

Curriculum-Based Deep Reinforcement Learning for Robust Biped Push Recovery

Luciana Ercolanese¹, Riccardo Aliotta^{2,3}, Pierluigi Arpentì², Fabio Ruggiero², Vincenzo Lippiello²

Abstract—Push recovery—the ability to regain balance after unexpected external disturbances—is a key capability for deploying biped robots in real-world environments. Existing approaches often rely on simplified template models, which can be overly conservative and fail to capture realistic recovery behaviors under disturbances that vary in location, magnitude, and duration. This paper presents a push-recovery framework that combines a sequential curriculum learning strategy with massively parallel deep reinforcement learning. By training the policy on progressively more challenging disturbance scenarios, the proposed method enables robust recovery from pushes applied at different body locations with varying intensity and duration. Simulation results demonstrate strong performance, including demanding dual-disturbance cases with simultaneous external forces.

Index Terms—Intelligent Systems, Neural Control, Robotics

I. INTRODUCTION

In recent years, humanoid robots have attracted growing attention due to their wide range of potential applications [1], from social robotics and human–robot interaction [2] to operation in hazardous environments such as search-and-rescue missions [3]. These scenarios require safe mobility in unstructured environments, which remains challenging for humanoids. Due to their inherently unstable single-support gait, biped robots must maintain dynamic balance while coping with terrain variations and unexpected external forces.

In this context, push recovery—the ability to withstand sudden disturbances without falling—is a fundamental capability for robots operating in human-centered environments [4]. Beyond humanoids, this capability is also critical for assistive exoskeletons supporting locomotor-impaired users, where robustness to unexpected perturbations is essential for safety and reliability [5]. Push recovery has traditionally been addressed through model-based approaches, ranging from simplified template models to whole-body control (WBC) frameworks. However, maintaining real-time tractability often requires simplified dynamics, e.g., template models in model predictive control (MPC), which poorly capture the nonlinear and hybrid nature of biped locomotion. As a result, such methods often produce conservative behaviors and remain limited to disturbances applied at specific locations or within restricted force ranges.

This work was partially supported by the Italian Ministry of University and Research, under the complementary actions to the NRRP ‘Fit4MedRob - Fit for Medical Robotics’ Grant (# PNC0000007).

¹ Authors are with The BioRobotics Institute, Scuola Superiore Sant’Anna, 56025 Pontedera, Pisa, Italy, ²Authors are with PRISMA Lab and CREATE Consortium, Department of Electrical Engineering and Information Technology, University of Naples Federico II, Naples, Italy, ³Authors are with Department of Agricultural Sciences, University of Naples Federico II, Portici, Italy. Corresponding author’s email: [pierluigi.arpentì@unina.it]

At the other end of the spectrum, Deep Reinforcement Learning (DRL) enables the emergence of recovery behaviors directly from task interaction, without relying on restrictive modeling assumptions.

In this work, we propose a DRL-based push-recovery strategy founded on a curriculum-learning framework, in which a humanoid policy is trained by progressively increasing disturbance magnitude over 12 distinct body contact points. The training process is carried out within a privileged asymmetric actor–critic architecture, where the critic has access to disturbance information that is not available to the actor at deployment. Extensive evaluations under randomized disturbances demonstrate strong robustness, showing that the learned policy can reject forces up to five times larger than those tolerated by a baseline walking controller, even in challenging dual-disturbance scenarios.

The remainder of the paper is organized as follows: Section II reviews related work, Section III introduces the DRL background, Section IV presents the proposed framework, Section V reports the simulation results, and Section VI concludes the paper.

II. RELATED WORKS

Traditionally, push recovery has been addressed using model-based control strategies. Among the most established approaches are capture point methods, where swing-foot placement is computed to maintain balance [6]. These approaches are typically grounded in the Linear Inverted Pendulum Model (LIPM), enabling stabilization mechanisms such as ankle, hip, and stepping strategies [4]. In parallel, WBC frameworks have been developed to achieve robust locomotion behaviors, including stair climbing [7] and disturbance rejection [8]. MPC is now widely adopted for push recovery: early formulations regulated the center of mass (CoM) within the support polygon [9], while more recent approaches enable adaptive stepping and dynamic behaviors such as jumping [10], [11].

However, these methods typically rely either on accurate and computationally demanding dynamic models or on simplified template models (e.g., the LIPM) to maintain real-time tractability. This mismatch with the highly nonlinear and hybrid nature of biped locomotion often leads to conservative behaviors and limited robustness to disturbances applied at specific locations or within restricted force ranges.

DRL has recently emerged as an alternative paradigm for robust legged locomotion. In particular, massively parallel DRL (MPDRL) exploits GPU-based simulation to train large populations of robots simultaneously, significantly reducing

training time. This approach has demonstrated strong results in quadruped locomotion [12] and humanoid push recovery [13]. Robustness is typically achieved through domain randomization, exposing policies to diverse disturbances and terrain conditions [14]–[16]. However, learning complex behaviors from scratch remains challenging, often resulting in slow convergence. Curriculum learning addresses this issue by progressively increasing task difficulty, enabling incremental skill acquisition. Nevertheless, although effective in legged locomotion [17], [18], prior curriculum-learning approaches often operate only implicitly, by shaping the distribution of environment parameters and relying on Proximal Policy Optimization (PPO) stochasticity, rather than by enforcing a structured and controlled progression of disturbance complexity.

Building on these insights, the contribution of this work lies not in a new learning algorithm but in a structured training procedure. We introduce a disturbance-aware curriculum that progressively increases push magnitude, duration, and application point while leveraging MPDRL for scalable training. Combined with a privileged asymmetric actor–critic architecture, this framework enables learning robust whole-body push recovery without requiring explicit force sensing at deployment.

III. METHODOLOGICAL BACKGROUND

In this section, we briefly recall the main concepts underlying the proposed push-recovery framework; the reader is referred to the cited literature for further details.

A. Massively Parallel Deep Reinforcement Learning

DRL combines reinforcement learning with neural networks to approximate policies and value functions in high-dimensional spaces. The environment is modeled as a Markov Decision Process (MDP) $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, $P(s' | s, a)$ the transition probability, $r(s, a)$ the reward function, and $\gamma \in [0, 1)$ the discount factor. The goal is to learn a policy $\pi_\theta(a | s)$ maximizing the expected discounted return $J(\theta) = \mathbb{E}_{\pi_\theta}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$. The associated state-value and action-value functions are $V^\pi(s) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s]$ and $Q^\pi(s, a) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a]$, respectively. In particular, a key advantage of MPDRL is its ability to exploit large-scale domain randomization. By simulating thousands of environments in parallel and varying physical parameters, MPDRL exposes the policy to a broad range of dynamics and interaction conditions, thereby improving generalization and robustness.

B. Asymmetric Actor-Critic Policy Architecture

A common approach to solving MDPs relies on temporal-difference (TD) methods, which update value estimates using the observed reward and the estimated value of the next state. Actor–critic algorithms combine TD-based value learning with policy optimization. In this framework, the actor

represents the policy $\pi_\theta(a | s)$, while the critic approximates a value function $V_\phi(s)$ or $Q_\phi(s, a)$ with parameters ϕ . The policy parameters are updated using the policy gradient theorem $\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(a_t | s_t) A^\pi(s_t, a_t)]$, where $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$ is the advantage function.

In practice, the actor (parameters $\theta \in \mathbb{R}^{n_\theta}$) defines a stochastic policy and selects actions, while the critic (parameters $\phi \in \mathbb{R}^{n_\phi}$) evaluates them through a value function. The critic is trained via TD updates, whereas the actor is optimized through policy gradients using the critic’s estimates as a learning signal. Among policy-gradient methods, PPO [19] is widely adopted due to its simplicity and stability. Unlike Trust Region Policy Optimization, which enforces a hard Kullback-Leibler divergence constraint via second-order optimization, PPO employs a clipped objective to limit policy updates. Let $\pi_{\theta_{\text{old}}}(a | s)$ denote the previous policy and $\pi_\theta(a | s)$ the updated one. The probability ratio $r_t(\theta) = \frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)}$ is used to define the surrogate objective $L^{\text{CLIP}}(\theta) = \mathbb{E}_t[\min(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t)]$, where ϵ (typically 0.2) limits deviations from the previous policy.

In particular, the asymmetric actor–critic framework [20] can be interpreted as a form of privileged learning built upon actor–critic methods such as PPO. During training, the critic receives privileged disturbance information available only in simulation, while the actor is restricted to realistic onboard observations. This improves value estimation and accelerates learning without requiring additional sensing at deployment.

C. Curriculum Learning

A major limitation of reinforcement learning is slow convergence when complex tasks are presented at full difficulty from the outset. Curriculum learning [21] addresses this issue by progressively increasing task difficulty according to training progress, thereby improving learning efficiency and accelerating convergence.

IV. PUSH-RECOVERY LEARNING FRAMEWORK

This section presents the main contribution of the paper: the integration of an asymmetric actor–critic framework with curriculum learning for push recovery.

A. Disturbance-Aware Asymmetric Actor–Critic

Let $\mathbf{F}_{\text{ext}} \in \mathbb{R}^3$ denote the external pushing force. In the proposed framework, only the critic has access to disturbance information, while the actor remains force-agnostic, resulting in an asymmetric actor–critic architecture. This choice is motivated by deployment: although disturbance magnitude $\|\mathbf{F}_{\text{ext}}\|$ and direction $\angle \mathbf{F}_{\text{ext}}$ are readily available in simulation, obtaining them on a real robot would require additional sensing or online estimation, increasing system complexity and cost. Restricting this privileged information to the critic during training improves value estimation and accelerates learning, while preserving a deployable actor based solely on realistic onboard measurements. A summary of the observation vector and its partition between actor and critic is reported in Table I.

TABLE I: Observation vector components.

Symbol	Description	Dim.	Privileged
$\mathbf{v}$	Base linear velocity	3	✗
ω	Base angular velocity	3	✗
$\mathbf{v}_{xy}^{\text{cmd}}$	Commanded x and y linear velocities	2	✗
ω_z^{cmd}	Commanded z angular velocity	1	✗
$\mathbf{a}_{t-1}$	Previous actions	21	✗
$\mathbf{q}_j$	Joint positions	21	✗
$\dot{\mathbf{q}}_j$	Joint velocities	21	✗
$\mathbf{g}_{\text{proj}}$	Projected gravity vector	3	✗
$\ \mathbf{F}_{\text{ext}}\ $	External force magnitude	1	✓
$\angle \mathbf{F}_{\text{ext}}$	External force direction (xy plane)	1	✓

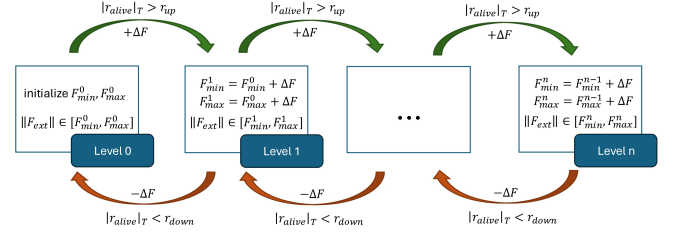
TABLE II: Reward components and expressions.

Component	Expression
Alive bonus	1
Linear velocity tracking	$\exp(-\ \mathbf{v}_{xy}^{\text{cmd}} - \mathbf{v}_{xy}\ ^2 / \sigma_{\text{track}}^2)$
Angular velocity tracking	$\exp(-(\omega_z^{\text{cmd}} - \omega_z)^2 / \sigma_{\text{track}}^2)$
Base orientation penalty	$\ \mathbf{g}_{\text{proj}, xy}\ ^2$
Base height penalty	$(z_{\text{base}} - z_{\text{target}})^2$
Action magnitude cost	$\ \mathbf{a}\ ^2$
Torque magnitude cost	$\ \boldsymbol{\tau}_j\ ^2$
Joint positions at limit cost	$\sum_{i=1}^n \mathbf{1}(q_{j,i}^{\text{scaled}} > 0.98)$
Both feet airborne cost	$\mathbf{1}(\bar{c}_L \wedge \bar{c}_R)$

The reward function does not explicitly account for push forces; instead, it promotes stable and natural locomotion. It rewards accurate velocity tracking, upright posture, pelvis stabilization, and low control effort, while penalizing undesirable behaviors such as both feet being simultaneously off the ground. A survival term encourages the robot to remain upright and active. The total reward is computed as a weighted sum of the terms listed in Table II.

B. Disturbance-Aware Curriculum Learning

Curriculum learning is employed to gradually increase task difficulty. Each parallel environment is monitored through the average alive reward, and—based on its recent performance—is promoted to a greater difficulty level, kept at the current one, or demoted. For environments exposed to external disturbances, the disturbance magnitude is progressively increased as the agent improves. Each environment is assigned a disturbance level $\ell \in \{0, 1, 2, \dots\}$, which determines the interval from which the disturbance magnitude $\|\mathbf{F}_{\text{ext}}\|$ is sampled. At the base level $\ell = 0$, the magnitude is drawn from $\|\mathbf{F}_{\text{ext}}\| \sim \mathcal{U}[F_{\min}^0, F_{\max}^0]$, where $F_{\min}^0$ and $F_{\max}^0$ define the minimum and maximum disturbance magnitudes at level zero. For higher levels $\ell > 0$, the sampling interval is increased by a fixed step ΔF $F_{\min}^\ell = F_{\min}^0 + \ell \Delta F$, $F_{\max}^\ell = F_{\max}^0 + \ell \Delta F$, yielding intervals of the form $\mathcal{U}[F_{\min}^\ell, F_{\max}^\ell]$ and enabling a smooth progression from easier to more challenging push-recovery tasks. A synthetic representation of the implemented curriculum-learning strategy is shown in Fig. 1. Promotion and demotion are driven by recent performance. Each environment maintains a moving average of the alive reward r_{alive} over the last N_{steps} time steps. If this average exceeds a promotion threshold r_{up} , the environment is promoted to level $\ell \rightarrow \ell + 1$. Conversely, if it falls below a demotion threshold r_{down} , the

Fig. 1: Flow diagram of the curriculum-learning procedure used during training, illustrating the progression from level 0 to level n as disturbance difficulty increases.

environment is demoted, with the constraint $\ell \geq 0$ to preserve a minimum difficulty level. This adaptive mechanism increases disturbance difficulty only when the policy can sustain stable locomotion, encouraging robustness without degrading gait quality. To improve robustness, external disturbances are applied every 5 s. Forces are randomized in magnitude (starting at 100 N and increasing via the curriculum), direction (uniform in the horizontal plane), duration (impulsive: 0.2 s, or sustained: 1–4 s), and point of application (uniformly sampled from 12 locations: pelvis, torso, shoulders, arms, forearms, thighs, and shins). To maintain training diversity and avoid catastrophic forgetting phenomena, 50% of the environments are disturbance-free, 25% experience impulsive disturbances, and the remaining 25% experience sustained disturbances. All randomizations in this work are performed by sampling from uniform distributions bounded by predefined minimum and maximum values. The static and dynamic friction coefficients, as well as the mass of the robot’s base link (pelvis), are independently sampled at the beginning of each episode. In particular, the nominal base-link mass $\bar{m}_{\text{base}} \in \mathbb{R}$ is perturbed within a symmetric range of amplitude Δm . In addition, velocity commands are resampled every T_{resample} seconds to provide varied locomotion targets. These randomized conditions are independent of the curriculum level, i.e., their sampling ranges remain constant throughout training. Episodes terminate either when the robot falls (incurring a penalty) or when the maximum episode duration is reached.

C. Learning-Based Control Architecture

The robot control architecture, shown in Fig. 2, adopts a two-level hierarchy to balance computational efficiency and actuator responsiveness. The high-level policy operates at frequency f_{policy} , while the actuators and low-level proportional-derivative (PD) controllers run at the higher frequency f_{PD} . The high-level controller is the neural-network policy π_θ introduced in Section IV-A, trained with the reward and curriculum-learning strategy described in Sections IV-A and IV-B. It receives the observation vector $\mathbf{o}_t$ reported in Table I and outputs joint position offsets $\Delta \mathbf{q}_j$ for the $n = 21$ actuated DoFs, from which desired joint positions are computed as $\mathbf{q}_j^{\text{des}} = \mathbf{q}_j^{\text{default}} + \Delta \mathbf{q}_j$. These references are tracked by the low-level PD controller, which generates the torque command $\boldsymbol{\tau}_j = \mathbf{K}_p (\mathbf{q}_j^{\text{des}} - \mathbf{q}_j) - \mathbf{K}_d \dot{\mathbf{q}}_j \in \mathbb{R}^n$, where $\mathbf{K}_p, \mathbf{K}_d \in \mathbb{R}^{n \times n}$ are positive definite stiffness and damping matrices. The result-

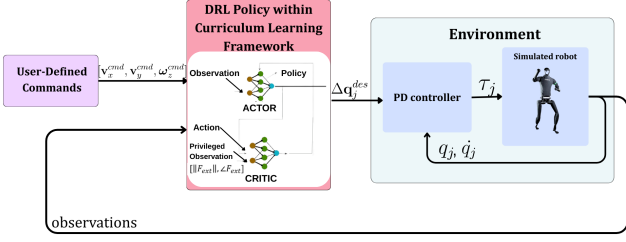


Fig. 2: Schematic representation of the proposed control architecture integrating the high-level policy with the low-level PD controller.

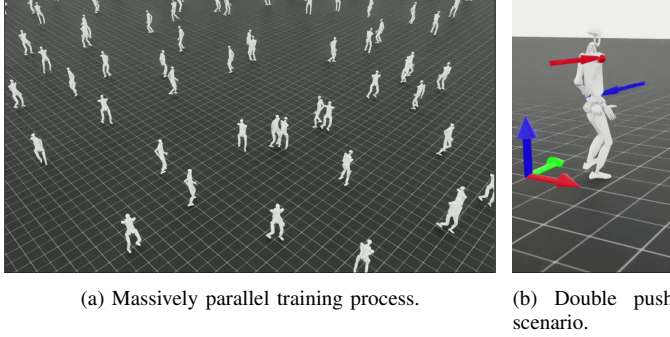


Fig. 3: Training setup for push-recovery learning. (a) Massively parallel DRL training with simulated Unitree H1 robots. (b) Illustration of the double-push disturbance scenario, where the red and blue arrows indicate two distinct forces applied simultaneously at different body locations.

ing torque is clipped to predefined limits to ensure hardware safety and actuator compliance.

V. SIMULATION RESULTS

Although the proposed architecture is not platform-specific, validation was carried out on the simulated Unitree H1 humanoid. Its human-scale dimensions make robust locomotion more challenging than on lighter platforms, providing a demanding benchmark for push recovery. An additional motivation was the availability of open-source software supporting both reinforcement learning and classical control methods.

Simulations were performed in NVIDIA Isaac Lab [22], which enables MPDRL through large-scale GPU-parallel simulation. Training showed rapid convergence: performance stabilized within the first 700 PPO iterations, while the curriculum progressively exposed the policy to stronger disturbances. Around iteration 1700, agents reached consistent performance, indicating completion of curriculum progression and final policy refinement.

A. Training Details

The policy was trained for 2000 PPO iterations using 4096 parallel environments and 32 steps per environment, yielding a batch size of 131,072. Both actor and critic were implemented as three-layer MLPs with hidden sizes [400, 200, 100] and ELU activations. Training was carried out with the `rsl-r1` framework [23] on an Intel i9-14900HX CPU and NVIDIA RTX 4090 GPU, requiring about 60 minutes per policy. The adopted PPO hyperparameters are reported in Table III.

TABLE III: PPO hyperparameters.

Training parameter	Value
Batch size	131072
Mini-batch size	26214
Number of epochs	5
Number of mini-batches	4
Clip range	0.2
Discount factor	0.99
GAE discount factor	0.95
Desired KL-divergence	0.008
Learning rate	adaptive

B. Single-Push Scenario

To evaluate the proposed policy, we compare it against a baseline inspired by the manufacturer’s native RL controller. In the baseline, only the 12 lower-limb DoFs are actuated, as commonly assumed in the literature [24], and no curriculum learning is used. The reward design remains comparable, including velocity tracking, posture maintenance, pelvis stabilization, control effort minimization, joint-limit penalties, airborne penalties, and an alive bonus.

1) *Velocity Tracking Performance under Perturbation:* This simulation compares the push-recovery (PR) policy and the baseline (BL) in tracking a constant forward velocity under external disturbances. A longitudinal force is applied to the torso at $t = 2.5$ s for 2 s: 100 N for the PR policy and 20 N for the baseline, reflecting their different robustness limits. The resulting linear and angular velocity tracking errors are shown in Fig. 4a and Fig. 4b. Before the disturbance, both controllers achieve similar tracking performance, with the baseline showing slightly lower steady-state linear-velocity error. After the perturbation, however, the PR policy recovers faster and restores accurate linear velocity tracking, whereas the baseline exhibits a larger and more persistent degradation. For angular velocity tracking, the two policies exhibit comparable behavior to the linear velocity case, as shown in Fig. 4b.

2) *Disturbance Rejection Analysis:* To assess the robustness of the PR policy, we performed randomized trials in which external forces were applied to different body parts. Disturbances were uniformly sampled in magnitude, horizontal-plane direction, and duration chosen from $\{0.1, 0.2, 0.5, 1, 2, 4\}$ s. Results are summarized in Fig. 5 and Fig. 6. Fig. 5 shows robustness as a function of disturbance magnitude and direction for each body part. To identify robust and non-robust regions, we performed a density-ratio analysis based on Kernel Density Estimation (KDE). Let $p_s(\|F_{\text{ext}}\|, \angle F_{\text{ext}})$ and $p_u(\|F_{\text{ext}}\|, \angle F_{\text{ext}})$ denote the estimated densities of successful and unsuccessful trials, respectively. The ratio $r(\|F_{\text{ext}}\|, \angle F_{\text{ext}}) = \frac{p_s(\|F_{\text{ext}}\|, \angle F_{\text{ext}})}{p_u(\|F_{\text{ext}}\|, \angle F_{\text{ext}})}$ provides a compact visualization of policy robustness over the disturbance space. The results show an improvement of the PR policy over the baseline, with success probability increasing by up to fivefold in several regions. The maximum rejectable force depends strongly on the point of application: the robot tolerates the largest forces on the shins and thighs (up to about 300 N), whereas upper-body disturbances are less

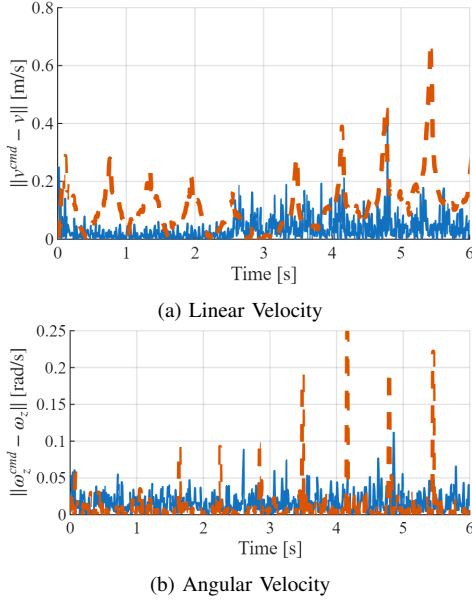


Fig. 4: Norm of the linear (a) and angular (b) velocity tracking errors with respect to the commanded velocity. The blue solid line denotes the push-recovery policy, and the orange dashed line the baseline policy. In both cases, a longitudinal force is applied at $t = 2.5$ s for 2 s.

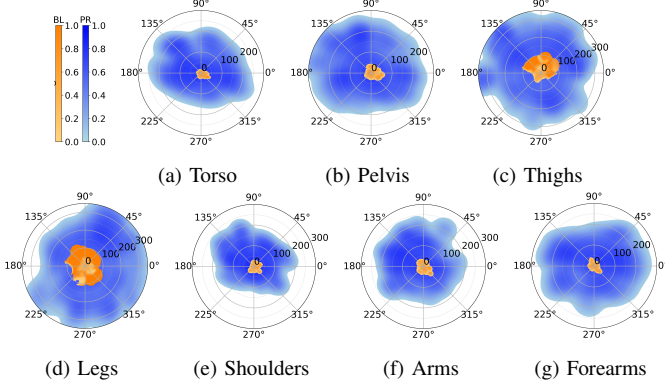


Fig. 5: Density ratio estimated via Gaussian KDE for different body parts, as a function of disturbance magnitude $\|F_{ext}\|$ and direction $\angle F_{ext}$, independently of the disturbance duration. For visualization purposes, only the values greater than 0.05 are represented.

tolerable because they generate larger destabilizing torques. No strong directional preference is observed for pushes applied to the pelvis, thighs, and shins, while for upper-body segments slightly greater robustness is observed along the sagittal direction. Fig. 6 reports the maximum successfully rejected force versus disturbance duration and body part. For most segments, this value remains nearly constant across durations, indicating robustness over a broad temporal range. A modest degradation is observed for the longest disturbances (4 s), suggesting that prolonged sustained pushes remain the most challenging case.

C. Dual-Push Scenario

A final and more demanding evaluation assessed the robustness of the push-recovery policy under simultaneous, spatially distributed perturbations. In this dual-disturbance scenario, two

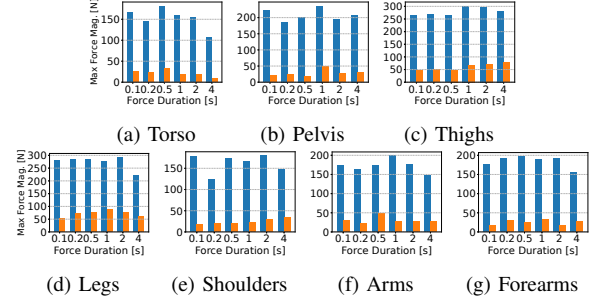


Fig. 6: Maximum rejected force as a function of duration of the force for different body parts, independently of the angle of application.

TABLE IV: Dual-disturbance test summary. Each row reports two simultaneous external forces characterized by magnitude, direction, and application point (abbreviated for space: r.=right, l.=left, sho.=shoulder, thi.=thigh, fore.=forearm). Both forces share the same duration. Dotted lines indicate omitted trials.

Force 1			Force 2			Dur. [s]	Succ.
Mag. [N]	Ang. [rad]	Body	Mag. [N]	Ang. [rad]	Body		
185.96	0.182	r. sho.	178.96	0.016	l. arm	0.5	✗
152.21	0.260	l. thi.	138.18	0.515	r. arm	4.0	✗
66.41	1.608	r. fore.	108.28	0.017	pelvis	2.0	✓
83.96	2.017	r. sho.	32.72	0.800	r. arm	0.2	✓
74.60	5.403	l. arm	9.06	5.911	r. sho.	4.0	✓
...							
8.08	3.375	pelvis	82.47	2.249	l. arm	2.0	✓
120.32	0.640	l. leg	156.36	0.117	torso	0.1	✗
181.83	1.426	l. arm	128.45	0.750	r. thi.	2.0	✗
6.42	0.920	l. leg	38.27	2.627	r. sho.	0.5	✓
117.09	3.823	l. arm	74.92	0.572	r. arm	2.0	✓
Total Survival							82.9%

external forces were applied concurrently to different body parts, yielding compound perturbations that are substantially more challenging than standard single-push tests. Each trial was defined by the magnitudes, directions, and application points of the two forces, together with a common duration. Force magnitudes were uniformly sampled up to 200 N, while durations were drawn from the same discrete set used in previous experiments. This test was performed only on the PR policy, since the baseline could not tolerate comparable force magnitudes even under single disturbances. Over 150 randomized trials, the PR policy maintained balance in about 82% of the cases. Representative disturbance tuples and outcomes are reported in Table IV. Failures mainly occurred when large forces acted simultaneously on critical lower-body segments or when opposing directions generated destabilizing moments beyond the controller’s recovery capability. Despite these challenging conditions, the overall success rate highlights strong robustness and the ability of the learned policy to generalize beyond its training distribution through coordinated whole-body responses.

D. Discussion

To counteract high-magnitude disturbances, especially when applied to the upper torso, the proposed policy sometimes

produced upper-body postures that appeared unnatural. In particular, unlike both the baseline policy and human-like motion, the learned behavior occasionally resulted in forearm configurations that were not bent at approximately 90° with respect to the upper arms, or in nearly fully extended arm postures, which are more typical of human motion. These configurations were not explicitly penalized, since doing so caused a noticeable degradation in push-recovery performance in simulation. Moreover, from a practical standpoint, they did not compromise locomotion stability or induce self-collisions with other body parts, as joints limit positions were explicitly incorporated into the reward design.

VI. CONCLUSIONS AND FUTURE WORK

This paper presented a push-recovery framework for humanoid locomotion based on massively parallel deep reinforcement learning and a structured curriculum-learning strategy. The main contribution is not a new RL algorithm, but a procedural training framework that progressively increases locomotion difficulty through disturbances of increasing magnitude, duration, and application point, enabling robust recovery while preserving stable walking. The approach was validated in simulation on the Unitree H1 in NVIDIA Isaac Lab. Results showed rapid convergence and clear improvements over a baseline locomotion policy, with the proposed controller tolerating significantly stronger perturbations and achieving perfect recovery in challenging randomized disturbance tests. Future work will focus on transfer to the real robot, extension to more complex interaction scenarios, and integration of biomechanics-inspired insights from passive dynamic walking [25] and physically grounded control principles [26], [27] to further improve robustness, safety, energy efficiency, and physical interoperability preserving strong push-recovery.

REFERENCES

- [1] Y. Tong, H. Liu, and Z. Zhang, "Advancements in humanoid robots: A comprehensive review and future prospects," *IEEE/CAA Journal of Automatica Sinica*, vol. 11, no. 2, pp. 301–328, 2024.
- [2] J. Fox and A. Gambino, "Relationship development with humanoid social robots: Applying interpersonal theories to human–robot interaction," *Cyberpsychology, Behavior, and Social Networking*, vol. 24, no. 5, pp. 294–299, 2021.
- [3] A. Shamsah, J. Jiang, Z. Yoon, S. Coogan, and Y. Zhao, "Terrain-aware model predictive control of heterogeneous bipedal and aerial robot coordination for search and rescue tasks," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 12 352–12 358.
- [4] B. Stephens, "Humanoid push recovery," in *2007 7th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2007, pp. 589–595.
- [5] Z. O. Orhan, M. Shafiee, V. Juillard, J. C. Oliveira, A. Ijspeert, and M. Bouri, "Exorecovery: Push recovery with a lower-limb exoskeleton based on stepping strategy," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 3248–3255.
- [6] J. Pratt, J. Carff, S. Drakunov, and A. Goswami, "Capture point: A step toward humanoid push recovery," in *2006 6th IEEE-RAS international conference on humanoid robots*. IEEE, 2006, pp. 200–207.
- [7] S. Caron, A. Kheddar, and O. Tempier, "Stair climbing stabilization of the hrp-4 humanoid robot using whole-body admittance control," in *2019 International conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 277–283.
- [8] D. Kim, S. J. Jorgensen, J. Lee, J. Ahn, J. Luo, and L. Sentis, "Dynamic locomotion for passive-ankle biped robots and humanoids using whole-body locomotion control," *The International Journal of Robotics Research*, vol. 39, no. 8, pp. 936–956, 2020.
- [9] B. J. Stephens and C. G. Atkeson, "Push recovery by stepping for humanoid robots with force controlled joints," in *2010 10th IEEE-RAS International conference on humanoid robots*. IEEE, 2010, pp. 52–59.
- [10] J. Li, Z. Le, J. Ma, and Q. Nguyen, "Adapting gait frequency for posture-regulating humanoid push-recovery via hierarchical model predictive control," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 12 330–12 336.
- [11] Y. Liu, J. Shen, J. Zhang, X. Zhang, T. Zhu, and D. Hong, "Design and control of a miniature bipedal robot with proprioceptive actuation for dynamic behaviors," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8547–8553.
- [12] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," in *Conference on robot learning*. PMLR, 2022, pp. 91–100.
- [13] K. Savevska and A. Ude, "Learning push recovery strategies for humanoid robots using deep reinforcement learning," in *2025 IEEE International Conference on Advanced Robotics (ICAR)*. IEEE, 2025, pp. 8–15.
- [14] D. Ferigo, R. Camoriano, P. M. Viceconte, D. Calandriello, S. Traversaro, L. Rosasco, and D. Pucci, "On the emergence of whole-body strategies from humanoid robot push-recovery learning," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8561–8568, 2021.
- [15] I. Radosavovic, T. Xiao, B. Zhang, T. Darrell, J. Malik, and K. Sreenath, "Real-world humanoid locomotion with reinforcement learning," *Science Robotics*, vol. 9, no. 89, p. eadi9579, 2024.
- [16] A. Duburcq, F. Schramm, G. Bo  ris, N. Bredeche, and Y. Cheval  yre, "Reactive stepping for humanoid robots using reinforcement learning: Application to standing push recovery on the exoskeleton atalante," in *2022 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2022, pp. 9302–9309.
- [17] G. B. Margolis, G. Yang, K. Paigwar, T. Chen, and P. Agrawal, "Rapid locomotion via reinforcement learning," *The International Journal of Robotics Research*, vol. 43, no. 4, pp. 572–587, 2024.
- [18] S. Li, G. Wang, Y. Pang, P. Bai, S. Hu, Z. Liu, L. Wang, and J. Li, "Learning agility and adaptive legged locomotion via curricular hindsight reinforcement learning," *Scientific Reports*, vol. 14, no. 1, p. 28089, 2024.
- [19] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
- [20] L. Pinto, M. Andrychowicz, P. Welinder, W. Zaremba, and P. Abbeel, "Asymmetric actor critic for image-based robot learning," *arXiv preprint arXiv:1710.06542*, 2017.
- [21] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.
- [22] M. Mittal, P. Roth, J. Tigue, A. Richard, O. Zhang, P. Du, A. Serrano-Munoz, X. Yao, R. Zurbr  gg, N. Rudin *et al.*, "Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning," *arXiv preprint arXiv:2511.04831*, 2025.
- [23] C. Schwarke, M. Mittal, N. Rudin, D. Hoeller, and M. Hutter, "Rsl-rl: A learning library for robotics research," *arXiv preprint arXiv:2509.10771*, 2025.
- [24] C. Yang, K. Yuan, W. Merkt, T. Komura, S. Vijayakumar, and Z. Li, "Learning whole-body motor skills for humanoids," in *2018 IEEE-RAS 18th international conference on humanoid robots (humanoids)*. IEEE, 2018, pp. 270–276.
- [25] Z. Xie, X. Gong, P. Arpent, R. Zou, S. Zhou, W. Zhang, Z. Zhang, and Y. Wang, "A novel model analysis of the effects of upper body dynamics on passive dynamic biped walkers," *Multibody System Dynamics*, pp. 1–23, 2025.
- [26] P. Arpent, F. Ruggiero, and V. Lippiello, "Interconnection and damping assignment passivity-based control for gait generation in underactuated compass-like robots," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 9802–9808.
- [27] P. Arpent, A. Donaire, F. Ruggiero, and V. Lippiello, "Energy pumping-and-damping for gait robustification of underactuated planar biped robots within the hybrid zero dynamics framework," in *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2021, pp. 415–421.