Towards Adaptive Humanoid Control via Multi-Behavior Distillation and Reinforced Fine-Tuning

Yingnan Zhao[1,7], **Xinmiao Wang**[1,2], **Dewei Wang**[2,3], **Xinzhe Liu**[2,4], **Dan Lu**[1,7*], **Qilong Han**[1,7], **Peng Liu**[5], **Chenjia Bai**[2,6*]

[1]College of Computer Science and Technology, Harbin Engineering University
[2]Institute of Artificial Intelligence (TeleAI), China Telecom
[3]School of Information Science and Technology, University of Science and Technology of China
[4]School of Information Science and Technology, ShanghaiTech University
[5]College of Computer Science and Technology, Harbin Institute of Technology
[6]Shenzhen Research Institute of Northwestern Polytechnical University
[7]National Engineering Laboratory for Modeling and Emulation in E-Government, Harbin Engineering University
{zhaoyingnan, wangxinmiao}@hrbeu.edu.cn, bz24006012@mail.ustc.edu.cn, liuxzh12023@shanghaitech.edu.cn,
{ludan, hanqilong}@hrbeu.edu.cn, pengliu@hit.edu.cn, baicj@chinatelecom.cn

Abstract

Humanoid robots are promising to learn a diverse set of human-like locomotion behaviors, including standing up, walking, running, and jumping. However, existing methods predominantly require training independent policies for each skill, yielding behavior-specific controllers that exhibit limited generalization and brittle performance when deployed on irregular terrains and in diverse situations. To address this challenge, we propose *Adaptive Humanoid Control (AHC)* that adopts a two-stage framework to learn an adaptive humanoid locomotion controller across different skills and terrains. Specifically, we first train several primary locomotion policies and perform a multi-behavior distillation process to obtain a basic multi-behavior controller, facilitating adaptive behavior switching based on the environment. Then, we perform reinforced fine-tuning by collecting online feedback in performing adaptive behaviors on more diverse terrains, enhancing terrain adaptability for the controller. We conduct experiments in both simulation and real-world experiments in Unitree G1 robots. The results show that our method exhibits strong adaptability across various situations and terrains.

Website — https://ahc-humanoid.github.io

Introduction

Humanoid robots, due to their human-like morphology, are expected to possess various fundamental human-like locomotion abilities, such as walking, running, and standing up after a fall. Previous studies adopt methods such as whole-body control (Sentis and Khatib 2006), model predictive control (Li and Nguyen 2023), and reinforcement learning (RL) (Ernst and Louette 2024; Wang et al. 2025b) to enhance the locomotion capabilities of humanoid robots. Recently, RL has achieved a remarkable progress for humanoid locomotion (Gu, Wang, and Chen 2024; Gu et al. 2024a; Xie et al. 2025), supported by large-scale simulation (Makoviychuk et al. 2021; Zakka et al. 2025) and advanced policy gradient methods (Schulman et al. 2017; Engstrom et al. 2019).

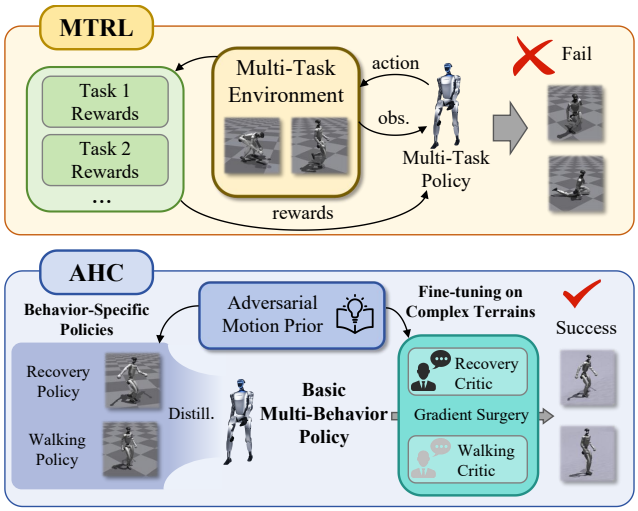

Figure 1: Comparison between multi-task RL and our proposed framework. Directly learning multiple skills via multi-task RL is challenging. Therefore, we adopt a two-stage framework consisting of behavior distillation and reinforced fine-tuning, enabling the acquisition of diverse humanoid robot skills and generalization to complex terrains.

However, existing locomotion algorithms predominantly require training an independent policy with a separately designed reward function, such as standing up (He et al. 2025b; Huang et al. 2025b), jumping (Tan et al. 2024), squatting (Ben et al. 2025), and walking (Gu, Wang, and Chen 2024). Such a training paradigm yields independent controllers that excel in performing specific skills, while they exhibit limited generalization abilities in both behavior diversity and terrain

*Corresponding authors.

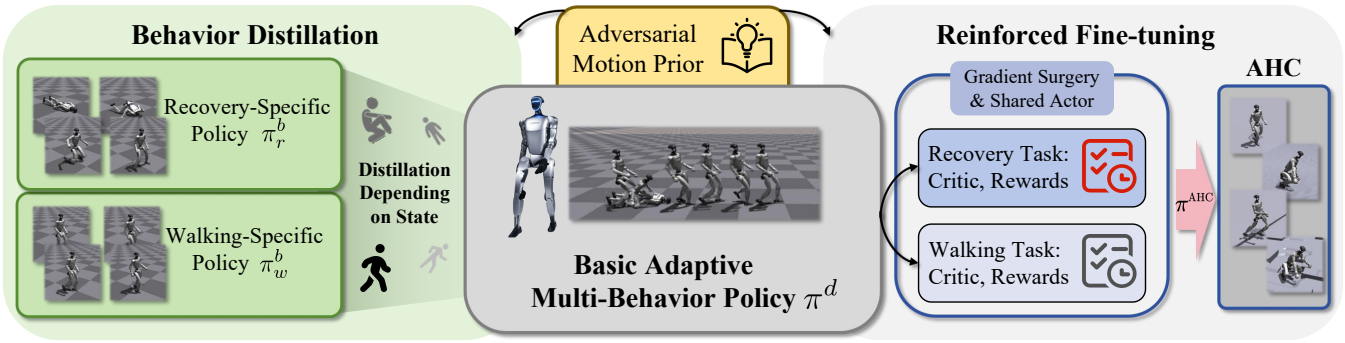

Figure 2: Overview of the proposed two-stage framework *Adaptive Humanoid Control*. In the first stage, we train two separate primary policies on flat terrain. These policies are then distilled into a basic multi-behavior policy via distillation. In the second stage, we perform reinforced fine-tuning on the distilled policy, employing gradient surgery to alleviate gradient conflicts and utilizing behavior-specific critics to provide more accurate value estimation.

adaptability. It is noted that directly extending the learning paradigm to multi-skill setting is challenging, which primarily arises from the conflicts of policy gradients caused by different reward functions (Yu et al. 2020; Liu et al. 2021), hindering the convergence of multi-skill policies.

Although several works focus on quadruped robot multi-skill learning using Mixture-of-Experts (MoE) or policy distillation (Wang et al. 2025a; Yang et al. 2020; Huang et al. 2025a), multi-skill learning for humanoid robots still remains challenging and requires further exploration (Zhuang, Yao, and Zhao 2024; Long et al. 2024a).

In this work, we propose a novel *Adaptive Humanoid Control (AHC)* method, as illustrated in Figure 1, which adopts a two-stage training framework that first learns a basic multi-behavior controller and then enhances its terrain adaptability to achieve adaptive humanoid control. In the first stage, we integrate human motion priors via Adversarial Motion Prior (AMP) with independent behavior-specific policy learning resulting several human-like basic controllers. Then, we adopt a policy distillation process to obtain a basic multi-behavior controller, which facilitates adaptive behavior switching based on the robot's state and also addresses the challenges in direct RL training of the multi-behavior policy. In the second stage, the basic controller collects trajectories and reward feedback in performing different behaviors in different terrains, which is used for continuously improving the terrain adaptability of each behavior via a RL tuning process. We divide the skills into separate tasks and formulate the training as a multi-task RL problem and adopt gradient projection to mitigate the impact of gradient conflicts.

We evaluate the obtained terrain-adapted multi-behavior policy in both simulation and a real-world Unitree G1 robot. Experimental results show that our controller adapts effectively to changes in environmental state (e.g. can stand up after falling and walk) and can perform robust locomotion on challenging terrains (e.g., stairs and slope).

Our contributions are summarized as follows: (i) Instead of direct training multi-behavior RL policy across diverse terrains, we employ motion-guided policy learning and supervised distillation to obtain the basic multi-behavior policy. (ii) We adopt sample-efficient RL fine-tune on the basic policy to obtain terrain-adaptive behaviors. (iii) We conduct extensive experiments to verify the learned controller is applicable in both simulation and real-world experiments.

Methodology

Preliminaries and Problem Definitions

In the first stage of *AHC*, each behavior-specific humanoid control problem is defined as a Markov Decision Process (MDP) defined as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state space and action space respectively, $\mathcal{P}(\cdot|\boldsymbol{s}, \boldsymbol{a})$ is the state transition function, $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{R}$ is the reward function and $\gamma \in [0, 1)$ is the reward discount factor. During training, the behavior-specific policy π^b learns to take a state $\boldsymbol{s}_t$ from $\mathcal{S}$ to output an action $\boldsymbol{a}_t$ to maximize the discounted return $\mathbb{E}\left[\sum_{t=1}^{T} \gamma^{t-1} R(\boldsymbol{s}_t, \boldsymbol{a}_t)\right]$. The adaptive humanoid control problem is formulated as a multi-task RL problem where each task can be seen as a MDP $\mathcal{M}_\mathrm{i} = (\mathcal{S}_\mathrm{i}, \mathcal{A}_\mathrm{i}, \mathcal{P}_\mathrm{i}, R_\mathrm{i}, \gamma_\mathrm{i})$, $\mathrm{i} \in [1, N]$. Since all tasks (i.e., behaviors) are conducted under a unified environment setup and the controller needs to perform different behavior according to the states ($\mathcal{S} = \bigcup_\mathrm{i} \mathcal{S}_\mathrm{i}, \mathcal{S}_\mathrm{i} \cap \mathcal{S}_\mathrm{j} = \emptyset$ for $\mathrm{i} \neq \mathrm{j}$), each MDP differs only in its reward function R_i and state space $\mathcal{S}_\mathrm{i}$. The objective that behavior-adaptive policy needs to optimize is

$$\sum_{i=1}^{N} \mathbb{E}_{\mathcal{P}, \pi_i} \left[\sum_{t=1}^{T} \gamma^{t-1} R_i(\boldsymbol{s}_t^i, \boldsymbol{a}_t) \right], \boldsymbol{s}_t^i \in \mathcal{S}_i. \tag{1}$$

Behavior-specific policies take privileged information s_t^{priv} and robot proprioception s_t^{prop} as inputs since they are not directly deployed in a real robot. In the distillation process, we distill the knowledge of behavior-specific policies ($\pi_1^\mathrm{b}, \pi_2^\mathrm{b}$, ...) into a basic multi-behavior policy π^d, which aims to perform adaptive behaviors using only proprioception s_t^{prop} according to different situations (e.g. recovering from fails and walking). The distilled policy for RL fine-

tuning using only s_t^{prop} to predict the action $\boldsymbol{a}_t$ for robot control by a PD controller.

The robot proprioception s_t^{prop} consists of the following parts:

$$s_t^{\text{prop}} = [\boldsymbol{\omega}_t, \boldsymbol{g}_t, \boldsymbol{c}_t, \boldsymbol{q}_t, \dot{\boldsymbol{q}}_t, \boldsymbol{a}_{t-1}] \in \mathbb{R}^{69}, \tag{2}$$

where $\boldsymbol{\omega}_t \in \mathbb{R}^3$ is the base angular velocity, $\boldsymbol{g}_t \in \mathbb{R}^3$ is the gravity vector in the base frame, $\boldsymbol{c}_t \in \mathbb{R}^3$ is the velocity command comprising linear velocities along the x- and y-axes and an angular velocity around the z-axis, $\boldsymbol{q}_t \in \mathbb{R}^{20}$ and $\dot{\boldsymbol{q}}_{\boldsymbol{t}} \in \mathbb{R}^{20}$ represent joint position and joint velocity, and $\boldsymbol{a}_{t-1} \in \mathbb{R}^{20}$ is the last action. The privileged information s_t^{priv} includes ground friction coefficient, motor controller gains, base mass and center of mass shift. The action $\boldsymbol{a}_t$ is converted into a PD target by $\boldsymbol{q}_t^{\text{target}} = \boldsymbol{q}^{\text{default}} + \alpha \boldsymbol{a}_t$ which is used to compute the motor torques:

$$\boldsymbol{\tau}_t = K_p \cdot (\boldsymbol{q}_t^{\text{target}} - \boldsymbol{q}^{\text{default}}) - K_d \cdot \dot{\boldsymbol{q}}_{\boldsymbol{t}}, \tag{3}$$

where K_p and K_d are the stiffness and damping coefficients of the PD controller, and α is a scalar to bound the action.

The remainder of this section is organized as follows: First, we introduce two fundamental behavior-specific policies: one for robust recovery from various fallen postures and another for human-like locomotion on flat terrain. These policies are then utilized for distilling a basic multi-behavior policy. Second, to enhance terrain adaptability, the distilled multi-behavior policy is fine-tuned across diverse terrains, enabling adaptive learning of both walking and recovery behaviors on more complex terrains. An overview of this two-stage framework is illustrated in Figure 2.

Multi-Behavior Distillation

Training an adaptive multi-behavior controller from scratch via online RL is challenging, as diverse behaviors require distinct environment settings and reward functions, often leading to poor policy convergence (Sodhani, Zhang, and Pineau 2021; Liu et al. 2021). This is primarily caused by issues such as gradient conflict and gradient imbalance. To address this issue and enable the policy to learn multiple primary behaviors, we first train two basic behavior-specific policies using Proximal Policy Optimization (PPO) (Schulman et al. 2017), which are then distilled into a basic multi-behavior policy π^d.

Falling Recovery Behavior Policy π_r^b Recovering from unexpected falls is an essential capability for ensuring the robustness and autonomy of humanoid robots. HoST (Huang et al. 2025b) has shown that a standing-up control policy can be trained via multiple critics and force curriculum. We also employ multiple critics (Mysore et al. 2022) for our basic standing-up policy π_r^b. In the surrogate loss for policy gradient, the advantage function is estimated using a weighted formulation $\hat{A} = \sum_{i=0}^{n} \omega_i \cdot (\hat{A}_i - \mu_{\hat{A}_i})/\sigma_{\hat{A}_i}$, where ω_i is a weighting coefficient, and $\mu_{\hat{A}_i}$ and $\sigma_{\hat{A}_i}$ correspond to the batch mean and standard deviation of the advantage from group i, respectively.

Different from HoST, we let the policy focus on recovery on flat terrains. Specifically, the robot is initialized in either a supine or prone position, with additional joint initialization noise introduced to encourage learning robust recovery behaviors from various postures. To mitigate the negative impact caused by interference between sampled rollouts from different initial postures (Huang et al. 2025b) and to promote more natural standing-up motions, we introduce AMP-based (Peng et al. 2021; Escontrela et al. 2022) reward function using a discriminator which determines whether an episode is a positive sample (i.e., from the reference motion) or a negative sample (i.e., from the robot). The output of the discriminator can be used to guide the robot to recover in a smooth and plausible manner. Consequently, the standing-up policy π_r^b can robustly recover from diverse abnormal postures and learns behaviors from the reference motion, such as leveraging the arms to support the ground during the standing-up process. The AMP-based reward formulation and the objective of the discriminator can be found in Appendix A.

Walking Behavior Policy π_w^b Locomotion across diverse terrains is also essential for humanoid robots. Although complex system designs and multi-stage training pipelines can enable robots to learn diverse locomotion skills (Lin et al. 2025; Zhuang, Yao, and Zhao 2024), we only adopt a simple framework and reward functions design to enable the policy π_w^b to learn fundamental locomotion ability.

π_w^b is capable of walking on flat terrain in response to a velocity command $\boldsymbol{c}_t$. To enable the robot to move in a human-like manner and accelerate the convergence process, an AMP-based reward function is introduced similar to that of π_r^b. It is worth noticing that although the policy π_w^b is only capable of walking, after undergoing distillation and RL fine-tuning, our policy is able to adapt to diverse terrains while exhibiting significantly improved robustness to external disturbances. The PPO parameters, reward designs, network architecture and domain randomization terms for the basic behavior policies training are listed in Appendix B.

Behavior Distillation In the behavior distillation process, we use DAgger (Chen et al. 2020; Ross, Gordon, and Bagnell 2011) to distill different behavior knowledge into an MoE-based multi-behavior policy π^d, which eliminates the gradient conflicts resulting from different reward landscapes of various behaviors.

The MoE module can automatically assign different experts to learn distinct behaviors, enabling the policy to leverage such prior knowledge for efficient multi-behavior improvement and multi-terrain adaptability in subsequent RL fine-tuning stage. During the training process, the robot is initialized in a fallen or standing posture and the policy is supervised by π_r^b or π_w^b according to which behavior it should perform. The loss function of π^d is computed by:

$$\mathcal{L}_{\pi^d}(\boldsymbol{s}_t) = \begin{cases} \mathbb{E}_{s_t, \pi_d, \pi_r^b} \left[\| \boldsymbol{a}_t^{\pi^d} - \boldsymbol{a}_t^{\pi_r^b} \|_2^2 \right], & \boldsymbol{s}_t \in \mathcal{S}_r \\ \mathbb{E}_{s_t, \pi_d, \pi_w^b} \left[\| \boldsymbol{a}_t^{\pi^d} - \boldsymbol{a}_t^{\pi_w^b} \|_2^2 \right], & \boldsymbol{s}_t \in \mathcal{S}_w \end{cases}, \tag{4}$$

where $\boldsymbol{a}_t^{\pi^d}$, $\boldsymbol{a}_t^{\pi_r^b}$ and $\boldsymbol{a}_t^{\pi_w^b}$ are sampled from their corresponding policy, $\mathcal{S}_r$ and $\mathcal{S}_w$ represent standing-up and walking state space. The distillation process employs the same domain randomization as the behavior-specific policy training

and π^d takes only proprioception as input. The distillation process not only integrates the basic behaviors into a single policy but also enhances each of them individually. Specifically, π^d exhibits more robust walking performance, as it learns to recover from near-fall postures, and demonstrates a more natural standing pose after a stand-up behavior, which facilitates smoother transitions into walking.

RL Fine-Tuning

In the RL fine-tuning stage, we formulate the problem as a multi-task RL problem, where the policy π^{ft} is initialized with the distilled policy π^d from last stage and learns fail recovery task and walking task on complex terrains. To maintain human-likeness, we also adopt an AMP-based reward in this stage using the same reference motion as in the previous stage. Leveraging the MoE module and prior knowledge in basic multi-behavior policy π^d, the gradient conflict problem is alleviated, enabling efficient learning of adaptive behaviors on various terrains. We fine-tune the policy π^{ft} using PPO on two GPUs, where each GPU handles one task under its corresponding environment setup. The policies for different task share the same set of parameters.

Behavior-Specific Critics and Shared Actor In the PPO algorithm, the policy gradient is computed using normalized advantages, which stabilizes the surrogate loss by standardizing the scale of the advantage estimates. However, the value loss relies on unnormalized return targets, since the critic must approximate the true expected return to provide meaningful value estimates.

To prevent tasks with larger reward scales from dominating gradient updates and hindering others' learning (Chen et al. 2018; Hessel et al. 2019), we use behavior-specific critics with a shared actor during fine-tuning. By assigning a separate critic to each task, we isolate value function learning for task-specific reward functions which enables more accurate value estimation and allows for customized critic architectures for each task (e.g., multiple critics for standing-up behavior). Meanwhile, a shared actor is updated using policy gradients aggregated across tasks, allowing skill transfer and terrain adaptability.

Eliminating Gradient Conflict in Multi-task Learning While the use of behavior-specific critics addresses discrepancies in gradient magnitude, the shared actor still aggregates potentially conflicting gradients from different tasks.

To alleviate this issue, we apply Projecting Conflicting Gradients (PCGrad) (Yu et al. 2020) to resolve gradient conflicts during optimization. For each pair of task gradients, if they conflict (i.e., their cosine similarity is negative), the gradient of one task is projected onto the normal plane of the other, removing the conflicting direction while preserving progress on the remaining subspace. Specifically, given two task gradients $\mathbf{g}_i$ and $\mathbf{g}_j$, the projected gradient is computed as:

$$\mathbf{g}_i = \mathbf{g}_i - \frac{\mathbf{g}_i \cdot \mathbf{g}_j}{\|\mathbf{g}_j\|^2}\mathbf{g}_j \tag{5}$$

In practice, we integrate PCGrad before the actor update step: each task computes its local actor gradient on its dedicated GPU, and all gradients are then communicated to the main process, where PCGrad is applied to perform gradient surgery. After performing the optimizer step with the conflict-free gradients on the main process, we broadcast the updated parameters back to all workers. PCGrad allows the shared actor to learn multiple task-specific skills without gradient conflicts, ensuring efficient multi-task RL learning. The detailed process can be found in Appendix B.

Terrain Curriculum Following the previous work (Rudin et al. 2022), we adopt the terrain curriculum to improve learning efficiency and adaptability on diverse terrains. An automatic level adjustment mechanism modulates terrain difficulty based on task-specific performance. We design flat, slope, hurdle and discrete terrains for both tasks. The maximum slope inclination is 16.6°. Hurdle terrains are composed of regularly spaced vertical obstacles with a maximum height $0.1m$. The discrete terrain consists of randomly placed rectangular blocks. Specifically, we generate 20 rectangular obstacles with width and lengths sampled between $0.5m$ and $2.0m$, and heights uniformly sampled between $0.03m$ and $0.15m$. We arrange the terrain map into a 10 × 12 grid of $8m\times$ $8m$ patches, with 10 difficulty levels and 3 columns per terrain type.

Experimental Results and Discussion

We perform training in the IsaacGym simulator, using 4096 parallel environments. We train the behavior-specific policies for 10,000 iterations, followed by 4,000 iterations of policy distillation. Fine-tuning for terrain adaptability runs an additional 10,000 iterations using online RL on two NVIDIA RTX 4090 GPUs. The 20-DoF action space (excluding waist joints) is used, with AMP rewards from retargeted motion capture for recovery and LAFAN1 data for locomotion. The policy is deployed on a Unitree G1 humanoid robot at 50Hz, tracked by a 500Hz PD controller converting joint positions to torques.

Simulation Results

Terrain Adaptability To evaluate the terrain-adaptive capability of our proposed *AHC* framework, we compare it with the following methods: (1) **HOMIE** (Ben et al. 2025): We adapt the lower-body locomotion policy from HOMIE to our setting by re-training it on our terrain settings. (2) **HoST** (Huang et al. 2025b): We train the standing-up controller from HoST on our terrain settings. (3) **Fail Recovery Policy** π_r^b: The fail recovery behavior policy trained in the first stage of our framework. (4) **Walking Policy** π_w^b: The walking behavior policy trained in the first stage of our framework. (5) **Basic Multi-Behaviors Policy** π^d: The basic multi-behavior policy from the first stage distillation process.

We construct four types of terrain patches for evaluation, each of size 8 $m\times$ 8 m, corresponding to the four terrain types used during training: flat, slope, hurdle, and discrete. On the slope terrain, the inclination angle is uniformly sampled between 12° and 16°. The hurdle obstacle heights uniformly sampled between $0.08m$ to $0.1m$. For the locomotion task, the hurdle terrain includes 3 obstacles, while the recovery task uses a more cluttered setup with 8 obstacles. The discrete terrain consists of randomly positioned rectangular

Method	Locomotion								Fail Recovery			
	Plane		Slope		Hurdle		Discrete		Plane	Slope	Hurdle	Discrete
	Succ.	Dist.	Succ.	Dist.	Succ.	Dist.	Succ.	Dist.	Succ.	Succ.	Succ.	Succ.
HOMIE (Ben et al. 2025)	0.802	6.421	0.599	4.795	0.407	3.259	0.442	3.603	-	-	-	-
HoST (Huang et al. 2025b)	-		-		-		-		0.997	0.978	0.911	0.843
Fail Recovery Policy	-		-		-		-		**1.000**	0.757	**0.999**	0.942
Walking Policy	0.971	7.782	0.000	2.197	0.161	2.561	0.127	3.788	-	-	-	-
Multi-Behaviors Policy	**0.993**	**7.969**	0.160	5.850	0.756	7.254	0.702	7.321	0.999	0.904	0.997	0.947
AHC	0.992	7.951	**0.975**	**7.987**	**0.922**	**7.866**	**0.969**	**7.951**	**1.000**	**1.000**	0.985	**0.969**

Table 1: Performance evaluation on locomotion and fail recovery tasks across different terrains. To compare our approach with the baselines, we report the traversal success rate (Succ.) and average traversing distance (Dist.) for the locomotion task, and the success rate (Succ.) for the fail recovery task.

blocks, with height variations ranging from $0.08m$ to $0.1m$. Each policy is evaluated separately on these terrain types to assess its performance.

We adopt **Success Rate (Succ.)** as the primary metric for both tasks. For the locomotion task, success rate is defined as the percentage of trials in which the robot traverses the full $8m$ terrain within $20s$ without termination. During evaluation, the robot is assigned a fixed forward velocity at the beginning of each episode, uniformly sampled from $0.4m/s$-$1.0m/s$. An episode is terminated if the robot walks off the current $8m\times8m$ terrain patch or falls irrecoverably. For policies that integrate both fail recovery and locomotion capabilities (i.e., multi-behavior policy π^d and *AHC* policy π^{AHC}), falling does not trigger termination, allowing the robot to autonomously recover and resume traversal. For recovery task, a trial is considered successful if the robot can stand up from a fallen posture and maintain balance without falling again within $10s$. In addition, we report **Traversing Distance (Dist.)**, defined as the average distance covered before termination, computed over all trials, including both successful and failed ones. All evaluations are conducted using 1000 parallel environments.

As shown in Table 1, our proposed *AHC* policy π^{AHC} outperforms both HOMIE and HoST on the locomotion and recovery tasks across a wide range of terrain types. In locomotion task, the performance gain is largely attributed to π^{AHC}'s ability to autonomously recover from falls and resume traversal, especially on terrains with high obstacle density (e.g., hurdle and discrete terrains). In addition, the incorporation of AMP provides motion priors that guide the policy toward learning stable and robust behaviors, contributing to π^{AHC}'s better performance compared to HoST. We further compare the multi-behavior policy π^d with the behavior-specific policies (i.e., π^b_r and π^b_w). The multi-behavior policy π^d demonstrates superior robustness in the locomotion task on complex terrains such as hurdle and discrete. This improvement stems from its seamless integration of walking and recovery behaviors within a single policy. These results underscore the promise of integrating complementary skills such as walking and recovery into a unified policy to achieve robust locomotion in challenging environments. To further enhance terrain adaptability, we apply RL to fine-tune π^d on diverse terrains, resulting in the final *AHC* policy π^{AHC}.

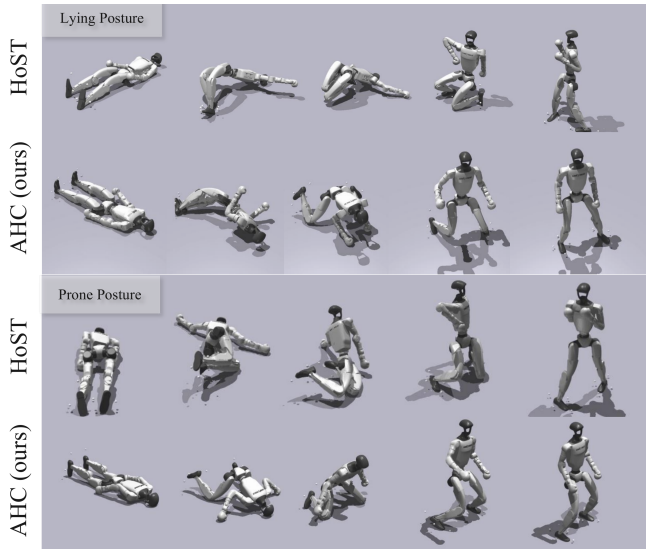

Figure 3: Comparison of recovery motions under AHC and HoST. We compare our AHC (with AMP) against the HoST (w/o AMP) in both lying and prone scenarios. AHC produces smoother recovery behaviors. This highlights the effectiveness of AMP in guiding the learning of naturalistic recovery motions.

This additional stage brings improvements on most terrain types across both tasks, particularly in slope and discrete terrains, highlighting the transferability of our two-stage training framework.

AMP for Standing-Up We compare *AHC*, which incorporates AMP, with the HoST (Huang et al. 2025b) baseline without human motion priors. As shown in Figure 3, we visualize a sequence of snapshots during the stand-up process under both approaches in lying and prone posture. Without AMP (i.e., HoST), the robot exhibits uncoordinated and jerky motions, relying on abrupt limb movements to return to a standing posture. In contrast, *AHC* policy generates a natural get-up motion, including leg folding, arm support, and trunk lifting. We further compare the joint velocity accelerate during the recovery. As shown in Figure 4, the velocity curves of key joints (i.e., hip and knee) for *AHC* pol-

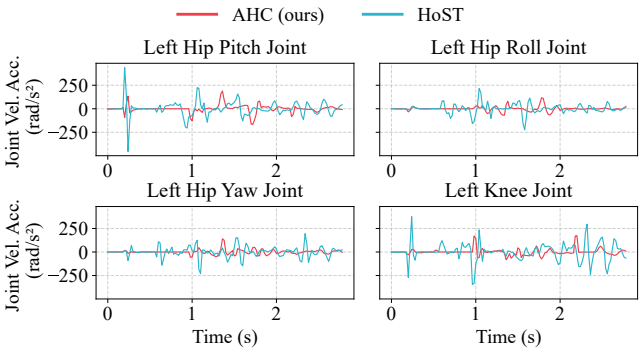

Figure 4: Joint acceleration analysis of the left leg during recovery. Acceleration profiles of hip and knee joints from the left leg illustrate that our AHC results in stable joint actuation, with notably fewer abrupt fluctuations compared to HoST.

Method	Cosine Similarity (↑)
AHC-SC-w/o-PC	0.247
AHC-SC-PC	**0.519**
AHC-BC-w/o-PC	0.334
AHC (ours)	**0.535**

Table 2: Gradient Cosine Similarity between Tasks across Different Ablation Settings. A lower similarity indicates higher conflict between task gradients.

icy exhibit fewer abrupt fluctuations compared to the policy trained with HoST. These results demonstrate that AMP helps shape the recovery controller towards producing stable motions, which are difficult to obtain with handcrafted reward functions.

Ablation on PCGrad and Behavior-Specific Critics We conduct a comprehensive ablation study to evaluate the contributions of the two key components introduced in the second-stage fine-tuning stage: PCGrad and behavior-specific critics update strategy. We examine four configurations: (1) **AHC-SC-w/o-PC**: a single shared critic without PCGrad. (2) **AHC-SC-PC**: a single shared critic with PCGrad. (3) **AHC-BC-w/o-PC**: behavior-specific critics without PCGrad. (4) **AHC (Ours)**: behavior-specific critics with PCGrad. In the single critic setting, a shared critic network is jointly optimized across all tasks.

To quantify the role of PCGrad in mitigating gradient conflicts during multi-task optimization, we compute the average cosine similarity between the gradients of the two tasks during the second-stage training. As shown in Table 2, PCGrad reduce gradient conflict, resulting in higher cosine similarity values in both critic settings. Notably, the use of behavior-specific critics also yield higher similarity, indicating that they help alleviate gradient conflicts between tasks. We further investigate the impact of adopting behavior-specific critics by monitoring the evolution of value loss during training. As illustrated in Figure 5, models

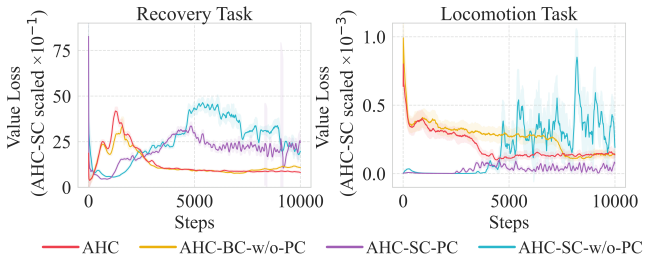

Figure 5: Value loss curves during the second-stage fine-tuning. Policies equipped with behavior-specific critics (AHC-BC-w/o-PC and *AHC*) indicate more stable value learning compared to their shared-critic counterparts (AHC-SC).

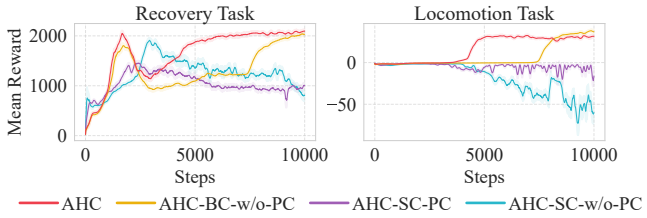

Figure 6: Training episode return curves during second-stage fine-tuning. With PCGrad and behavior-specific critics *AHC* achieve higher and more balanced returns across tasks.

with behavior-specific critics (AHC-BC-w/o-PC and AHC) achieve lower value loss compared to their shared-critic counterparts. This suggests that decoupling value learning for each task helps mitigate optimization difficulties caused by reward scale discrepancies. In addition, we visualize the training episode return curves in Figure 6 to evaluate how well each configuration balances learning across tasks. The shared-critic variants (AHC-SC) tend to neglect the locomotion task due to its smaller reward magnitude. In contrast, *AHC* maintains high returns for both tasks, demonstrating superior performance in multi-task learning. Notably, AHC exhibits faster convergence compared to the other settings. These results highlight the effectiveness of incorporating both PCGrad and behavior-specific critics in facilitating balanced and efficient optimization during the second stage.

Deployment Results

We deploy our trained policy to a Unitree G1 humanoid robot in real-world settings without additional fine-tuning, validating its effectiveness across diverse scenarios. A sequence of deployment snapshots is shown in Figure 7. To evaluate robustness and generalization, we conduct multiple trials in our laboratory environment under different conditions. For the recovery evaluation, we place the robot in both supine and prone init position on flat ground and inclined terrain. In all cases, the robot successfully recovers from various fallen postures, including moderate external disturbances. After each recovery, it stabilizes itself and smoothly transitions into a walking-ready posture, displaying natural and coordinated motion. For the locomotion task, we test the policy in two initialization settings: recovery followed

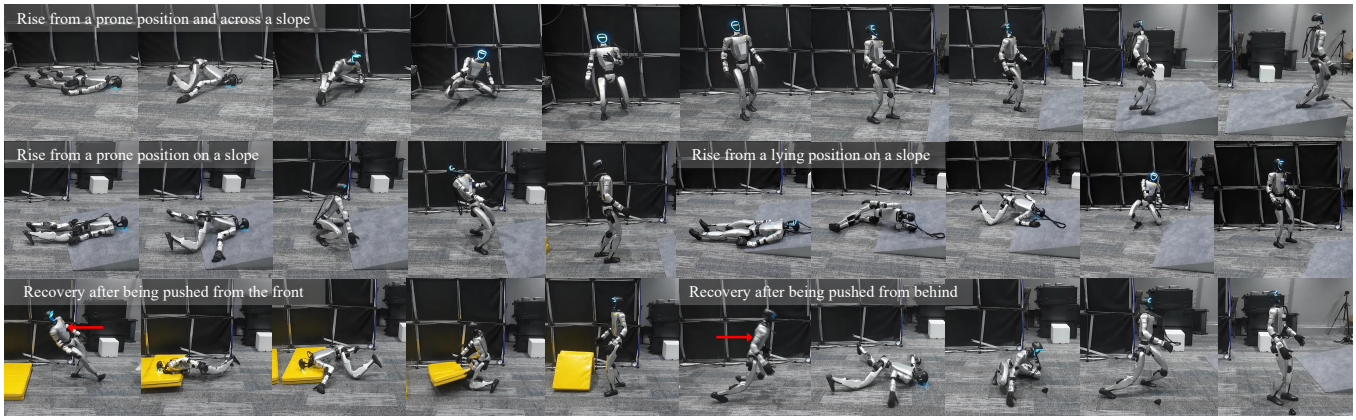

Figure 7: Snapshot of real-world deployment. The robot performs recovery and locomotion in diverse scenarios, including standing up from prone and lying positions on sloped terrain and recovering after external pushes during walking.

by walking, and directly starting from a standing posture. In both cases, the robot is able to walk stably on flat ground and inclined surfaces, demonstrating robust control and effective tracking of velocity commands. During walking, we apply external pushed in random directions to assess the robot's ability to maintain balance. The robot generally withstands perturbations and continues walking. Even when a fall does occur during walking, the robot is able to autonomously perform the recovery maneuver and resume locomotion across the terrain, exhibiting strong resilience and long-horizon autonomy. These results suggest that the learned policy not only bridges the sim-to-real gap effectively, but also integrate recovery and locomotion behaviors in a cohesive and robust manner.

Related Work

Humanoid Locomotion Deep reinforcement learning (RL) algorithms have enabled humanoid robots to perform robust and even highly complex locomotion behaviors in simulation environments (Makoviychuk et al. 2021; Todorov, Erez, and Tassa 2012). Without relying on external sensors, prior works have demonstrated several challenging behaviors, including coordinated control of the upper and lower body (Shi et al. 2025), whole-body locomotion (Radosavovic et al. 2024), robust traversal over complex terrains (Gu et al. 2024b), and highly flexible full-body teleoperation (Ben et al. 2025). Fundamental capabilities such as fall recovery have also been achieved through techniques like using multiple critics (Huang et al. 2025b). The RL policy with external sensors like depth cameras and LiDARs is able to perceive the environment, like terrains and obstacles, and learn to avoid obstacles and leap a gap (Long et al. 2024b; Ren et al. 2025). More extreme terrains with sparse footholds can be traversed by employing finer terrain perception or attention-based network designs (Wang et al. 2025c; He et al. 2025a). However, all these works focus solely on a single behavior, such as moving or recovery. In contrast, our proposed framework enables the robot to acquire multiple skills and autonomously select appropriate behaviors based on its current state.

Multi-Behavior Learning in Robots Mastering multiple behaviors is essential for enhancing the adaptability and practical application of robots. Policy distillation allows the integration of skills from multiple expert policies into a single policy, enabling diverse behaviors for navigating complex terrains (Zhuang et al. 2023; Zhuang, Yao, and Zhao 2024). Alternatively, hierarchical frameworks can select among multiple skill policies to facilitate efficient multi-skill traversal (Hoeller et al. 2024). HugWBC (Xue et al. 2025) leverages input signals such as gait frequency and foot contact patterns to guide the policy, allowing it to exhibit different behaviors in response to varying commands. MoE-Loco (Huang et al. 2025a) adopts an MoE architecture to reduce gradient conflicts in multi-skill RL, thereby improving training efficiency. MoRE (Wang et al. 2025b) further enhances policy performance by incorporating AMP-based rewards and external sensor inputs. However, prior works typically achieve multi-behavior capabilities either through explicit control signals or by combining behaviors with high similarity (e.g., stair climbing and gap jumping). In contrast, our proposed framework integrates highly diverse behaviors into a single unified policy and enables the robot to autonomously switch between them based on its state.

Conclusion and Future Work

In this paper, we propose a two-stage framework, *Adaptive Humanoid Control* (*AHC*). The first stage distills a basic multi-behavior policy, while the second stage fine-tunes it for terrain adaptability. The resulting controller enables robust locomotion across diverse terrains and effective recovery from various types of falls. By integrating an MoE architecture, gradient projection techniques, and behavior-specific critics, our approach enhances multi-task learning efficiency and mitigates gradient conflicts. Extensive simulation and real-world experiments validate the robustness and adaptability of the proposed *AHC* policy. Future work will explore augmenting perceptual capabilities with external sensors and expanding the behavior category for even greater generalization.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No.62306242, No.62302120), the Young Elite Scientists Sponsorship Program by CAST (Grant No.2024QNRC001), the Yangfan Project of the Shanghai (Grant No.23YF11462200), and the Heilongjiang Key R&D Program of China (Grant No.GA23A915).

References

Ben, Q.; Jia, F.; Zeng, J.; Dong, J.; Lin, D.; and Pang, J. 2025. HOMIE: Humanoid Loco-Manipulation with Isomorphic Exoskeleton Cockpit. In *Robotics: Science and Systems*.

Chen, D.; Zhou, B.; Koltun, V.; and Krähenbühl, P. 2020. Learning by cheating. In *Conference on robot learning*, 66–75. PMLR.

Chen, Z.; Badrinarayanan, V.; Lee, C.-Y.; and Rabinovich, A. 2018. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, 794–803. PMLR.

Engstrom, L.; Ilyas, A.; Santurkar, S.; Tsipras, D.; Janoos, F.; Rudolph, L.; and Madry, A. 2019. Implementation matters in deep rl: A case study on ppo and trpo. In *ICLR*.

Ernst, D.; and Louette, A. 2024. Introduction to reinforcement learning. *Feuerriegel, S., Hartmann, J., Janiesch, C., and Zschech, P*, 111–126.

Escontrela, A.; Peng, X. B.; Yu, W.; Zhang, T.; Iscen, A.; Goldberg, K.; and Abbeel, P. 2022. Adversarial motion priors make good substitutes for complex reward functions. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 25–32. IEEE.

Gu, X.; Wang, Y.-J.; and Chen, J. 2024. Humanoid-Gym: Reinforcement Learning for Humanoid Robot with Zero-Shot Sim2Real Transfer. *arXiv preprint arXiv:2404.05695*.

Gu, X.; Wang, Y.-J.; Zhu, X.; Shi, C.; Guo, Y.; Liu, Y.; and Chen, J. 2024a. Advancing Humanoid Locomotion: Mastering Challenging Terrains with Denoising World Model Learning. In *RSS*.

Gu, X.; Wang, Y.-J.; Zhu, X.; Shi, C.; Guo, Y.; Liu, Y.; and Chen, J. 2024b. Advancing Humanoid Locomotion: Mastering Challenging Terrains with Denoising World Model Learning. arXiv:2408.14472.

He, J.; Zhang, C.; Jenelten, F.; Grandia, R.; BÄcher, M.; and Hutter, M. 2025a. Attention-Based Map Encoding for Learning Generalized Legged Locomotion. arXiv:2506.09588.

He, X.; Dong, R.; Chen, Z.; and Gupta, S. 2025b. Learning Getting-Up Policies for Real-World Humanoid Robots. In *Robotics: Science and Systems*.

Hessel, M.; Soyer, H.; Espeholt, L.; Czarnecki, W.; Schmitt, S.; and Van Hasselt, H. 2019. Multi-task deep reinforcement learning with popart. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3796–3803.

Hoeller, D.; Rudin, N.; Sako, D.; and Hutter, M. 2024. Anymal parkour: Learning agile navigation for quadrupedal robots. *Science Robotics*, 9(88): eadi7566.

Huang, R.; Zhu, S.; Du, Y.; and Zhao, H. 2025a. MoE-Loco: Mixture of Experts for Multitask Locomotion. arXiv:2503.08564.

Huang, T.; Ren, J.; Wang, H.; Wang, Z.; Ben, Q.; Wen, M.; Chen, X.; Li, J.; and Pang, J. 2025b. Learning Humanoid Standing-up Control across Diverse Postures. In *Robotics: Science and Systems*.

Li, J.; and Nguyen, Q. 2023. Multi-Contact MPC for Dynamic Loco-Manipulation on Humanoid Robots. In *American Control Conference (ACC)*, 1215–1220. IEEE.

Lin, S.; Qiao, G.; Tai, Y.; Li, A.; Jia, K.; and Liu, G. 2025. HWC-Loco: A Hierarchical Whole-Body Control Approach to Robust Humanoid Locomotion. *arXiv preprint arXiv:2503.00923*.

Liu, B.; Liu, X.; Jin, X.; Stone, P.; and Liu, Q. 2021. Conflict-averse gradient descent for multi-task learning. *Advances in Neural Information Processing Systems*, 34: 18878–18890.

Long, J.; Ren, J.; Shi, M.; Wang, Z.; Huang, T.; Luo, P.; and Pang, J. 2024a. Learning Humanoid Locomotion with Perceptive Internal Model. arXiv:2411.14386.

Long, J.; Ren, J.; Shi, M.; Wang, Z.; Huang, T.; Luo, P.; and Pang, J. 2024b. Learning humanoid locomotion with perceptive internal model. *arXiv preprint arXiv:2411.14386*.

Makoviychuk, V.; Wawrzyniak, L.; Guo, Y.; Lu, M.; Storey, K.; Macklin, M.; Hoeller, D.; Rudin, N.; Allshire, A.; Handa, A.; et al. 2021. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*.

Mysore, S.; Cheng, G.; Zhao, Y.; Saenko, K.; and Wu, M. 2022. Multi-critic actor learning: Teaching rl policies to act with style. In *International Conference on Learning Representations*.

Peng, X. B.; Ma, Z.; Abbeel, P.; Levine, S.; and Kanazawa, A. 2021. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4).

Radosavovic, I.; Xiao, T.; Zhang, B.; Darrell, T.; Malik, J.; and Sreenath, K. 2024. Real-world humanoid locomotion with reinforcement learning. *Science Robotics*, 9(89): eadi9579.

Ren, J.; Huang, T.; Wang, H.; Wang, Z.; Ben, Q.; Pang, J.; and Luo, P. 2025. Vb-com: Learning vision-blind composite humanoid locomotion against deficient perception. *arXiv preprint arXiv:2502.14814*.

Ross, S.; Gordon, G.; and Bagnell, D. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 627–635. JMLR Workshop and Conference Proceedings.

Rudin, N.; Hoeller, D.; Reist, P.; and Hutter, M. 2022. Learning to walk in minutes using massively parallel deep reinforcement learning. In *Conference on robot learning*, 91–100. PMLR.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Sentis, L.; and Khatib, O. 2006. A Whole-Body Control Framework for Humanoids Operating in Human Environments. In *ICRA*, 2641–2648. Orlando, FL, USA: IEEE.

Shi, J.; Liu, X.; Wang, D.; Lu, O.; Schwertfeger, S.; Sun, F.; Bai, C.; and Li, X. 2025. Adversarial Locomotion and Motion Imitation for Humanoid Policy Learning. *arXiv preprint arXiv:2504.14305*.

Sodhani, S.; Zhang, A.; and Pineau, J. 2021. Multi-Task Reinforcement Learning with Context-based Representations. arXiv:2102.06177.

Tan, R.; Li, X.; Ni, F.; Zhou, D.; Ji, Y.; and Shao, X. 2024. Versatile Jumping of Humanoid Robots via Curriculum-Assisted Reinforcement Learning. In *2024 China Automation Congress (CAC)*, 2502–2508. IEEE.

Todorov, E.; Erez, T.; and Tassa, Y. 2012. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, 5026–5033. IEEE.

Wang, D.; Bai, C.; Li, C.; Shi, J.; Ding, Y.; Zhang, C.; and Zhao, B. 2025a. Skill-Nav: Enhanced Navigation with Versatile Quadrupedal Locomotion via Waypoint Interface. *arXiv preprint arXiv:2506.21853*.

Wang, D.; Wang, X.; Liu, X.; Shi, J.; Zhao, Y.; Bai, C.; and Li, X. 2025b. MoRE: Mixture of Residual Experts for Humanoid Lifelike Gaits Learning on Complex Terrains. *arXiv preprint arXiv:2506.08840*.

Wang, H.; Wang, Z.; Ren, J.; Ben, Q.; Huang, T.; Zhang, W.; and Pang, J. 2025c. BeamDojo: Learning Agile Humanoid Locomotion on Sparse Footholds. In *Robotics: Science and Systems (RSS)*.

Xie, W.; Bai, C.; Shi, J.; Yang, J.; Ge, Y.; Zhang, W.; and Li, X. 2025. Humanoid Whole-Body Locomotion on Narrow Terrain via Dynamic Balance and Reinforcement Learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Xue, Y.; Dong, W.; Liu, M.; Zhang, W.; and Pang, J. 2025. A Unified and General Humanoid Whole-Body Controller for Fine-Grained Locomotion. In *Robotics: Science and Systems (RSS)*.

Yang, C.; Yuan, K.; Zhu, Q.; Yu, W.; and Li, Z. 2020. Multi-expert learning of adaptive legged locomotion. *Science Robotics*, 5(49).

Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836.

Zakka, K.; Tabanpour, B.; Liao, Q.; Haiderbhai, M.; Holt, S.; Luo, J. Y.; Allshire, A.; Frey, E.; Sreenath, K.; Kahrs, L. A.; et al. 2025. Mujoco playground. *arXiv preprint arXiv:2502.08844*.

Zhuang, Z.; Fu, Z.; Wang, J.; Atkeson, C.; Schwertfeger, S.; Finn, C.; and Zhao, H. 2023. Robot Parkour Learning. In *Conference on Robot Learning (CoRL)*.

Zhuang, Z.; Yao, S.; and Zhao, H. 2024. Humanoid Parkour Learning. In *8th Annual Conference on Robot Learning*.