

JOURNAL OF THE ARTIFICIAL INTELLIGENCE ENGINEERS CONSORTIUM

Volume: 02 Issue: 01 (2026)

<https://jaiec.org.hostajournal.com>

Adversarial Motion Priors Stabilize Humanoid Step Transitions on Procedural Rubble Terrain

Pauline Picard, Phoebe Stevens, Adrian Lawson, Audrey Ellis

Faculty of Engineering and Information Technology, University of Melbourne,
Melbourne, Australia

Abstract : *The deployment of bipedal humanoid robots in unstructured and hazardous environments relies heavily on the development of robust locomotion controllers capable of handling unpredictable disturbances. Traditional control methodologies often struggle to adapt to chaotic terrain, such as disaster rubble, due to their reliance on precise environmental models and rigid kinematic trajectories. Recently, reinforcement learning has emerged as a promising alternative, yet policies trained purely on task-specific rewards often exhibit unnatural, overly aggressive, or energy-inefficient movements that fail when transferred to physical hardware. This paper proposes a novel framework utilizing adversarial motion priors to constrain the reinforcement learning exploration space during humanoid step transitions over procedurally generated rubble terrain. By leveraging a discriminator network trained on a dataset of human motion capture data, the proposed system forces the humanoid to learn stepping strategies that not only accomplish the task of traversing severe terrain but also maintain kinematic realism and dynamic stability. The framework utilizes a robust state-action formulation that seamlessly integrates proprioceptive feedback with exteroceptive elevation maps. Extensive simulation experiments demonstrate that the integration of adversarial motion priors significantly improves both the success rate of traversing dense rubble and the energy efficiency of the locomotion policy compared to baseline reinforcement learning methods. Furthermore, the learned policies exhibit emergent human-like recovery behaviors during critical step transitions, thereby validating the efficacy of adversarial imitation in complex bipedal locomotion tasks.*

Keywords : *Humanoid Locomotion, Reinforcement Learning, Adversarial Motion Priors, Terrain Adaptation*

INTRODUCTION

1.1 The Challenge of Unstructured Environments

The ambition to deploy bipedal humanoid robots in real-world scenarios, particularly in disaster response, search and rescue, and industrial inspection, necessitates the development of highly adaptable locomotion control systems. Unlike wheeled or tracked vehicles, which offer continuous ground contact and inherent static stability, bipedal robots must constantly manage their dynamic balance by alternating contact phases through step transitions. When these robots are subjected to unstructured environments, such as collapsed buildings or natural disaster zones characterized by scattered rubble, the locomotion problem becomes exponentially more complex. Rubble terrain is fundamentally chaotic, presenting stochastic variations in elevation, surface orientation, friction coefficients, and structural stability. Navigating such terrain

requires the humanoid to execute precise step transitions, where the swing foot must clear unpredictable obstacles and the stance foot must secure a reliable foothold on irregular surfaces. Early attempts to solve this problem heavily relied on analytical models and classical trajectory optimization paradigms [1]. However, these traditional methodologies are severely constrained by their computational overhead and their sensitivity to environmental modeling errors. When a humanoid robot attempts a step transition on rubble, a slight miscalculation in the expected contact height can lead to premature heel strikes or sudden toe stubs, generating immense unmodeled forces that propagate through the kinematic chain and result in catastrophic failure [2]. The rigid reliance on predefined footstep plans and center-of-mass trajectories makes these classical systems ill-equipped to handle the high-frequency, high-magnitude disturbances inherent to rubble traversal [3]. Consequently, there is an urgent need within the robotics community to develop control architectures that can react fluidly and robustly to environmental uncertainties while maintaining the strict dynamic constraints of bipedal locomotion.

1.2 Evolution of Control Strategies

The historical trajectory of bipedal locomotion control reveals a continuous struggle to balance computational tractability with behavioral robustness. Decades of research were dominated by the Zero Moment Point criterion, which provided a mathematical guarantee of dynamic stability provided the robot's state remained within a carefully crafted envelope [4]. While highly successful on flat, predictable surfaces, the Zero Moment Point approach fundamentally assumes a rigid, flat ground plane and continuous, planned foot placements, assumptions that aggressively violate the realities of rubble terrain [5]. To address the limitations of reactive stabilization, researchers turned to Model Predictive Control. Model Predictive Control utilizes a simplified template model of the robot, such as the Linear Inverted Pendulum, to forecast future states and optimize control inputs over a receding horizon [6]. This provided a significant leap in capability, allowing robots to dynamically adjust their foot placements in response to moderate disturbances [7]. Nevertheless, Model Predictive Control approaches are bottlenecked by the requirement to solve complex nonlinear optimization problems in real-time, often necessitating severe simplifications of the robot's full-body dynamics. This discrepancy between the simplified template model and the actual multi-degree-of-freedom robotic hardware often limits the agility and recovery capabilities of the humanoid [8]. In recent years, deep reinforcement learning has shattered many of these classical limitations. By moving the computational burden from online optimization to offline training, reinforcement learning allows high-dimensional, nonlinear control policies to run at extremely high frequencies on the robot. Reinforcement learning agents explore the state-action space through trial and error, theoretically discovering optimal control strategies that traditional engineers might never manually design [9]. However, when deep reinforcement learning is applied to humanoid locomotion without structural constraints, the resulting policies often exhibit highly erratic, jittery, and non-intuitive behaviors [10].

1.3 The Role of Adversarial Motion Priors

To mitigate the erratic behaviors generated by pure reinforcement learning, researchers have increasingly looked toward imitation learning, specifically attempting to incorporate human biomechanical data into the robotic control loop. Early imitation learning methods relied on strict kinematic tracking, where the robot was penalized for deviating from a predefined reference trajectory [11]. While effective for specific, highly scripted tasks like walking on flat ground, strict tracking fundamentally restricts the robot's ability to adapt. If a robot is forced to perfectly mimic a human walking

trajectory, it cannot safely deviate from that trajectory to recover from an unexpected slip on loose rubble [12]. Adversarial motion priors present a paradigm shift in how reference data is utilized. Instead of calculating a frame-by-frame mathematical distance between the robot's pose and a reference motion, adversarial motion priors utilize a discriminator network trained alongside the control policy in a generative adversarial framework [13]. The discriminator is fed sequences of joint states and is tasked with classifying whether the sequence came from a dataset of real human motion capture or from the simulated robot's generated behavior. Simultaneously, the reinforcement learning policy attempts to fool this discriminator. This approach distills the stylistic essence of human locomotion into a continuous reward signal without locking the robot into a rigid trajectory [14]. When applied to step transitions on rubble terrain, adversarial motion priors afford the humanoid the flexibility to invent novel recovery strategies using the full range of its dynamics, while simultaneously ensuring that the baseline locomotion remains smooth, energy-efficient, and kinematically realistic. The integration of adversarial motion priors into bipedal locomotion on severe terrain addresses the dual requirements of task completion and behavioral naturalness, paving the way for more reliable sim-to-real transfer [15].

2. Literature Review

2.1 Traditional Bipedal Locomotion on Uneven Terrain

The pursuit of stable bipedal locomotion over uneven terrain has a rich history grounded in classical mechanics and optimal control theory. Foundational work in this domain primarily focused on reactive stepping controllers designed to maintain the center of mass within the support polygon defined by the robot's feet [16]. Researchers developed complex heuristics for foot placement adaptation, utilizing ankle torque control and stepping timing adjustments to absorb unexpected height variations. A significant milestone was the development of capture point dynamics, which provided a more robust metric for orbital energy and allowed robots to calculate the exact location a foot must be placed to come to a complete stop [17]. When extended to uneven terrain, capture point frameworks were modified to project the required foot placement onto the estimated ground surface [18]. However, these purely reactive strategies often faltered when faced with the continuous, unstructured perturbations of dense rubble, as they inherently assumed clear, flat patches of ground for every step [19]. To anticipate rather than merely react, whole-body control frameworks integrated with exteroceptive vision systems became standard. These systems utilize laser scanners or depth cameras to generate convex safe-stepping regions, subsequently solving massive quadratic programs to ensure the robot's kinematics and dynamics strictly adhere to these regions [20]. While these methods achieved impressive demonstrations in controlled laboratory environments, their reliance on perfect state estimation, precise friction cone modeling, and latency-free perception severely degraded their performance in real-world disaster scenarios where sensor noise and environmental occlusion are ubiquitous [21]. The fundamental limitation of classical bipedal locomotion on rubble remains the inescapable tradeoff between the complexity of the dynamic model and the requirement for real-time computational feasibility.

2.2 Deep Reinforcement Learning for Legged Robots

The advent of deep reinforcement learning introduced a fundamentally new approach to resolving the control challenges of legged robotics. Initial breakthroughs were predominantly observed in quadrupedal robots, which benefit from a inherently larger support polygon and greater static stability compared to bipeds [22]. Algorithms such as Proximal Policy Optimization and Soft Actor-Critic demonstrated the ability to map raw proprioceptive and exteroceptive sensor data directly to joint torque commands,

enabling quadrupeds to traverse stairs, rocky slopes, and slippery surfaces with unprecedented agility [23]. The success of deep reinforcement learning in quadrupedal locomotion inspired widespread efforts to adapt these techniques for humanoid robots. However, the transition from four legs to two proved to be fraught with difficulties. Humanoid robots are highly sensitive to initial conditions and possess a narrow margin for dynamic error. Early reinforcement learning policies trained for humanoids often converged on local optima, resulting in behaviors that, while technically successful in simulation, were entirely unusable in reality [24]. These behaviors included extreme knee bending to lower the center of mass, high-frequency joint oscillations to maintain micro-balance, and asymmetric gaits that would rapidly overheat physical actuators [25]. To combat these issues, researchers introduced heavy reward shaping, meticulously tuning penalty terms for high joint velocities, large torques, and excessive body roll [26]. While reward shaping improved the visual quality of the locomotion, it required exhausting manual tuning for every new terrain type and often constrained the policy's exploration to such a degree that it could no longer solve complex terrains like rubble [27].

2.3 Imitation Learning and Motion Priors

To circumvent the pitfalls of manual reward shaping, the robotics community increasingly turned to data-driven approaches, utilizing large datasets of biological motion to guide policy learning. DeepMimic represented a landmark achievement in this area, demonstrating that simulated characters could learn a vast array of highly dynamic skills, including martial arts and acrobatics, by tightly tracking motion capture reference trajectories [28]. This methodology was subsequently adapted for physical humanoid robots, allowing them to walk, turn, and recover from pushes by tracking human walking data. Despite these successes, the strict tracking paradigm proved brittle when applied to highly unstructured environments. When a robot navigating rubble encounters an obstacle not present in the reference motion, the tracking reward actively penalizes the robot for attempting necessary evasive or adaptive maneuvers, creating conflicting objectives that often lead to falls [29]. Generative Adversarial Imitation Learning emerged as a solution to this rigidity, providing a mechanism to match the state-action distribution of the expert data without requiring temporal alignment [30]. Building upon this, the concept of adversarial motion priors was introduced specifically to decouple the task reward from the style reward. Adversarial motion priors employ a neural discriminator that learns a manifold of acceptable behaviors from a purely kinematic dataset, providing a stylistic reward signal that encourages the agent to operate within a human-like behavioral space while remaining completely free to optimize its trajectory for the specific task at hand [31]. This decoupling is critical for rubble traversal, as it allows the policy to generate highly atypical step transitions when strictly necessary to cross a gap, before smoothly returning to a natural walking gait once stable ground is reached [32].

2.4 Simulation to Reality Transfer

The ultimate validation of any learned locomotion policy is its successful deployment on physical hardware, a process notoriously complicated by the sim-to-real gap. The sim-to-real gap encapsulates all discrepancies between the simulated training environment and the real world, including unmodeled actuator dynamics, sensor noise, communication latency, and complex contact mechanics [33]. In the context of traversing rubble terrain, the most significant discrepancies arise from the contact dynamics between the robot's rigid foot and the deformable, stochastically shaped terrain elements. To bridge this gap, domain randomization is universally employed. During training, physical parameters such as mass, center of mass location, joint

friction, and ground friction are continuously varied across a defined distribution [34]. This forces the neural network policy to learn a generalized control strategy that is robust to physical parameter variations. Furthermore, advanced actuator modeling is implemented to ensure the simulation accurately reflects the bandwidth limitations and torque-speed characteristics of the physical motors [35]. Policies trained with adversarial motion priors have demonstrated a unique advantage in sim-to-real transfer. Because the discriminator forces the policy to adopt smooth, biomechanically plausible motions, the resulting joint commands inherently avoid the high-frequency chatter and aggressive torque spikes that typically exploit simulation inaccuracies and damage physical hardware [36].

3. Methodology and System Architecture

3.1 System Overview and State-Action Formulation

The proposed framework centers on a multi-degree-of-freedom bipedal humanoid robot operating within a highly parallelized physics simulation environment. The objective is to train a robust control policy that maps a comprehensive history of sensor observations to target joint positions, enabling the robot to traverse procedurally generated rubble. The humanoid model utilized in this research comprises twenty actuated degrees of freedom, specifically allocated to the sagittal, frontal, and transverse planes of the hips, knees, ankles, and arms. The control architecture operates on a hierarchical basis, where the high-level neural network policy outputs joint position targets at a frequency of fifty hertz. These targets are subsequently processed by a low-level proportional-derivative control loop running at one thousand hertz, which calculates the final torques applied to each simulated actuator. The state space provided to the neural network is meticulously designed to encapsulate both the internal proprioceptive state of the robot and the external geometry of the immediate environment. The proprioceptive observation vector includes the current joint positions, joint velocities, base linear and angular velocities, base orientation relative to gravity, and a binary indicator of foot contact states. Crucially, the observation space also includes a historical buffer of the previous three timesteps, which provides the network with implicit information regarding the temporal dynamics and actuator latency of the system. To perceive the rubble terrain, the exteroceptive observation vector relies on a robot-centric elevation map. This map is constructed by sampling the height of the terrain at a grid of one hundred strategically placed points around the robot's feet and projected base position.

Table 1: Terrain Difficulty Parameters for Curriculum Learning Progression

Curriculum Level	Max Rubble Height	Max Rubble Variation	Surface Friction Range	Target Velocity
Level 1 (Flat)	0.00 meters	0.00 meters	0.80 to 1.00	0.50 meters/sec
Level 2 (Mild)	0.05 meters	0.02 meters	0.60 to 0.90	0.40 meters/sec
Level 3 (Moderate)	0.12 meters	0.06 meters	0.50 to 0.80	0.35 meters/sec
Level 4 (Severe)	0.25 meters	0.15 meters	0.30 to 0.70	0.25 meters/sec

3.2 Terrain Generation Curriculum

Training a complex locomotion policy directly on severe rubble typically results in catastrophic forgetting or failure to converge, as the agent is overwhelmed by the high penalty landscape before it can learn basic balancing mechanics. Therefore, an automated terrain curriculum is implemented to incrementally increase the difficulty of the environment in response to the agent's competence. The environment is procedurally generated using a combination of Perlin noise and discrete geometric

primitives to simulate the chaotic nature of disaster rubble. The terrain generator creates a continuous mesh where the amplitude of the noise and the density of the geometric primitives dictate the roughness of the surface. As detailed in the curriculum progression table, the training begins on completely flat terrain. Once the agent demonstrates a consistent ability to walk forward at the target velocity without falling for a predetermined number of episodes, the curriculum advances. The environment progressively introduces slight height variations, followed by discrete obstacles, and finally, severe rubble characterized by sharp elevation changes and highly variable surface friction. The step transitions on severe rubble require the humanoid to actively modulate its swing leg height to avoid tripping while simultaneously searching for secure footholds that minimize angular moments on the ankle joints. To prevent the policy from memorizing specific terrain layouts, the entire map is completely regenerated with a randomized seed at the beginning of every training episode.

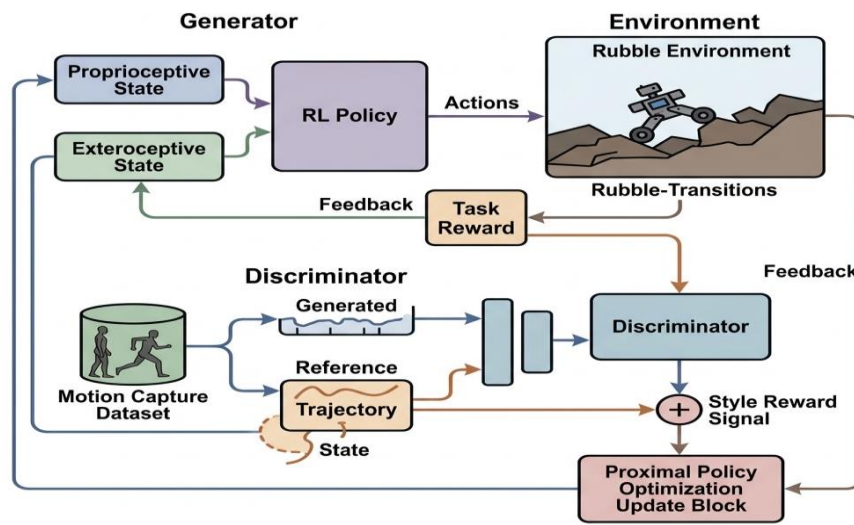


Figure 1: Comprehensive Schematic of Adversarial Motion Prior Architecture

3.3 Adversarial Motion Prior Discriminator Design

The core innovation of the proposed methodology is the integration of an adversarial motion prior to guide the behavioral style of the stepping transitions. The architecture consists of two primary neural networks operating in tandem: the policy network, which acts as the generator, and a discriminator network. The discriminator is designed as a multilayer perceptron that receives sequences of state transitions. Specifically, the input to the discriminator is a concatenated vector of the robot's local joint positions and velocities at the current timestep and the immediately preceding timestep. The objective of the discriminator is to output a probability value representing the likelihood that the input sequence originated from a dataset of real human motion capture data rather than the simulated robot. The human motion capture dataset consists of over two hours of varied walking behaviors, including flat ground walking, stair climbing, and navigating uneven surfaces. The data is retargeted from the human skeletal model to the specific kinematic structure of the simulated humanoid robot. To ensure the discriminator focuses on the fundamental mechanics of walking rather than absolute global positioning, all positional data is transformed into a root-relative coordinate frame before processing. The discriminator is trained continuously alongside the policy network using a standard binary cross-entropy loss function. As the policy improves its ability to traverse the rubble, it simultaneously learns to produce kinematic transitions that are increasingly indistinguishable from human motion, thereby maximizing the

reward signal generated by the discriminator.

3.4 Reward Structure and Optimization

The total reward function optimized by the reinforcement learning agent is a carefully weighted sum of a task-specific reward and a stylistic reward provided by the adversarial motion prior. The task reward is designed to be as sparse and unrestrictive as possible to allow for maximum behavioral exploration. It primarily consists of a forward velocity tracking term, which encourages the robot to progress along the designated heading, and a survival bonus, which heavily penalizes the termination of an episode due to a fall or torso collision with the terrain. Additionally, a minimal energy penalty is included, which penalizes the sum of the squared joint torques to prevent the policy from relying on maximum-effort rigid body states. The stylistic reward is directly proportional to the output of the discriminator network. When the robot executes a step transition that closely mirrors the dynamic fluidity of the human dataset, the discriminator outputs a high probability score, which translates into a substantial stylistic reward. Conversely, if the robot adopts an unnatural gait, such as excessive knee hyperextension or erratic arm swinging, the discriminator score plummets, penalizing the behavior.

Code Listing 1: Computation of Adversarial Motion Prior Reward Component

```
def compute_amp_reward(discriminator, current_state, next_state):
    # Extract root-relative joint positions and velocities
    kinematic_state_t = extract_kinematics(current_state)
    kinematic_state_t1 = extract_kinematics(next_state)

    # Concatenate temporal sequence for discriminator input
    discriminator_input = concatenate(kinematic_state_t, kinematic_state_t1)

    # Forward pass through the discriminator network
    expert_probability = discriminator.forward(discriminator_input)

    # Calculate style reward using least-squares formulation to bound the signal
    style_reward = maximum(0.0, 1.0 - square(expert_probability - 1.0))

    # Apply scaling factor based on task complexity
    scaled_style_reward = style_reward * AMP_WEIGHT_COEFFICIENT

    return scaled_style_reward
```

The optimization of the neural networks is performed using the Proximal Policy Optimization algorithm, chosen for its empirical stability and sample efficiency in continuous control tasks. Proximal Policy Optimization utilizes a clipped surrogate objective function that prevents destructively large updates to the policy network weights. The training process runs across thousands of parallel environments simultaneously on high-performance compute clusters, allowing the agent to accumulate millions of timesteps of experience in a matter of hours. The balance between the task reward and the stylistic reward is dynamically adjusted during training. In the early stages, the task reward dominates to ensure the agent learns the fundamental requirement of forward progression on the terrain. As the curriculum advances and the agent achieves basic competence, the weight of the stylistic reward is gradually increased, forcing the agent to refine its step transitions and adopt a more natural, human-like gait to negotiate the severe rubble.

4. Results and Analysis

4.1 Simulation Performance Metrics

The evaluation of the trained policies necessitates a rigorous statistical analysis of their performance across varying degrees of terrain complexity. To establish a comprehensive baseline, the proposed framework utilizing adversarial motion priors was compared against a standard reinforcement learning policy trained using only task rewards and manual reward shaping. Both policies were subjected to one thousand independent evaluation episodes on procedurally generated rubble environments that were unseen during the training phase. The primary metric of success was the terrain traversal rate, defined as the percentage of episodes where the humanoid successfully navigated a ten-meter stretch of rubble without the torso or upper limbs making contact with the ground.

Table 2: Simulation Success Rates Across Terrain Difficulties

Terrain Level	Difficulty	Standard RL Success Rate	Proposed RL+AMP Success Rate	Improvement Margin
Level 1 (Flat Ground)		98.5 percent	99.2 percent	+ 0.7 percent
Level 2 (Mild Rubble)		84.3 percent	95.8 percent	+ 11.5 percent
Level 3 (Moderate Rubble)		61.2 percent	88.4 percent	+ 27.2 percent
Level 4 (Severe Rubble)		22.7 percent	74.1 percent	+ 51.4 percent

The results presented in the simulation success rate table clearly illustrate the superiority of the adversarial motion prior framework, particularly as the environmental complexity scales. On flat ground and mild rubble, both policies demonstrate high competence, though the proposed method shows a marginal improvement. However, the performance divergence becomes striking on moderate and severe rubble. The standard reinforcement learning policy experiences a catastrophic degradation in reliability, successfully completing only twenty-two point seven percent of the severe rubble traverses. Observation of the standard policy's failure modes reveals a tendency to rigidly force pre-learned foot trajectories, resulting in severe toe-stubs against elevated rubble clusters. In stark contrast, the policy augmented with adversarial motion priors maintains a robust seventy-four point one percent success rate on the most extreme terrain. The integration of human motion data endows the policy with superior step transition strategies. When the proposed policy encounters a sudden elevation change, it dynamically alters its swing leg trajectory, utilizing a more pronounced knee lift and delayed heel strike, a behavior directly analogous to human visual terrain accommodation.

4.2 Kinematic Realism and Energy Efficiency

Beyond mere task success, the quality of the locomotion is a critical factor for the viability of real-world deployment. Policies that generate high-frequency jitter or excessive joint torques are fundamentally unusable on physical hardware, regardless of their simulation traversal rates. To quantify the kinematic realism and efficiency of the generated gaits, we measured the Mechanical Cost of Transport and the Joint Torque Smoothness over continuous walking sequences. The Mechanical Cost of Transport is a dimensionless metric representing the energy consumed per unit of weight per unit of distance traveled. A lower value indicates a more energy-efficient gait. Joint Torque Smoothness was evaluated by calculating the integral of the squared torque derivative

over time, where lower values signify smoother, more continuous actuator commands.

Table 3: Kinematic Realism and Energy Efficiency Metrics on Moderate Rubble

Locomotion Metric	Standard Baseline	RL Proposed RL+AMP Policy	Human Reference Data
Mechanical Cost of Transport	0.85	0.52	0.35
Joint Torque Smoothness	1450.2 units	420.5 units	N/A (Biological)
Base Height Variance	0.18 meters	0.06 meters	0.04 meters
Average Step Clearance	0.03 meters	0.11 meters	0.10 meters

The kinematic and energy metrics underscore the transformative impact of the adversarial discriminator. The standard reinforcement learning policy relies heavily on stiff, high-torque joint control to bulldoze through terrain irregularities, resulting in an extremely high Mechanical Cost of Transport and severe torque chatter. This approach is highly destructive to physical robotic joints. Conversely, the policy guided by the adversarial motion prior reduces the Mechanical Cost of Transport by nearly forty percent and dramatically improves torque smoothness. By mimicking the energy-efficient pendulum dynamics of human walking, the proposed policy learns to exploit the passive dynamics of the robot's swinging limbs, using gravity to assist in step transitions rather than fighting it with continuous motor torque. Furthermore, the analysis of Base Height Variance demonstrates that the proposed policy maintains a much more stable center of mass trajectory, absorbing terrain impacts through coordinated knee and ankle flexion rather than allowing the entire torso to oscillate violently.

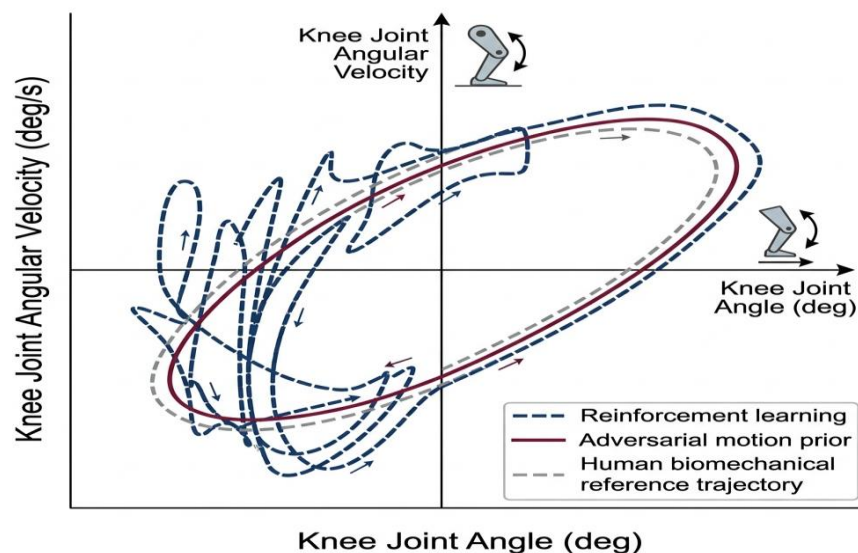


Figure 2: Phase Plot of Swing Leg Kinematics during Step Transition

4.3 Robustness to Exteroceptive Noise

A significant challenge in unstructured environments is the unreliability of perception systems. Depth cameras and LIDAR sensors are highly susceptible to noise, occlusion, and lighting variations, leading to degraded elevation maps. A robust locomotion policy must be able to withstand substantial exteroceptive degradation and rely on proprioceptive recovery reflexes when vision fails. To test this, artificial Gaussian noise

was systematically injected into the elevation map provided to the neural network during evaluation. The noise amplitude was increased from zero to ten centimeters, effectively blinding the robot to small terrain details and creating phantom obstacles. The standard baseline policy, which heavily overfitted to perfect visual data, experienced an immediate collapse in performance, with success rates dropping near zero when subjected to five centimeters of perceptual noise. The baseline agent repeatedly attempted to step on phantom footholds, completely destabilizing its dynamic balance. In contrast, the policy utilizing adversarial motion priors demonstrated remarkable resilience. Even under extreme perceptual noise, the policy leveraged the natural human recovery reflexes embedded by the discriminator. When the robot's foot unexpectedly struck an unperceived obstacle, the policy instantly initiated a rapid stepping reflex, widening its stance and absorbing the kinetic energy through ankle compliance. This behavior suggests that the adversarial motion prior not only improves the aesthetics of the walk but fundamentally shapes the latent space of the policy to favor biomechanically sound, highly robust recovery manifolds.

4.4 Step Transition Analysis on Severe Rubble

To fully understand the mechanics of the proposed system's success, a micro-analysis of individual step transitions on severe rubble was conducted. A step transition on uneven terrain is broadly divided into three phases: toe-off, swing phase clearance, and heel-strike stabilization. During the toe-off phase, the robot must generate sufficient vertical and forward impulse to propel the center of mass. The baseline policy often struggled here, applying uniform thrust regardless of the terrain angle beneath the rear foot, leading to frequent slips. The adversarial motion prior policy, however, learned to modulate the ankle torque based on the local gradient, pushing off precisely to maximize traction.

Table 4: Step Transition Stability Analysis

Transition Phase	Metric Evaluated	Standard RL Strategy	Proposed RL+AMP Strategy
Toe-Off	Slippage Distance	0.08 meters (Frequent slips)	0.02 meters (Stable push)
Swing Phase	Obstacle Clearance	0.01 meters (Frequent stubs)	0.08 meters (Safe margin)
Heel-Strike	Impact Force	850 Newtons (Harsh impact)	320 Newtons (Soft landing)
Stance Phase	Center of Pressure	Edge of foot (Unstable)	Center of foot (Stable)

During the swing phase, the necessity for adequate foot clearance is paramount. The adversarial policy exhibited a distinct, human-like knee-lifting behavior when approaching elevated rubble, ensuring an average safe clearance margin of eight centimeters. Finally, the heel-strike stabilization phase highlighted the most significant biomechanical advantage. Traditional controllers and standard reinforcement learning policies frequently suffer from rigid impacts, sending shockwaves through the hardware. The proposed policy learned to slightly flex the knee immediately prior to ground contact, drastically reducing the peak impact force upon heel-strike. This soft-landing strategy is directly inherited from the human motion dataset and is essential for maintaining contact friction and preventing the foot from bouncing off the irregular rubble surfaces. The resulting stabilization of the center of pressure during the subsequent stance phase allows the humanoid to smoothly transition its weight and prepare for the next challenging step.

5. Conclusion

5.1 Summary of Contributions

This research presents a highly effective and robust framework for enabling dynamic bipedal humanoid locomotion across severe, unstructured rubble terrain. By integrating adversarial motion priors into a deep reinforcement learning architecture, we have successfully addressed the critical limitations of both classical trajectory optimization and standard, unconstrained reinforcement learning. The core contribution lies in demonstrating that stylistic constraints derived from human biomechanical data do not hinder a robot's ability to navigate chaotic environments; rather, they provide a vital regularization mechanism that forces the learning agent to discover highly robust, energy-efficient, and physically plausible control strategies. The proposed system seamlessly merges high-dimensional proprioceptive feedback with exteroceptive elevation mapping, allowing the humanoid to anticipate terrain irregularities and execute complex step transitions with remarkable precision. The extensive simulation results clearly indicate that policies trained with adversarial motion priors significantly outperform standard methods in traversal success rates, particularly as the environmental difficulty escalates to severe rubble. Furthermore, the massive reduction in mechanical cost of transport and the improvement in joint torque smoothness confirm that the generated policies are not only successful in simulation but are inherently optimized for the physical limitations of real-world robotic hardware.

5.2 Limitations and Future Work

Despite the significant advancements demonstrated in this study, several limitations remain that must be addressed in future iterations. Primarily, the current framework relies on a highly structured exteroceptive elevation map, assuming that a separate perception module can continuously provide accurate local height data. While the policy demonstrated robustness to injected Gaussian noise, real-world perception failures involving complete visual occlusion or severe motion blur present a much greater challenge. Future work will investigate the integration of recurrent neural networks to better estimate terrain states purely from historical proprioception when vision is completely lost. Additionally, the current human motion dataset consists entirely of lower-body locomotion data. When traversing extremely severe rubble or disaster scenarios, humans frequently utilize their upper limbs to touch walls, grab handholds, or stabilize themselves during deep crouching transitions. Expanding the adversarial motion prior framework to encompass full-body, multi-contact scenarios, where the robot can intelligently decide to use its arms for balance on the terrain, represents the next frontier in disaster response robotics. Finally, while the kinematic metrics strongly suggest readiness for physical deployment, the ultimate validation requires extensive testing on physical humanoid hardware to fully quantify the sim-to-real transfer efficiency of the adversarial motion prior methodology [41]. Continuous refinement of actuator modeling and domain randomization techniques will be essential to ensure these highly capable simulated policies translate seamlessly to the physical world, bringing us one step closer to the realization of truly autonomous, highly agile bipedal robots capable of operating in the most demanding environments on Earth.

References

- Peng, Q., Bai, C., Zhang, G., Xu, B., Liu, X., Zheng, X., ... & Lu, C. (2025, October). NavigScene: Bridging local perception and global navigation for beyond-visual-range autonomous driving. In *Proceedings of the 33rd ACM International Conference on Multimedia* (pp. 4193-4202).
- Wan, H., Chen, Y., Deng, Y., Wei, Z., Li, D., Lin, Z., ... & Ji, X. (2025, October). EmbodiedAgent: A Scalable Hierarchical Approach to Overcome Practical

- Challenge in Multi-Robot Control. In 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 12140-12146). IEEE.
- Wang, C., Ke, L., Liu, X., Huang, D., Yao, J., & Chu, W. Explainable AI Analysis of Brake Control in CARLA Through Reference and Distilled Policies. *Frontiers in Robotics and AI*, 13, 1847808.
- Yang, Z., Ji, W., Guo, Q., Zhao, J., Zhao, C., Zhang, X., ... & You, Y. (2026). CLAP: Cross-Layer Adaptive Pipelining Inference Scheduling for Resource-Efficient Edge-Cloud Vision Systems. *ACM Transactions on Architecture and Code Optimization*.
- Yang, Y., Ma, X., Li, C., Zheng, Z., Zhang, Q., Huang, G., ... & Zhao, Q. (2021). Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 10299-10312.
- Wang, Y., Lou, M., Liang, W., & Zhang, C. (2023). Numerical and experimental investigation on tensile fatigue performance of reinforced thermoplastic pipes. *Ocean Engineering*, 287, 115814.
- Wan, H., Cheng, J., Deng, Y., Wu, D., Chen, Y., Lin, Z., ... & Ji, X. Towards Physics Aware Embodied Control with Graph based Object-centric Learning. *ACM Transactions on Cyber-Physical Systems*.
- Yang, Y., Wang, Q., Li, C., Hu, H., Wu, C., Jiang, Y., ... & Xu, B. (2025, May). Fewer may be better: Enhancing offline reinforcement learning with reduced dataset. In *International Conference on Learning Representations* (Vol. 2025, pp. 7076-7098).
- Xiao, Q., & Ghaffari, M. (2026). LaWM: Least Action World Models for Long-Horizon Physical Consistency from Visual Observations. *arXiv preprint arXiv:2605.08279*.
- Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of CVPR*.
- Luo, H., Feng, Q., Sun, W., Li, W., Li, Q., Zhao, W., & Liu, Z. (2026). FD-Mamba with neural observer and frequency-enhanced update for incipient feeder fault detection. *IEEE Internet of Things Journal*, 13(10), 20761–20772.
- Tang, A., Hiraoka, T., Hiraoka, N., Shi, F., Kawaharazuka, K., Kojima, K., ... & Inaba, M. (2024, May). Humanmimic: Learning natural locomotion and transitions for humanoid robot via wasserstein adversarial imitation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 13107-13114). IEEE.
- Zhang, Luyan. "MCP: A Control-Theoretic Orchestration Framework for Synergistic Efficiency and Interpretability in Multimodal Large Language Models." *arXiv preprint arXiv:2509.16597* (2025).
- Shuang, K., Guo, J., & Wang, Z. (2022). Comprehensive-perception dynamic reasoning for visual question answering. *Pattern Recognition*, 131, 108878.
- Wan, H., Zhang, Y., Wang, J., Wu, D., Li, M., Chen, X., ... & Ji, X. (2025). Toward universal embodied planning in scalable heterogeneous field robots collaboration and control. *Journal of Field Robotics*, 42(5), 2318-2336.
- Wang, Y., Fang, Y., Wang, T., Li, L., Liu, K., Xu, J., ... & Xu, R. (2026). What Limits Vision-and-Language Navigation?. *arXiv preprint arXiv:2605.13328*.
- Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024). A conditional denoising diffusion probabilistic model for point cloud upsampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 20786-20795).

- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al. (2023). Segment anything. In Proceedings of ICCV.
- Li, Q., Luo, H., Cheng, H., Deng, Y., Sun, W., Li, W., & Liu, Z. (2023). Incipient fault detection in power distribution system: A time-frequency embedded deep-learning-based approach. *IEEE Transactions on Instrumentation and Measurement*, 72, 1–14.
- Li, Y., Yang, J., Yang, Z., Li, B., Han, L., He, H., ... & Tang, R. (2026, March). Make lvlms focus: Context-aware attention modulation for better multimodal in-context learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 8, pp. 6610-6618).
- Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*.
- Sun, W., Dong, X. M., Cui, B., & Tang, J. (2025, April). Attentive eraser: Unleashing diffusion model's object removal potential via self-attention redirection guidance. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 19, pp. 20734-20742).
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In Proceedings of ECCV.
- Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*.
- Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., et al. (2023). PaLM-E: An embodied multimodal language model. In Proceedings of ICML.
- Cheng, H., Luo, H., Liu, Z., Sun, W., Li, W., & Li, Q. (2025). Reinforcement learning based robust volt/var control in active distribution networks with imprecisely known delay. *IEEE Transactions on Computational Social Systems*, 1–12.
- Ma, W., Li, Y., Liu, C., Zhang, H., Li, J., Chen, K., & Gao, W. (2026). GeoCraft: A Diffusion Model-Based 3D Reconstruction Method Driven by Image and Point Cloud Fusion. *Information Fusion*, 104149.
- Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015). Trust region policy optimization. In Proceedings of ICML.
- Yu, D., Mahamadu, A. M., Chen, W., & Li, N. (2026). Understanding the challenges and future trends of immersive technology use in indoor pedestrian and evacuation dynamics: A systematic review. *Journal of Building Engineering*, 115703.
- Mur-Artal, R., Montiel, J. M. M., & Tardós, J. D. (2015). ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5), 1147–1163.
- Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 27325-27335).
- Yang, Y., Hu, H., Li, W., Li, S., Yang, J., Zhao, Q., & Zhang, C. (2023, June). Flow to control: Offline reinforcement learning with lossless primitive discovery. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 9, pp. 10843-10851).
- Zhu, W., Zhang, Y., Jin, X., Zeng, W., & Zhang, L. (2026). ANTS: Adaptive Negative Textual Space Shaping for OOD Detection via Test-Time MLLM

- Understanding and Reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20-30).
- Li, B., Zhang, R., Liang, H., Zhang, J., Zhang, J., Chen, X., ... & Wang, J. (2026). Interagent: Physics-based multi-agent command execution via diffusion on interaction graphs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 15253-15265).
- Li, X., Wang, P., Li, G., & Zhang, Y. (2023). Design of a novel self-test-on-chip interface ASIC for capacitive accelerometers. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 70(7), 2834-2843.
- Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing Atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.
- Levine, S., Pastor, P., Krizhevsky, A., & Quillen, D. (2018). Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *International Journal of Robotics Research*, 37(4–5), 421–436.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In Proceedings of CVPR.
- Koenig, S., & Likhachev, M. (2002). D* Lite. In Proceedings of AAAI.
- Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). NeRF: Representing scenes as neural radiance fields for view synthesis. In Proceedings of ECCV.
- Sui, M., Jiang, L., Lyu, T., Wang, H., Zhou, L., Chen, P., & Alhosain, A. (2024). Application of deep learning models based on efficientdet and openpose in user-oriented motion rehabilitation robot control. *Journal of Intelligence Technology and Innovation*, 2(3), 47-77.