 7.34}$ |
| Viol./step $\downarrow$ | $0.086 \pm 0.142$ | $0.110 \pm 0.197$ | $\mathbf{0.003 \pm 0.008}$ | $0.069 \pm 0.077$ | $0.040 \pm 0.058$ | $\mathbf{0.004 \pm 0.009}$ |
| Eps. with viol. $\downarrow$ | 91.7% | 92.8% | **54.7%** | 99.0% | 94.1% | **61.7%** |
| Mean $\ell$ $\uparrow$ | $\mathbf{+0.088 \pm 0.015}$ | $+0.044 \pm 0.024$ | $+0.086 \pm 0.011$ | $+0.079 \pm 0.013$ | $-0.008 \pm 0.038$ | $\mathbf{+0.098 \pm 0.002}$ |
| Vel. error (m/s) $\downarrow$ | $0.190 \pm 0.085$ | $\mathbf{0.175 \pm 0.070}$ | $0.235 \pm 0.097$ | $0.210 \pm 0.043$ | $0.180 \pm 0.062$ | $\mathbf{0.101 \pm 0.025}$ |

TABLE II

DEPTH-BASED COLLISION AVOIDANCE ON THE CORRIDOR AND SCATTERED-OBSTACLE TERRAINS OVER $N \approx 1{,}000$ EPISODES (MEAN $\pm$ STD). FOR THE CORRIDOR THE REPORTED SAFETY VALUE IS THE CLEARANCE MARGIN $\ell$; FOR THE SCATTERED-OBSTACLE TERRAIN IT IS THE LEARNED SAFETY VALUE $V^{\phi}_{\text{safe}}$. $\downarrow$ LOWER IS BETTER; $\uparrow$ HIGHER IS BETTER. BEST PER TERRAIN IN **BOLD**.

| | **Corridor** | | | **Scattered-obstacle** | | |
|---|---|---|---|---|---|---|
| Metric | Reward Shaping | PPO Lagrangian | Ours | Reward Shaping | PPO Lagrangian | Ours |
| Return $\uparrow$ | $30.3 \pm 8.7$ | $32.0 \pm 6.6$ | $\mathbf{33.9 \pm 2.4}$ | $\mathbf{27.13 \pm 3.29}$ | $26.13 \pm 4.49$ | $24.06 \pm 4.76$ |
| Collision rate $\downarrow$ | 13.9% | 12.1% | **4.4%** | 4.6% | 4.5% | **2.7%** |
| Eps. with collision $\downarrow$ | 83.9% | 80.8% | **71.6%** | 47.2% | 46.5% | **45.5%** |
| Mean safety value $\uparrow$ | $-0.046 \pm 0.102$ | $-0.062 \pm 0.075$ | $\mathbf{+0.086 \pm 0.038}$ | $0.292 \pm 0.461$ | $0.322 \pm 0.281$ | $\mathbf{0.339 \pm 0.289}$ |
| Min safety value $\uparrow$ | $-1.975$ | $-1.649$ | $\mathbf{-0.538}$ | $-0.861$ | $-0.841$ | $\mathbf{-0.807}$ |
| Vel. error (m/s) $\downarrow$ | $0.217 \pm 0.161$ | $0.152 \pm 0.142$ | $\mathbf{0.085 \pm 0.088}$ | $\mathbf{0.082 \pm 0.108}$ | $0.105 \pm 0.126$ | $0.096 \pm 0.142$ |
| Mean \|lateral vel\| (m/s) | $0.049 \pm 0.056$ | $0.052 \pm 0.057$ | $\mathbf{0.085 \pm 0.066}$ | $0.033 \pm 0.157$ | $0.043 \pm 0.057$ | $0.078 \pm 0.083$ |
| Collision reduction vs. baseline | – | 12.9% | **68.3%** | – | 2.2% | **41.3%** |

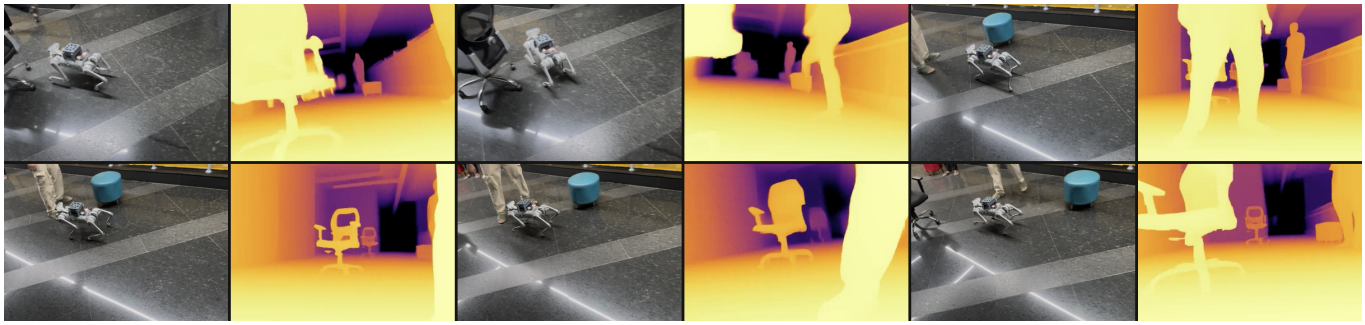

Fig. 2. **Qualitative snapshots of obstacle interaction.** Snapshots of the quadruped robot safely navigating around obstacles. Our reachability-guided null-space policy enables smooth steering compared to baselines.

## VI. DISCUSSION AND CONCLUSION

We present a training-time reachability-guided policy optimization framework using asymmetric null-space projections to satisfy physical safety constraints without compromising legged locomotion performance. By isolating safety corrections from locomotion goals via reachability-aware filtering, our formulation resolves critical task-safety trade-offs that standard scalar penalty profiles fail to balance. When evaluated in simulation and deployed on a quadrupedal robot (Unitree Go1), our method reduces joint limit violations by over 94% and corridor collisions by 68.3% while maintaining agile velocity tracking, all with zero runtime overhead under standard vision-proprioceptive policy architectures.

*a) Limitations & Future Work:* First, our framework utilizes neural reachability approximations, acting as a guided regularization rather than providing formal analytical guarantees. Second, the projection is currently confined to a single task-safety ranking; handling multiple competing constraints simultaneously requires multi-objective projection metrics. Finally, noisy task gradients can slow down safety convergence. Future work will investigate non-linear damping profiles and multi-constraint scaling vectors to deliver scalable, mathematically integrated safety for high-dimensional robotic systems.

**Full Paper with Appendix:** Here

**Simulation and Hardware Videos are available at** https://safeloco-rl.github.io/.

## REFERENCES


[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning (ICML)*, 2017.

[2] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *AAAI Conference on Artificial Intelligence*, 2018.

[3] Somil Bansal and Claire J. Tomlin. DeepReach: A deep learning approach to high-dimensional reachability. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[4] Somil Bansal, Mo Chen, Sylvia Herbert, and Claire J. Tomlin. Hamilton–Jacobi reachability: A brief overview and recent advances. In *IEEE Conference on Decision and Control (CDC)*, 2017.

[5] Elliot Chane-Sane, Joseph Amigo, Thomas Flayols, Ludovic Righetti, and Nicolas Mansard. SoloParkour: Constrained reinforcement learning for visual locomotion from privileged experience. In *Conference on Robot Learning (CoRL)*, 2024.

[6] Richard Cheng, Gábor Orosz, Richard M. Murray, and Joel W. Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *AAAI Conference on Artificial Intelligence*, 2019.

[7] Benjamin Eysenbach, Shixiang Gu, Julian Ibarz, and Sergey Levine. Leave no trace: Learning to reset for safe and autonomous reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2018.

[8] Jaime F. Fisac, Neil F. Lugovoy, Vicenç Rubies-Royo, Shromona Ghosh, and Claire J. Tomlin. Bridging Hamilton–Jacobi safety analysis and reinforcement learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[9] Tairan He, Chong Zhang, Wenli Xiao, Guanqi He, Changliu Liu, and Guanya Shi. Agile but safe: Learning collision-free high-speed legged locomotion. In *Robotics: Science and Systems (RSS)*, 2024.

[10] Kai-Chieh Hsu, Vicenç Rubies-Royo, Claire J. Tomlin, and Jaime Fernandez Fisac. Safety and liveness guarantees through reach-avoid reinforcement learning. *arXiv preprint arXiv:2112.12288*, 2023.

[11] Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26), 2019.

[12] Yunho Kim, Hyunsik Oh, Jeonghyun Lee, Jinhyeok Choi, Gwanghyeon Ji, Moonkyu Jung, Donghoon Youm, and Jemin Hwangbo. Not only rewards but also constraints: Applications on legged robot locomotion. *IEEE Transactions on Robotics*, 2024.

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[14] Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science Robotics*, 5(47), 2020.

[15] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[16] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Gu, Miles Macklin, Gavriel David, Alex Knyazev, Lukasz Hanzel, Hao State, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2110.19106*, 2021.

[17] Gabriel B. Margolis and Pulkit Agrawal. Walk these ways: Tuning robot control for generalization with multiplicity of behavior. In *Conference on Robot Learning (CoRL)*, 2023.

[18] I.M. Mitchell, A.M. Bayen, and C.J. Tomlin. A time-dependent hamilton-jacobi formulation of reachable sets for continuous dynamic games. *IEEE Transactions on Automatic Control*, 50(7):947–957, 2005. doi: 10.1109/TAC.2005.851439.

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

[20] Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. Asymmetric actor critic for image-based robot learning. In *Robotics: Science and Systems (RSS)*, 2018.

[21] Abhishek Ranjan, Shreenabh Agrawal, Aayush Jain, Pushpak Jagtap, Shishir Kolathaya, et al. Barrier functions inspired reward shaping for reinforcement learning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10807–10813. IEEE, 2024.

[22] Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 2019.

[23] Nikita Rudin, David Hoeller, Philipp Reist, and Marco Hutter. Learning to walk in minutes using massively parallel deep reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2022.

[24] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[25] Aditya Shirwatkar, Satyam Gupta, and Shishir Kolathaya. VIP-Loco: A visually guided infinite horizon planning framework for legged locomotion. *arXiv preprint arXiv:2603.14345*, 2026.

[26] Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID Lagrangian methods. *arXiv preprint arXiv:2007.03964*, 2020.

[27] Manan Tayal, Aditya Singh, Pushpak Jagtap, and Shishir Kolathaya. Semi-supervised safe visuomotor policy synthesis using barrier certificates. In *2025 IEEE 64th Conference on Decision and Control (CDC)*, pages 4831–4837. IEEE, 2025.

[28] Mumuksh Tayal, Manan Tayal, and Ravi Prakash. Safe flow Q-learning: Offline safe reinforcement learning with reachability-based flow policies. *arXiv preprint arXiv:2603.15136*, 2026.

[29] Brijen Thananjeyan, Ashwin Balakrishna, Suraj Nair, Michael Luo, Krishnan Srinivasan, Minho Hwang, Joseph E. Gonzalez, Julian Ibarz, Chelsea Finn, and Ken Goldberg. Recovery RL: Safe reinforcement learning with learned recovery zones. *IEEE Robotics and Automation Letters*, 2021.

[30] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.