X-Loco: Towards Generalist Humanoid Locomotion Control via Synergetic Policy Distillation

Dewei Wang[1, 2] Xinmiao Wang[2, 3] Chenyun Zhang[2] Jiyuan Shi[2] Yingnan Zhao[3]
Chenjia Bai[2 *] Xuelong Li[2 *]
[1]University of Science and Technology of China [2]Institute of Artificial Intelligence (TeleAI), China Telecom
[3]Harbin Engineering University
[*]Corresponding author Website: x-loco-humanoid.github.io

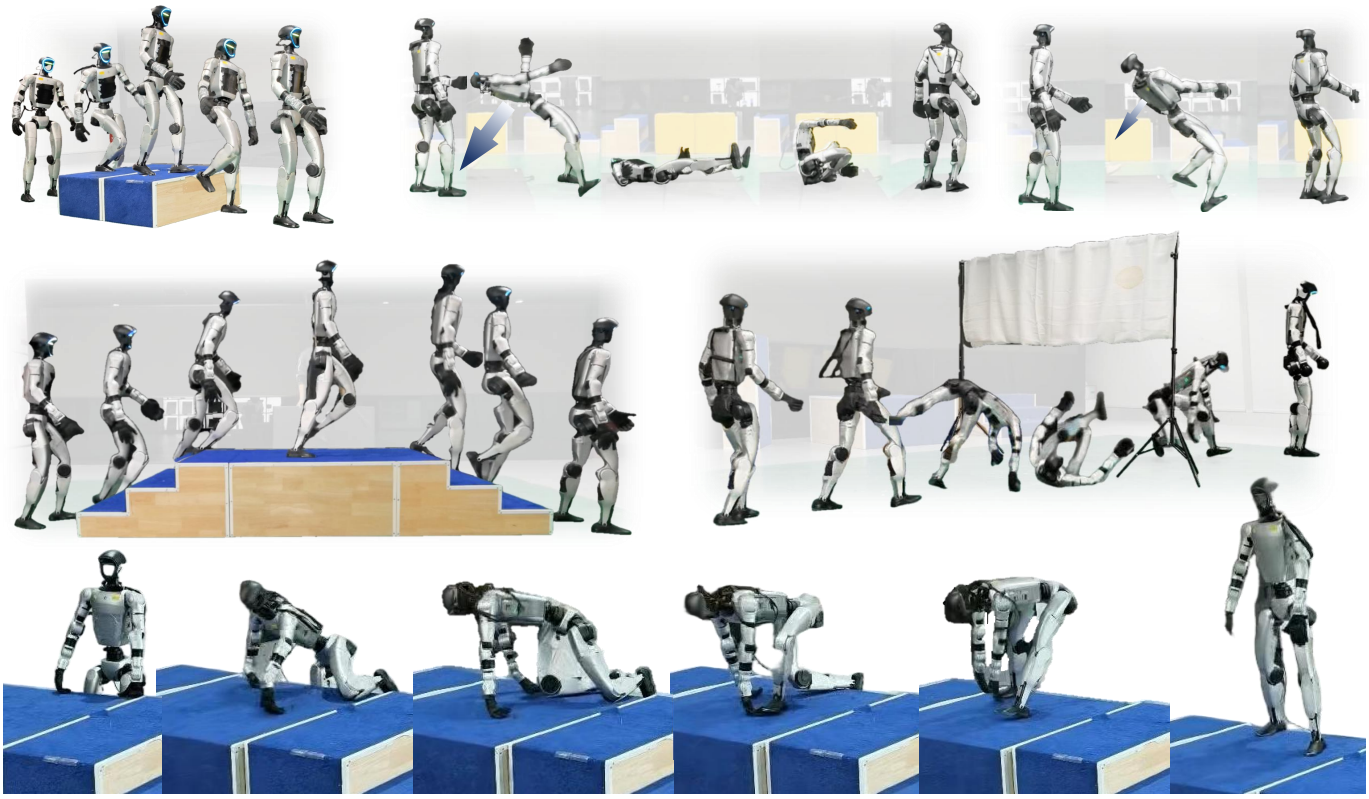


Fig. 1: X-Loco achieves vision-based generalist humanoid locomotion control. Relying solely on velocity commands without reference motions, X-Loco leverages proprioception and exteroception to traverse complex terrains, performing behaviors such as climbing up and down stairs, fall recovery, box climbing, and forward rolls.

***Abstract*—While recent advances have demonstrated strong performance in individual humanoid skills such as upright locomotion, fall recovery and whole-body coordination, learning a single policy that masters all these skills remains challenging due to the diverse dynamics and conflicting control objectives involved. To address this, we introduce X-Loco, a framework for training a vision-based generalist humanoid locomotion policy. X-Loco trains multiple oracle specialist policies and adopts a synergetic policy distillation with a case-adaptive specialist selection mechanism, which dynamically leverages multiple specialist policies to guide a vision-based student policy. This design enables the student to acquire a broad spectrum of locomotion skills, ranging from fall recovery to terrain traversal and whole-body coordination skills. To the best of our knowledge, X-Loco is the first framework to demonstrate vision-based humanoid locomotion that jointly integrates upright locomotion, whole-body coordination and fall recovery, while operating solely under velocity commands without relying on reference motions. Experimental results show that X-Loco achieves superior performance, demonstrated by tasks such as fall recovery and terrain traversal. Ablation studies further highlight that our framework effectively leverages specialist expertise and enhances learning efficiency.**

I. Introduction

Humanoid robots equipped with advanced algorithms have garnered significant attention due to their potential for anthropomorphic motion and versatile manipulation [52, 11, 39, 6, 47, 4, 19, 2]. As a fundamental capability, locomotion control has witnessed rapid progress facilitated by

reinforcement learning (RL) and high-fidelity physical simulators [49, 43, 13, 36, 33, 64, 22, 19]. Recent advancements have demonstrated remarkable results, ranging from traversing challenging terrains [54, 29, 46, 14] and executing robust fall recovery [8, 20, 56] to agilely tracking highly dynamic motions [19, 28, 60, 55, 15, 68].

Despite these advancements, a key limitation of many existing humanoid locomotion control methods lies in their predominant focus on isolated or specific categories of locomotion skills, such as exteroception-based terrain traversal [54, 29, 46, 71] or fall recovery [22, 56, 30, 48]. This functional fragmentation renders these methods incapable of handling complex scenarios, such as autonomously resuming locomotion after a fall. Furthermore, these methods fail to achieve whole-body coordination skills, which prevents the robot from traversing challenging terrains. While motion tracking [15, 61, 66, 60, 59] and teleoperation paradigms [17, 65, 25] facilitate diverse whole-body coordination behaviors, their inference relies heavily on reference motions or human intervention via specialized devices, leaving the robot unable to autonomously perceive its surroundings and perform locomotion. Consequently, these approaches lack the necessary autonomy required for self-contained and versatile humanoid locomotion controllers.

Developing a vision-based generalist locomotion controller capable of integrating locomotion skills ranging from terrain traversal to fall recovery, alongside contact-rich skills like box climbing, remains an open challenge in humanoid robotics. To realize such a versatile controller, several challenges must be addressed: First, reward engineering for humanoid locomotion remains inherently labor-intensive [22, 57], as formulating reward functions capable of eliciting diverse motor skills poses a significant challenge. Second, achieving whole-body coordination skills [59, 15] without reference motion introduces substantial complexity, as the absence of guidance leads to inefficient exploration within high-dimensional state-action spaces. Third, beyond mastering diverse skills, the controller needs to leverage visual perception to understand the environment and generate appropriate, real-time responses to conditions. Lastly, while head-mounted cameras facilitate perception, camera rendering in current simulations is limited by slow or non-independent rendering across parallel environments. Achieving fast and decoupled camera rendering is essential for the efficient training of vision-based policies.

To address these challenges, we introduce X-Loco, a framework that develops a generalist humanoid controller by distilling three privileged specialist policies, each optimized for a distinct capability: upright locomotion, fall recovery, and whole-body coordination, such as rolling and box climbing. Specifically, we implement a *Case-Adaptive Specialist Selection* (CASS) mechanism that dynamically queries the most relevant specialist policy based on the robot's state and the surrounding terrains giving the action guidance for the student policy. As acquiring multiple locomotion skills simultaneously is often hindered by the untrained policy inability to cover the specialists' optimal state-action distribution, we introduce *Specialist Annealing Rollout* (SAR), a dynamic ratio-based data collection strategy that incorporates a proportion of rollouts from the specialist policies for training. This mixing ratio decays as the distillation loss converges, shifting the focus toward the student policy's own explorations while reducing noisy data in the early stages of training. To enhance robustness, *Stochastic Fall Injection* (SFI) is implemented by applying active external disturbances during locomotion, rather than just initializing fallen states. This mechanism forces the policy to adapt to unexpected loss of balance, requiring the emergence of a transition between locomotion and emergency fall recovery.

In summary, by distilling specialist policy expertise into a generalist policy, X-Loco expands the capabilities of existing humanoid locomotion controllers. Our primary contributions are summarized as follows:

- We develop X-Loco, which trains three locomotion specialist policies sharing a unified action space and integrates their diverse motor skills into a generalist policy.
- We introduce a synergetic distillation paradigm to consolidate expertise from multiple specialist policies into a single generalist policy, effectively mitigating interference between diverse locomotion skills.
- We introduce SAR to ensure efficient expert knowledge internalization through adaptive rollout mixing, and SFI to expose the policy to failure regimes, thereby enabling transitions between locomotion and fall recovery.
- We deploy X-Loco on the Unitree G1 robot, demonstrating its superior performance and stability across diverse terrains and challenging scenarios, thereby validating the framework's robustness and sim-to-real transferability.

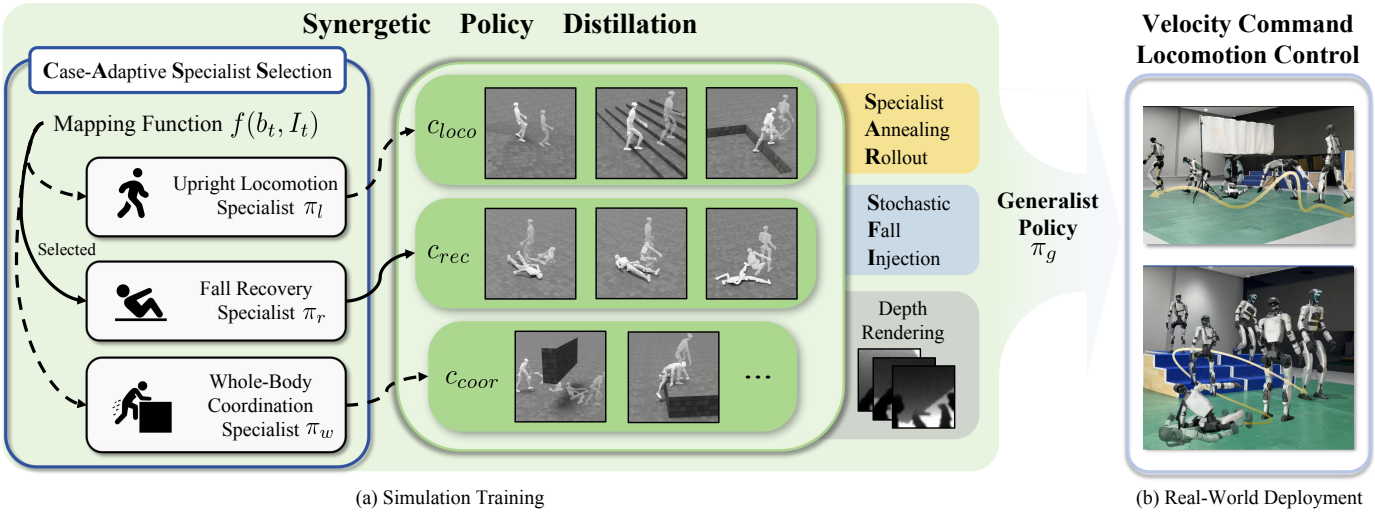
II. Related Work

A. Learning-based Humanoid Control

1) Upright Locomotion: Humanoid locomotion has evolved from blind locomotion on uneven terrains [13, 40, 67, 13], visual locomotion across complex terrains [54, 46, 29] and multiple locomotion skill learning [71, 57]. By decoupling the control of the upper and lower limbs [5, 45, 57], humanoid robots can achieve complex loco-manipulation tasks that require coordination between the entire body. However, the aforementioned methods fall short of addressing complex tasks such as fall recovery and high-difficulty whole-body coordinated movements.

2) Fall Recovery: Falling is an inevitable occurrence for legged robots. RL provides a robust and generalizable framework for fall recovery [24, 22, 8]. FIRM [56] has explored the synergy between fall-safety and fall-recovery and AHC [69] utilizes multi-task RL to simultaneously achieve both autonomous fall recovery and stable walking. Despite these advancements, integrating fall recovery with a visual locomotion controller remains an open challenge.

3) Motion Tracking: Leveraging motion retargeting [3, 63, 18] and motion imitation [37, 51, 31], humanoid robots can perform highly dynamic behaviors [19, 55, 68, 28] and universal motion tracking [15, 66, 60, 9] given reference motions.



(a) Simulation Training (b) Real-World Deployment

Fig. 2: **Overview of X-Loco**. (a) X-Loco integrates the capabilities of three specialist policies into a vision-based generalist policy via Synergetic Policy Distillation. (b) X-Loco can perform diverse locomotion skills in the real world.

Exteroceptive information is integrated into a motion tracking framework to enable complex loco-manipulation [61] and sensorimotor locomotion [1]. Our method further eliminates the dependency on reference motions during inference, enabling the robot to autonomously perform whole-body coordination movements based on exteroceptive perception.

B. Multiple Locomotion Skills Learning

While developing a single policy for multiple skills is often hindered by gradient interference and task objective trade-offs [26, 62]. Mixture-of-Experts (MoE) [7, 23] has been shown to mitigate gradient conflict in multi-skill learning and effectively facilitate the acquisition of diverse locomotion gaits [54, 21]. Methods such as MELA [58] and MTAC [44] utilize a hierarchical structure to decouple individual motor skills, which are subsequently coordinated by a high-level module. Locomotion skills with minor morphological variations can be acquired by one-stage framework leveraging reward function design [57, 35] and curriculum learning [10, 53].

In legged locomotion, policy distillation is widely utilized for privileged information learning [10, 17, 9] and sim-to-real transfer [61, 27]. Policy distillation enables dynamic parkour behaviors by training specialized teachers for different terrains and distilling their expertise into a single policy [70, 71]. In contrast to existing works, our framework further integrates whole-body coordination and fall recovery into a generalist policy, autonomously executed based solely on exteroceptive perception and velocity commands.

III. Problem Formulation

We formulate the humanoid locomotion control as a Markov Decision Process (MDP), defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma)$ where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, respectively. $\mathcal{P}(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t)$ represents the state transition probability, $R(\boldsymbol{s}_t, \boldsymbol{a}_t) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor. We employ proximal policy optimization (PPO) [43] and Generalized Advantage Estimation (GAE) [42] to train specialist policies, each optimized to maximize the expected discounted cumulative return $J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(\boldsymbol{s}_t, \boldsymbol{a}_t)\right]$.

1) State Space: We partition the state space $\mathcal{S}$ into three components: the privileged state $\boldsymbol{s}_t^{\text{priv}}$, which contains information available only in simulation, proprioceptive information $\boldsymbol{s}_t^{\text{prop}}$ and exteroceptive information which in our setting consists of depth images $\boldsymbol{d}_t \in \mathbb{R}^{64\times64}$. The proprioceptive information $\boldsymbol{s}_t^{\text{prop}}$ primarily comprises:

$$\boldsymbol{s}_t^{\text{prop}} = [\boldsymbol{\omega}_t, \boldsymbol{g}_t, \boldsymbol{q}_t, \dot{\boldsymbol{q}}_t, \boldsymbol{a}_{t-1}] \in \mathbb{R}^{78}, \tag{1}$$

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, $\boldsymbol{q}_t \in \mathbb{R}^{23}$ and $\dot{\boldsymbol{q}}_t \in \mathbb{R}^{23}$ represent joint position and joint velocity, and $\boldsymbol{a}_{t-1} \in \mathbb{R}^{23}$ is the last action. The privileged state $\boldsymbol{s}_t^{\text{priv}}$ is defined as:

$$\boldsymbol{s}_t^{\text{priv}} = [\boldsymbol{v}_t, b_t, \boldsymbol{e}_t, \boldsymbol{f}_t, \boldsymbol{h}_t, \boldsymbol{m}_t], \tag{2}$$

where $\boldsymbol{v}_t \in \mathbb{R}^3$ is the base linear velocity, b_t denotes the head height, and $\boldsymbol{e}_t \in \mathbb{R}^6$, $\boldsymbol{f}_t \in \mathbb{R}^6$ represent the positions and velocities of the hands and feet relative to the base frame, respectively. $\boldsymbol{h}_t \in \mathbb{R}^{143}$ denotes the local terrain height samples, while $\boldsymbol{m}_t \in \mathbb{R}^{52}$ indicates the motion command.

2) Action Space: To ensure knowledge transfer during distillation, all policies share a consistent action space $\mathcal{A} \in \mathbb{R}^{23}$ and employ the same Proportional Derivative (PD) parameters to map the action to joint target positions: $\boldsymbol{q}_t^{\text{target}} = \boldsymbol{q}^{\text{default}} + \boldsymbol{\alpha} \cdot \boldsymbol{a}_t$, where $\boldsymbol{q}^{\text{default}}$ represents the pre-defined default pose and $\boldsymbol{\alpha}$ is a scaling factor following Liao et al. [28]. The joint torques $\boldsymbol{\tau}_t$ are then computed as:

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

where K_p and K_d denote the stiffness and damping coefficients, respectively.

IV. METHOD

In this section, we introduce the pipeline of the X-Loco framework as shown in Fig 2. We first train three specialist policies: upright locomotion π_l, fall recovery π_r, and whole-body coordination π_w. Subsequently, we propose a synergetic policy distillation to effectively consolidate these specialized capabilities into a single generalist policy π_g.

A. Specialist Policies Training

All specialists share a unified action space and consistent DoFs. While these specialists are undeployable in real world due to their dependency on privileged state, they achieve peak performance in their respective domains.

1) Upright Locomotion Specialist: The upright locomotion specialist π_l is designed to establish the robot's basic mobility, enabling navigation of diverse terrains, such as stairs, pits, and gaps, while following velocity commands $\boldsymbol{c}_t = [v_x, v_y, \omega_z]^\top$. We employ two encoders to process observations, alongside an actor network to predict actions. Specifically, the policy incorporates a history encoder to compress historical states comprising $\{\boldsymbol{s}_i^{\text{prop}}, \boldsymbol{v}_i, \boldsymbol{e}_i, \boldsymbol{f}_i\}_{i=t-9}^{t}$ over ten consecutive time steps into a latent representation. Simultaneously, an elevation map encoder processes $\boldsymbol{h}_t$ to provide a compact geometric understanding of the surrounding terrain. To ensure behavioral naturalness and enhance exploration efficiency, we incorporate the Adversarial Motion Prior (AMP) [38, 50] as a style reward to guide the policy. Consequently, π_l learns to track velocity commands with a natural gait across diverse terrains, providing upright locomotion guidance for the student policy.

2) Fall Recovery Specialist: The fall recovery specialist π_r is optimized for robot posture restoration from diverse fallen situations. Given that the objective of π_r is postural stabilization rather than velocity-tracking, the policy is designed to be agnostic to terrain information, i.e. $\boldsymbol{h}_t$. We adopt an architecture comprising only a history encoder compressing historical state similar as π_l and an actor network for π_r. To ensure the policy can recover from arbitrary failure cases, the robot is initialized in both supine and prone postures during training, with large joint position noise added to the initial states to simulate diverse falling conditions. Furthermore, an AMP-based style reward is integrated into the training of π_r for avoiding jerk motions and guide the policy's exploration. As a result, π_r is capable of executing recovery maneuvers, enabling the robot to regain its footing from any fallen state.

3) Whole-Body Coordination Specialist: While existing locomotion policies often lack whole-body coordination ability, we develop the whole-body coordination specialist π_w focusing on these skills following a motion tracking paradigm. The π_w consists of an elevation map encoder and an actor network. The former processes $\boldsymbol{h}_t$ into a latent representation, which together with $\boldsymbol{s}_t^{\text{prop}}$ and $\boldsymbol{m}_t$ is fed into the actor network to predict actions. We employ the elevation map encoder for local terrain perception, thereby enhancing the policy's generalization capability when tracking motions that involve terrain interaction. π_w is specialized in tracking motions that necessitate whole-body coordination motion like box climbing and rolling, rather than aiming for universal motion imitation. The subsequent distillation phase transfers these skills into a generalist policy π_g and eliminate its dependence on reference motions during inference.

B. Synergetic Policy Distillation

In the distillation process, we employ a student policy using MoE architecture that maps $\boldsymbol{s}_t^{\text{prop}}$, $\boldsymbol{c}_t$ and $\boldsymbol{d}_t$ to target joint positions, acquiring the capabilities from various specialist policies. To consolidate skills from these specialists into a generalist policy, our synergetic policy distillation employs CASS to dynamically switch specialists based on the robot's state and terrain, while SFI and SAR collectively enhance robustness and training efficiency.

1) Case-Adaptive Specialist Selection: CASS defines three cases according to the robot state and local terrain: recovery, upright locomotion, and coordinated maneuvers. Each case is associated with a specific specialist. Let $\mathcal{C} = \{(c_{\text{rec}}, \pi_r), (c_{\text{loco}}, \pi_l), (c_{\text{coor}}, \pi_w)\}$ denotes the set of all cases and their corresponding specialist. The specific case assigned to the robot is determined by the mapping function $f(b_t, I_t)$, which evaluates the robot's head height b_t and the current terrain context I_t and outputs a specific case index from $\mathcal{C}$. A detailed definition for the mapping function is provided in Appendix C. During distillation, the target action a_t^* is selected as:

$$a_t^* = \sum_{(i,\pi_i)\in\mathcal{C}} \mathbb{I}(i = f(b_t, I_t)) \cdot \pi_i(\boldsymbol{s}_{i,t}), \tag{4}$$

where $\boldsymbol{s}_{i,t}$ represents the input of the specialist policy and $\mathbb{I}$ is the indicator function. The termination criteria are consistent with the selected specialist policy to ensure that the collected trajectories remain within state-space manifold explored during specialist training. Notably, each specialist leverages clean privileged observations $\boldsymbol{s}_{i,t}$ for maximum performance, the student policy is trained using a noised non-privileged observation $o_{u,t}$. The distillation objective minimizes the mean squared error between the student's action and the selected target action:

$$\mathcal{L}_{\text{distill}} = \mathbb{E}[\|\pi_g(o_{u,t}) - a_t^*\|_2^2]. \tag{5}$$

Within the case c_{coor}, CASS adaptively modulates the motion commands fed into π_w, thereby providing the student policy with appropriate whole-body coordination skill guidance across different scenarios. Consequently, CASS provides the student policy with appropriate target action, enabling the generalist humanoid control learning including whole-body coordination, terrain traversal, and fall recovery based on proprioception and depth without requiring reference motions.

2) Specialist Annealing Rollout: Only with the CASS mechanism, training a single policy to master multiple specialist domains simultaneously is challenging. This difficulty stems from two primary issues: First, an untrained policy frequently generates trajectories that are out-of-distribution (OOD) relative to the specialists, which are difficult to filter via early termination. Second, contact-rich and whole-body

coordination skills are hard to acquire because errors in the early stages of a maneuver prevent the agent from experiencing subsequent parts of the skill. To mitigate these challenges, we introduce SAR, which samples actions applied for the environment during data collection by switching between the student and specialist policies according to a mixing ratio ρ:

$$a_{env} = b_t a_t^* + (1 - b_t)\pi_g(o_{g,t}), \tag{6}$$

where $b_t \sim \text{Bernoulli}(\rho)$ is a binary random variable governed by ρ. This mixed rollout strategy reduces the occurrence of low-quality state manifolds, allowing the student policy to observe and learn from the successful behaviors of complex motion sequences before it can execute them independently. The mixing ratio ρ follows an annealing schedule tied to the training progress. Specifically, we decrease ρ when the distillation loss $\mathcal{L}_{distill}$ falls below a predefined threshold ϵ:

$$\rho \leftarrow \max(0, \rho - \Delta\rho \cdot \mathbb{I}(\mathcal{L}_{distill} < \epsilon)) \tag{7}$$

By gradually shifting from expert-driven rollouts to autonomous exploration, SAR provides a scaffold for distillation, ensuring the student policy transitions to self-exploration only after attaining a foundational mastery of the specialist's knowledge.

3) Stochastic Fall Injection: Instead of solely initialize a subset of environments in fallen states, we propose SFI which introduces random external forces while the robot is walk or climbing, forcing the robot to enter the c_{rec} regime from normal states. To ensure realism, these external force injections are not entirely stochastic, they are conditioned on specific vulnerable scenarios, such as high-speed turns or rolling motions, to simulate unpredictable real-world accidents. To prevent the student policy from being compromised by corrupted data when external forces drive the robot into OOD states relative to π_r, we implement an additional termination criterion when SFI is triggered. We maintain a buffer of the robot's head height H over consecutive timesteps, the episode is terminated if the variance of the buffer falls below a predefined threshold indicating that the robot has become stuck during the stand-up process:

$$\mathbb{I}\left(f(b_t, I_t) = c_{rec}\right) \cdot \mathbb{I}\left(\text{Var}(H_{t-k:t}) < \delta\right), \tag{8}$$

where δ is a stability threshold. This criterion ensures that the distillation process remains focused on recoverable states by pruning trajectories where the π_r fails to progress.

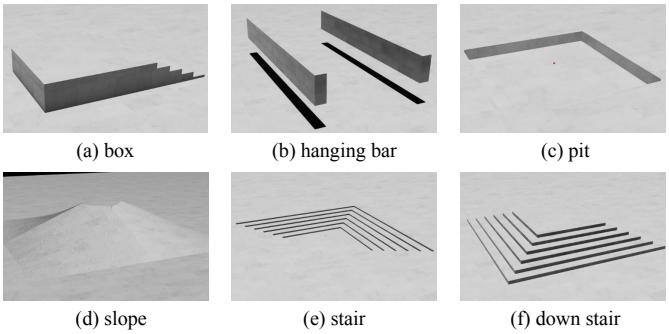
C. Training Details

This section details implementation techniques used to optimize the training process. Refer to Appendix A for the complete configurations regarding the simulation environment and hardware setup, as well as reward formulations and domain randomization.

1) Expert Ratio Scheduling: We implement a hysteresis-based annealing mechanism for the expert ratio ρ. Specifically, ρ decays by a step size of $1\text{e}{-4}$ per iteration only when the MSE loss falls below a lower threshold $\tau_{low} = 0.005$. Conversely, if the loss exceeds an upper threshold $\tau_{high} = 0.010$, the decay process is suspended until the performance recovers (i.e., loss $< \tau_{low}$). This adaptive scheduling dynamically modulates the intensity of expert assistance based on the student's real-time proficiency. The introduction of the hysteresis zone enhances the stability of ρ updates and prevents premature reduction of expert support.

2) Stabilizing Whole-Body Specialist Distillation: During the pre-training of π_w, we incorporate elevation maps into the observation space to facilitate the perception of local terrain geometry. To mitigate overfitting to specific terrain configurations, we randomize the robot's start position with x-y offsets sampled from $[-0.15\text{m}, 0.15\text{m}]$ on box terrains. For the distillation phase, we preserve data quality by adopting the specialist's termination criteria to exclude failed samples.

3) Camera Rendering: Current sensor rendering in simulations [33, 36] suffers from low throughput and a lack of visual isolation, where robots inadvertently perceive instances from parallel environments. Furthermore, acceleration frameworks such as NVIDIA Warp [32] often struggle to concurrently handle both dynamic and static meshes. We implement a high-performance depth rendering pipeline utilizing NVIDIA Warp to execute parallelized ray-casting. Our approach decouples the environment into static meshes representing the terrain and dynamic meshes representing the robot mesh. To optimize computational throughput, each agent's ray-caster selectively queries the global static mesh and its own local dynamic mesh.



(a) box (b) hanging bar (c) pit

(d) slope (e) stair (f) down stair

Fig. 3: Terrains used for training and evaluation.

V. Experiments

A. Experiment Setup

We perform training and evaluation in the IsaacLab [36] simulator. The trained policy is deployed on a Unitree G1 humanoid robot, operating at 50Hz predicting target joint positions. These targets are subsequently tracked by a 500Hz PD controller, which translates them into torques to drive the motors.

In the simulation evaluation, we categorize the test task into three categories: **Upright Locomotion**, **Whole-Body Coordination** (**WBC**), and **Recovery**. The full suite of evaluation terrains is visualized in Fig. 3 and the detailed physical specifications for each terrain type are summarized in Table II.

1) Baselines: To validate the locomotion performance of the proposed framework, we compare our framework with the following baselines:

TABLE I: Quantitative comparison of X-Loco against baselines and specialist policies. Bold numbers indicates the best performance other than the specialists, and - denotes that the method failed to complete the corresponding task.

Method	**Locomotion**							**Whole-Body Coordination**			**Recovery**
	Slope		Pit		Stairs			Hanging Bar	Box		Flat
	R_{succ}	D_{trav}	R_{succ}	D_{trav}	R_{succ}	D_{trav}	$\bar{R}_{\text{succ}}$	R_{succ}	R_{succ}	$\bar{R}_{\text{succ}}$	R_{succ}
BeyondMimic [28]	-	-	-	-	-	-	-	**1.000** ± .000	**0.916** ± .008	**0.958** ± .004	-
MoRE [54]	**0.992** ± .008	7.863 ± .018	0.844 ± .009	6.833 ± .114	0.926 ± .009	7.387 ± .013	0.921 ± .009	-	-	-	-
PPO [43]	0.823 ± .021	6.674 ± .152	0.793 ± .018	6.396 ± .184	0.781 ± .025	6.565 ± .141	0.799 ± .021	-	-	-	-
AHC [69]	0.968 ± .005	7.871 ± .022	0.403 ± .031	3.179 ± .254	0.278 ± .042	2.524 ± .311	0.550 ± .026	-	-	-	**1.000** ± .000
Locomotion Specialist	0.995 ± .002	7.989 ± .011	0.984 ± .010	7.914 ± .135	0.991 ± .011	7.963 ± .004	0.990 ± .008	-	-	-	-
Whole-Body Specialist	-	-	-	-	-	-	-	1.000 ± .000	1.000 ± .000	1.000 ± .000	-
Recovery Specialist	-	-	-	-	-	-	-	-	-	-	1.000 ± .000
Ours (X-Loco)	0.982 ± .010	**7.984** ± .033	**0.878** ± .015	**7.592** ± .084	**0.958** ± .007	**7.853** ± .011	**0.939** ± .011	0.873 ± .014	0.868 ± .018	0.871 ± .016	**1.000** ± .000

TABLE II: Terrain Configurations for Evaluation

Skill	Obstacle Properties	Ranges	Unit
Locomotion	slope incline	$[15, 20]$	°
	pit obstacle height	$[0.30, 0.40]$	m
	stair step height	$[0.10, 0.15]$	m
Whole-Body Coordination	obstacle vertical clearance	$[0.87, 0.95]$	m
	obstacle height	$[0.50, 0.65]$	m
Recovery	flat ground	-	-

- **BeyondMimic [28]:** A baseline focuses on whole-body coordination motion tracking relying solely on proprioception. We compare against it to highlight the necessity of exteroception for interacting with environments.
- **MoRE [54]:** This baseline employs a two-stage pipeline training a vision-based policy to achieve robust locomotion across complex terrains.
- **AHC [69]:** This baseline implements an adaptive blind control policy that unifies locomotion with autonomous fall recovery behaviors.
- **PPO [43]:** The baseline trains a policy using PPO with depth inputs to traverse across all terrains.
- **Our Specialists:** We evaluate the individual specialist policies (π_l, π_r and π_w) separately.

To assess the specific contributions of our proposed components, we also evaluate several ablation baselines:

- **Ours w/o CASS:** An ablation to evaluate whether CASS facilitates the acquisition diverse locomotion skills.
- **Ours w/o SFI:** An ablation to evaluate whether SFI helps to enhance the policy's robustness.
- **Ours w/o SAR:** An ablation to analyze SAR's contribution to the policy distillation process.
- **Ours w/o MoE:** Ablation of policy network with the MoE replaced by a Multi-Layer Perceptron (MLP).
- **MoE-N**: Variants of our framework with $N \in \{2, 3\}$ experts to investigate the sensitivity of performance to N, where **MoE-2** serves as our adopted architecture.

2) Metrics: We quantify performance using two primary metrics:

- **Success Rate (R_{succ}):** The percentage of episodes where the task objective is fully met. Success is defined as: traversing the entire $8\,\text{m}$ track without falling for the **Upright Locomotion** and **WBC** task; regaining a standing posture and balancing for $5\,\text{s}$ for the **Recovery** task.
- **Traversal Distance (D_{trav}):** The average distance traveled along the x-axis prior to failure or timeout is assessed in the **Upright Locomotion** tasks.

B. Simulation Results

1) Comparative Analysis: We conduct a comprehensive quantitative comparison between our proposed X-Loco, specialized specialists, and several baselines. The results are summarized in Table I.

a) Specialist vs. Baselines: As expected, the specialist policies consistently achieve the highest performance in their respective domains. For instance, the locomotion specialist achieves a near-perfect average success rate $\bar{R}_{\text{succ}}$ of 0.990 and the whole-body specialist achieves perfect performance with $\bar{R}_{\text{succ}}$ at 1.000. The superior performance over the baseline in the box climbing task stems from to the utilization of the information in $\boldsymbol{h}_t$. This results is primarily attributed to the use of privileged information and a focused training objective on a single category of tasks which allows these policies to establish the performance upper bound for each skill.

b) X-Loco vs. Baselines: Our generalist policy demonstrates significant versatility compared to the baselines. While BeyondMimic excels in **WBC** tasks with an average $\bar{R}_{\text{succ}}$ of 0.958 but lacks locomotion and recovery capabilities, and MoRE and AHC show competitive results in locomotion but fail completely in **WBC** task, X-Loco successfully masters all three tasks. In the **Upright Locomotion** task, X-Loco achieves an average success rate of 0.939. This performance outperforms the PPO baseline at 0.799 and AHC at 0.550 by a large margin, while remaining highly competitive with MoRE. In the **WBC** task, X-Loco maintains a success rate of 0.871 without relying on reference motions, whereas other baselines lacking such references fail to execute these tasks.

c) Specialist vs. X-Loco: The performance gap between X-Loco and the specialist policies is remarkably narrow. In the **Upright Locomotion** task, X-Loco recovers approximately 94.8% of the specialist's success rate. In the **Recovery** task, X-Loco matches the specialist with a perfect 1.000 success rate. While X-Loco exhibits the most pronounced performance degradation in the **WBC** tasks, this underscores the difficulty of mastering vision-based whole-body coordination skills without reliance on reference motions. These results indicate

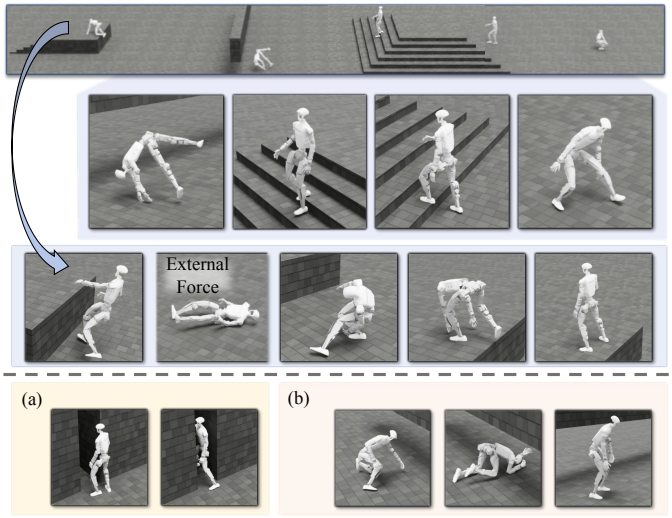


Fig. 4: **Top**: Testing the generalist policy on hybrid, challenging terrains. **Bottom**: Extensibility of the X-Loco to include vision-guided (a) lateral sidling and (b) kneeling crawling.

TABLE III: Quantitative results of ablation analysis on CASS and MoE architectures.

Method	Locomotion $\bar{R}_{\text{succ}}$	Whole-Body Coordination $\bar{R}_{\text{succ}}$	Recovery R_{succ}	Average $\bar{R}_{\text{succ}}$
Ours w/o CASS	0.903	0.446	**1.000**	0.783
Ours w/o MoE	0.692	0.561	0.996	0.749
MoE-3	0.628	0.709	**1.000**	0.779
MoE-2 (Ours)	**0.939**	**0.845**	**1.000**	**0.928**

that X-Loco can effectively distill and integrate diverse expert knowledge into a single policy without significant performance degradation, despite the inherent challenges of multi-task learning and potential gradient interference between different skills.

2) Ablation Analysis: We conduct a comprehensive ablation study on CASS, MoE architecture, SAR and SFI.

a) Effectiveness of CASS: To evaluate the contribution of CASS, we compare our method against the variant **Ours w/o CASS**. As reported in Table III, the absence of CASS leads to a drastic performance collapse in both **WBC** and **Upright Locomotion** tasks. The primary driver of this degradation is that without the selection mechanism, the distillation process lacks critical training samples representing the transition phases between heterogeneous motor patterns. Specifically, the student policy is not exposed to the specific state-action pairs required to bridge disparate skills. In summary, CASS is essential for providing the guidance to synthesize specialized specialist into a generalist policy, ensuring the emergent generalist policy maintains high fidelity to specialist capabilities while mastering the connective dynamics between them.

b) Ablations on MoE Architecture: We evaluate **Our w/o MoE** and **MoE-3** across all test tasks and report the average success rate. As shown in Table III, the **Ours w/o MoE** variant struggles to maintain high success rates across all tasks, with performance dropping significantly in **WBC**

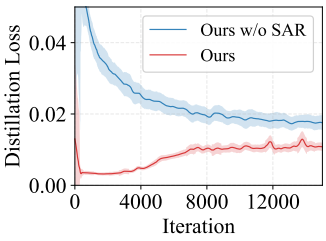


Fig. 5: Distillation loss curves of ablation on SAR.

Method	Recovery R_{succ}
Ours w/o SFI	0.912
Ours	**0.958**

TABLE IV: Recovery success rate of ablation on SFI with large external disturbance

tasks. This confirms that the MLP backbone lacks the capacity to reconcile state-action distributions from multiple specialists into a shared parameter space. While the **MoE-2** significantly recovers performance, further increasing the number of experts (**MoE-3**) leads to a regression, particularly in **Upright Locomotion** tasks. We attribute this to the fact that an excessive number of experts can lead to sparse updates and sub-optimal expert utilization during the distillation process.

c) Ablations on SAR & SFI: As shown in Fig. 5, our method demonstrates a clear advantage in distillation loss over the **Ours w/o SAR** baseline. In the early training phase, the loss decreases rapidly as the policy focuses on high-quality samples from the specialists. As ρ decays, the student policy begins to utilize its own rollout data including non-optimal or failed trajectories to learn to adapt complex situations, leading to a rise in loss. Ultimately, X-Loco achieves faster convergence to a lower terminal loss. This confirms that SAR effectively bridges specialist guidance and self-exploration, ensuring more efficient policy distillation. To evaluate the role of SFI, we assess the policy's performance in **Upright Locomotion** and **WBC** tasks under large external disturbances that lead to falls. We measure the success rate after falling to quantify the recovery capability. As shown in Table IV, the variant with SFI achieves a higher success rate, demonstrating that SFI significantly enhances the policy's ability to regain balance in unpredictable environments.

C. More Analysis

1) Hybrid Terrain: We further evaluated the robustness and skill integration of the generalist policy by constructing a challenging hybrid terrain sequence that requires continuous navigation across diverse obstacles. In this experiment, the robot is initialized in a fallen state and must sequentially perform fall recovery, climb up and down stairs, roll under a hanging bar, and finally climb a box. This is a long-horizon task using depth information, as shown in Fig 4. It can even recover after being pushed off a high platform and subsequently resume the climbing task. This successful traversal across such a challenging terrain proves that our framework effectively consolidates diverse skills into a single policy, enabling the robot to execute multifaceted movements while maintaining inherent adaptability.

2) Framework Extensibility: X-Loco enables the capability expansion of the generalist policy by expanding motions learned by the whole-body coordination specialist and sub-

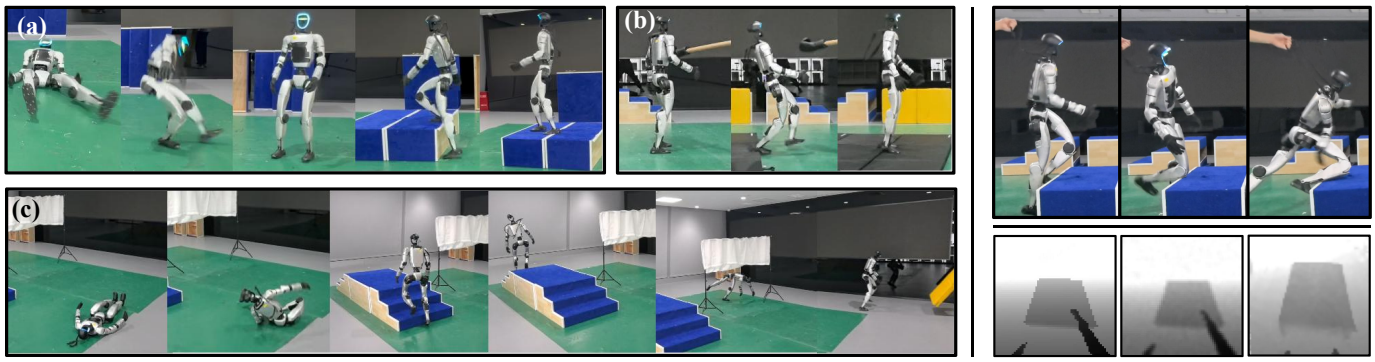


Fig. 6: **Left**: (a) fall recovery and platform traversal; (b) resilience to external disturbances; (c) a continuous sequence of recovery, stair climbing, and rolling under an overhead bar. **Right**: (Top) a failure case where the robot trips due to lack of camera randomization. (Bottom, from left to right) original depth in simulation, noisy depth, and processed real-world depth.

sequently adjusting CASS and terrains during the distillation process. Fig. 4 illustrates two examples of capability expansion: lateral movement through narrow gaps and crawling under low-overhanging obstacles. The results demonstrate that the distilled generalist policy π_g successfully mastered these skills: when encountering narrow gaps, the robot autonomously transitions to a sidling posture to traverse the terrain; when faced with low-clearance obstacles, the robot spontaneously adopts a crawling gait to pass under the obstacles. This capability underscores the versatility and extensibility of our framework. By simply enriching the specialist's training set, the generalist policy can effectively integrate new whole-body coordination skills, proving that X-Loco is a scalable solution for humanoid locomotion control.

D. Real-World Deployment

a) Performance on Hybird Terrains: To validate the robustness and versatility of X-Loco in the real world, we deployed the generalist policy on a Unitree G1 robot. Fig. 1 illustrates the performance of X-Loco across diverse challenging terrains, including a 90cm-high hanging bar and a 60cm-tall box. In a series of comprehensive tests, the robot successfully manifested a seamless transition between the three core locomotion abilities. Fig. 6 demonstrates X-Loco's performance across various composite terrains under disturbances. The policy successfully executes a series of continuous maneuvers, including recovering from a fall, traversing high platforms, climbing stairs, and performing a rolling motion to pass under a hanging bar. Furthermore, the experimental results confirm the policy's exceptional resilience. When subjected to severe external disturbances or manual pushes that result in a fall, the policy autonomously regains an upright posture and resumes locomotion. The experiment results demonstrate that the capabilities integrated via our distillation pipeline generalize effectively to physical hardware.

b) Depth Sim-to-Real Analysis: Transitioning the policy from simulation to the real world reveals a large depth sim-to-real gap, which significantly degrades the policy's performance as shown in Fig. 6. The policy suffers from perception-action mismatches caused by depth disparities, resulting in collisions with obstacles followed by falls or continuous foot scuffing against the ground during locomotion. To bridge the depth sim-to-real gap, we implement a multi-stage pipeline that synthesizes realistic depth data by injecting additive Gaussian noise and blurring to emulate real-world sensor. Furthermore, we employ domain randomization over the camera's extrinsic and intrinsic parameters, including random perturbations of its orientation, position, and field of view (FOV). For real-world depth images, we apply hole-filling and Gaussian filtering to mitigate sensor noise and data sparsity. As illustrated in Fig. 6, this preprocessing ensures consistent image alignment between simulation and reality, enabling the policy to achieve successful sim-to-real transfer.

VI. Conclusion

We present X-Loco, a framework that enables a single vision-based policy to achieve generalist humanoid control via velocity commands. By employing synergetic policy distillation, X-Loco consolidates three specialist policies into a unified agent. SAR and SFI are introduced to bridge the gap between expert guidance and autonomous exploration, and to enhance the policy's recovery resilience, respectively. Experimental results demonstrate that X-Loco achieves successful integration of diverse locomotion skills with performance comparable to that of the specialists, offering a scalable solution for generalist humanoid locomotion. Real-world deployments further confirm its capability to manage complex environments, pushing the boundaries of humanoid locomotion.

VII. Limitations and Future Work

Despite its advancements, X-Loco has limitations. Its performance is primarily constrained by a limited sensory horizon due to the narrow-FOV camera, and the difficulty of perfectly modeling real-world sensor noise. Additionally, since the distillation relies on behavior cloning, the policy is inherently bounded by the specialists' performance, making X-Loco less effective in edge cases lacking expert coverage.

Future work will follow two directions. First, we will integrate multi-modal sensing (e.g., RGB-D and LiDAR) to broaden the perceptual field and rectify misjudgments through

cross-modal calibration. Second, we aim to develop a hybrid learning framework that combines distillation with RL-based fine-tuning, allowing the policy to explore and generalize beyond specialist demonstrations.

References

[1] Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. Visual imitation enables contextual humanoid control. *arXiv preprint arXiv:2505.03729*, 2025.

[2] Hongjun An, Wenhan Hu, Sida Huang, Siqi Huang, Ruanjun Li, Yuanzhi Liang, Jiawei Shao, Yiliang Song, Zihan Wang, Cheng Yuan, et al. Ai flow: Perspectives, scenarios, and approaches. *Vicinagearth*, 3(1):1, 2026.

[3] Joao Pedro Araujo, Yanjie Ze, Pei Xu, Jiajun Wu, and C Karen Liu. Retargeting matters: General motion retargeting for humanoid motion tracking. *arXiv preprint arXiv:2510.02252*, 2025.

[4] Tamim Asfour, Pedram Azad, Nikolaus Vahrenkamp, Kristian Regenstein, Alexander Bierbaum, Kai Welke, Joachim Schroeder, and Ruediger Dillmann. Toward humanoid manipulation in human-centred environments. *Robotics and Autonomous Systems*, 56(1):54–65, 2008.

[5] Qingwei Ben, Feiyu Jia, Jia Zeng, Junting Dong, Dahua Lin, and Jiangmiao Pang. Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit. *arXiv preprint arXiv:2502.13013*, 2025.

[6] Aude Billard and Danica Kragic. Trends and challenges in robot manipulation. *Science*, 364(6446):eaat8414, 2019.

[7] Onur Celik, Aleksandar Taranovic, and Gerhard Neumann. Acquiring diverse skills using curriculum reinforcement learning with mixture of experts. In *International Conference on Machine Learning*, pages 5907–5933. PMLR, 2024.

[8] Penghui Chen, Yushi Wang, Changsheng Luo, Wenhan Cai, and Mingguo Zhao. Hifar: Multi-stage curriculum learning for high-dynamics humanoid fall recovery. *arXiv preprint arXiv:2502.20061*, 2025.

[9] Zixuan Chen, Mazeyu Ji, Xuxin Cheng, Xuanbin Peng, Xue Bin Peng, and Xiaolong Wang. Gmt: General motion tracking for humanoid whole-body control. *arXiv:2506.14770*, 2025.

[10] Xuxin Cheng, Kexin Shi, Ananye Agarwal, and Deepak Pathak. Extreme parkour with legged robots. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11443–11450. IEEE, 2024.

[11] Kourosh Darvish, Luigi Penco, Joao Ramos, Rafael Cisneros, Jerry Pratt, Eiichi Yoshida, Serena Ivaldi, and Daniele Pucci. Teleoperation of humanoid robots: A survey. *IEEE Transactions on Robotics*, 39(3):1706–1727, 2023.

[12] Alejandro Escontrela, Xue Bin Peng, Wenhao Yu, Tingnan Zhang, Atil Iscen, Ken Goldberg, and Pieter Abbeel. Adversarial motion priors make good substitutes for complex reward functions. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 25–32. IEEE, 2022.

[13] Xinyang Gu, Yen-Jen Wang, and Jianyu Chen. Humanoid-gym: Reinforcement learning for humanoid robot with zero-shot sim2real transfer. *arXiv preprint arXiv:2404.05695*, 2024.

[14] Xinyang Gu, Yen-Jen Wang, Xiang Zhu, Chengming Shi, Yanjiang Guo, Yichen Liu, and Jianyu Chen. Advancing humanoid locomotion: Mastering challenging terrains with denoising world model learning. *arXiv preprint arXiv:2408.14472*, 2024.

[15] Jinrui Han, Weiji Xie, Jiakun Zheng, Jiyuan Shi, Weinan Zhang, Ting Xiao, and Chenjia Bai. Kungfubot2: Learning versatile motion skills for humanoid whole-body control. *arXiv preprint arXiv:2509.16638*, 2025.

[16] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020.

[17] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *arXiv preprint arXiv:2406.08858*, 2024.

[18] Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8944–8951. IEEE, 2024.

[19] Tairan He, Jiawei Gao, Wenli Xiao, Yuanhang Zhang, Zi Wang, Jiashun Wang, Zhengyi Luo, Guanqi He, Nikhil Sobanbab, Chaoyi Pan, et al. Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. *arXiv preprint arXiv:2502.01143*, 2025.

[20] Xialin He, Runpei Dong, Zixuan Chen, and Saurabh Gupta. Learning getting-up policies for real-world humanoid robots. *arXiv preprint arXiv:2502.12152*, 2025.

[21] Runhan Huang, Shaoting Zhu, Yilun Du, and Hang Zhao. Moe-loco: Mixture of experts for multitask locomotion. *arXiv preprint arXiv:2503.08564*, 2025.

[22] Tao Huang, Junli Ren, Huayi Wang, Zirui Wang, Qingwei Ben, Muning Wen, Xiao Chen, Jianan Li, and Jiangmiao Pang. Learning humanoid standing-up control across diverse postures. *arXiv preprint arXiv:2502.08378*, 2025.

[23] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[24] Heejin Jeong and Daniel D Lee. Efficient learning of stand-up motion for humanoid robots with bilateral symmetry. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1544–1549. IEEE, 2016.

[25] Mazeyu Ji, Xuanbin Peng, Fangchen Liu, Jialong Li, Ge Yang, Xuxin Cheng, and Xiaolong Wang. Ex-

body2: Advanced expressive humanoid whole-body control, 2025. URL https://arxiv.org/abs/2412.13196.

[26] Dmitry Kalashnikov, Jacob Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. Mt-opt: Continuous multi-task robotic reinforcement learning at scale. *arXiv preprint arXiv:2104.08212*, 2021.

[27] Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. Rma: Rapid motor adaptation for legged robots. *arXiv preprint arXiv:2107.04034*, 2021.

[28] Qiayuan Liao, Takara E Truong, Xiaoyu Huang, Guy Tevet, Koushil Sreenath, and C Karen Liu. Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. *arXiv preprint arXiv:2508.08241*, 2025.

[29] Junfeng Long, Junli Ren, Moji Shi, Zirui Wang, Tao Huang, Ping Luo, and Jiangmiao Pang. Learning humanoid locomotion with perceptive internal model. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9997–10003. IEEE, 2025.

[30] Dingsheng Luo, Yaoxiang Ding, Zidong Cao, and Xihong Wu. A multi-stage approach for efficiently learning humanoid robot stand-up behavior. In *2014 IEEE international conference on mechatronics and automation*, pages 884–889. IEEE, 2014.

[31] Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10895–10904, 2023.

[32] Miles Macklin. Warp: A high-performance python framework for gpu simulation and graphics. https://github.com/nvidia/warp, March 2022. NVIDIA GPU Technology Conference (GTC).

[33] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

[34] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[35] Gabriel B Margolis and Pulkit Agrawal. Walk these ways: Tuning robot control for generalization with multiplicity of behavior. In *Conference on Robot Learning*, pages 22–31. PMLR, 2023.

[36] Mayank Mittal, Pascal Roth, James Tigue, Antoine Richard, Octi Zhang, Peter Du, Antonio Serrano-Muñoz, Xinjie Yao, René Zurbrügg, Nikita Rudin, et al. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025.

[37] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018.

[38] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4):1–20, 2021.

[39] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.

[40] Ilija Radosavovic, Tete Xiao, Bike Zhang, Trevor Darrell, Jitendra Malik, and Koushil Sreenath. Real-world humanoid locomotion with reinforcement learning. *Science Robotics*, 9(89):eadi9579, 2024.

[41] Nikita Rudin, David Hoeller, Philipp Reist, and Marco Hutter. Learning to walk in minutes using massively parallel deep reinforcement learning. In *Conference on robot learning*, pages 91–100. PMLR, 2022.

[42] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

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

[44] Nishaant Shah, Kshitij Tiwari, and Aniket Bera. Mtac: Hierarchical reinforcement learning-based multi-gait terrain-adaptive quadruped controller. *arXiv preprint arXiv:2401.03337*, 2023.

[45] Jiyuan Shi, Xinzhe Liu, Dewei Wang, Ouyang Lu, Sören Schwertfeger, Chi Zhang, Fuchun Sun, Chenjia Bai, and Xuelong Li. Adversarial locomotion and motion imitation for humanoid policy learning. *arXiv preprint arXiv:2504.14305*, 2025.

[46] Haolin Song, Hongbo Zhu, Tao Yu, Yan Liu, Mingqi Yuan, Wengang Zhou, Hua Chen, and Houqiang Li. Gait-adaptive perceptive humanoid locomotion with real-time under-base terrain reconstruction. *arXiv preprint arXiv:2512.07464*, 2025.

[47] Haoming Song, Delin Qu, Yuanqi Yao, Qizhi Chen, Qi Lv, Yiwen Tang, Modi Shi, Guanghui Ren, Maoqing Yao, Bin Zhao, et al. Hume: Introducing system-2 thinking in visual-language-action model. *arXiv preprint arXiv:2505.21432*, 2025.

[48] Jörg Stückler, Johannes Schwenk, and Sven Behnke. Getting back on two feet: Reliable standing-up routines for a humanoid robot. In *IAS*, pages 676–685, 2006.

[49] Richard S Sutton, Andrew G Barto, et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

[50] Annan Tang, Takuma Hiraoka, Naoki Hiraoka, Fan Shi, Kento Kawaharazuka, Kunio Kojima, Kei Okada, and Masayuki Inaba. Humanmimic: Learning natural loco-

motion and transitions for humanoid robot via wasserstein adversarial imitation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13107–13114. IEEE, 2024.

[51] Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions on Graphics (TOG)*, 43(6):1–21, 2024.

[52] Yuchuang Tong, Haotian Liu, and Zhengtao Zhang. Advancements in humanoid robots: A comprehensive review and future prospects. *IEEE/CAA Journal of Automatica Sinica*, 11(2):301–328, 2024.

[53] Dewei Wang, Chenjia Bai, Chenhui Li, Jiyuan Shi, Yan Ding, Chi Zhang, and Bin Zhao. Skill-nav: enhanced navigation with versatile quadrupedal locomotion via waypoint interface. *Vicinagearth*, 2(1):7, 2025.

[54] Dewei Wang, Xinmiao Wang, Xinzhe Liu, Jiyuan Shi, Yingnan Zhao, Chenjia Bai, and Xuelong Li. More: Mixture of residual experts for humanoid lifelike gaits learning on complex terrains. *arXiv preprint arXiv:2506.08840*, 2025.

[55] Weiji Xie, Jinrui Han, Jiakun Zheng, Huanyu Li, Xinzhe Liu, Jiyuan Shi, Weinan Zhang, Chenjia Bai, and Xuelong Li. Kungfubot: Physics-based humanoid whole-body control for learning highly-dynamic skills. *arXiv preprint arXiv:2506.12851*, 2025.

[56] Zhengjie Xu, Ye Li, Kwan-yee Lin, and Stella X Yu. Unified humanoid fall-safety policy from a few demonstrations. *arXiv preprint arXiv:2511.07407*, 2025.

[57] Yufei Xue, Wentao Dong, Minghuan Liu, Weinan Zhang, and Jiangmiao Pang. A unified and general humanoid whole-body controller for fine-grained locomotion. *arXiv e-prints*, pages arXiv–2502, 2025.

[58] Chuanyu Yang, Kai Yuan, Qiuguo Zhu, Wanming Yu, and Zhibin Li. Multi-expert learning of adaptive legged locomotion. *Science Robotics*, 5(49):eabb2174, 2020.

[59] Lujie Yang, Xiaoyu Huang, Zhen Wu, Angjoo Kanazawa, Pieter Abbeel, Carmelo Sferrazza, C Karen Liu, Rocky Duan, and Guanya Shi. Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. *arXiv preprint arXiv:2509.26633*, 2025.

[60] Kangning Yin, Weishuai Zeng, Ke Fan, Minyue Dai, Zirui Wang, Qiang Zhang, Zheng Tian, Jingbo Wang, Jiangmiao Pang, and Weinan Zhang. Unitracker: Learning universal whole-body motion tracker for humanoid robots. *arXiv preprint arXiv:2507.07356*, 2025.

[61] Shaofeng Yin, Yanjie Ze, Hong-Xing Yu, C Karen Liu, and Jiajun Wu. Visualmimic: Visual humanoid loco-manipulation via motion tracking and generation. *arXiv preprint arXiv:2509.20322*, 2025.

[62] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836, 2020.

[63] Kevin Zakka. Mink: Python inverse kinematics based on MuJoCo, December 2025. URL https://github.com/kevinzakka/mink.

[64] Kevin Zakka, Baruch Tabanpour, Qiayuan Liao, Mustafa Haiderbhai, Samuel Holt, Jing Yuan Luo, Arthur Allshire, Erik Frey, Koushil Sreenath, Lueder A Kahrs, et al. Mujoco playground. *arXiv preprint arXiv:2502.08844*, 2025.

[65] Yanjie Ze, Zixuan Chen, João Pedro Araújo, Zi ang Cao, Xue Bin Peng, Jiajun Wu, and C. Karen Liu. Twist: Teleoperated whole-body imitation system. *arXiv preprint arXiv:2505.02833*, 2025.

[66] Weishuai Zeng, Shunlin Lu, Kangning Yin, Xiaojie Niu, Minyue Dai, Jingbo Wang, and Jiangmiao Pang. Behavior foundation model for humanoid robots. *arXiv preprint arXiv:2509.13780*, 2025.

[67] Qiang Zhang, Peter Cui, David Yan, Jingkai Sun, Yiqun Duan, Gang Han, Wen Zhao, Weining Zhang, Yijie Guo, Arthur Zhang, et al. Whole-body humanoid robot locomotion with human reference. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11225–11231. IEEE, 2024.

[68] Zhikai Zhang, Jun Guo, Chao Chen, Jilong Wang, Chenghuai Lin, Yunrui Lian, Han Xue, Zhenrong Wang, Maoqi Liu, Jiangran Lyu, et al. Track any motions under any disturbances. *arXiv preprint arXiv:2509.13833*, 2025.

[69] Yingnan Zhao, Xinmiao Wang, Dewei Wang, Xinzhe Liu, Dan Lu, Qilong Han, Peng Liu, and Chenjia Bai. Towards adaptive humanoid control via multi-behavior distillation and reinforced fine-tuning. *arXiv preprint arXiv:2511.06371*, 2025.

[70] Ziwen Zhuang, Zipeng Fu, Jianren Wang, Christopher Atkeson, Sören Schwertfeger, Chelsea Finn, and Hang Zhao. Robot parkour learning. In *Conference on Robot Learning (CoRL)*, 2023.

[71] Ziwen Zhuang, Shenzhe Yao, and Hang Zhao. Humanoid parkour learning. *arXiv preprint arXiv:2406.10759*, 2024.

APPENDIX

A. Experimental Setup Details

1) Simulation Environment and Robot Configuration: We conduct training and evaluation in the Isaac Lab simulator [36]. The physics simulation operates at 200 Hz, while the policy runs at 50 Hz, with a decimation factor of 4. For real-world evaluation, we use the Unitree G1 humanoid robot, which is equipped with an Intel RealSense D435i and has 23 actuated degrees of freedom (6 per leg, 4 per arm, and 3 in the waist). The robot is equipped with a Jetson Orin NX for onboard computation. We assign different K_p, K_d parameters and action scales for each joint as shown in Table V, following the setup in [28].

TABLE V: PD Controller Parameters and Action Scaling

Joint Names	K_p (N·m/rad)	K_d (N·m·s/rad)	Action Scale
Hip Roll, Knee	99.10	6.31	0.35
Hip Pitch, Hip Yaw, Waist Yaw	40.18	2.56	0.55
Ankle (P/R), Waist (R/P)	28.50	1.81	0.44
Shoulder (P/R/Y), Elbow	14.25	0.91	0.44

2) Reward Formulation: The formulation of reward functions and their respective weights for the three specialist policies ($\pi_\text{l}, \pi_\text{r}, \pi_\text{w}$) are detailed in Table IX–XI. To facilitate the emergence of naturalistic motion patterns, we incorporate an Adversarial Motion Prior (AMP) reward into the training of the upright locomotion (π_l) and fall recovery (π_r) specialists. The mathematical formalization of the AMP reward and its optimization objectives are provided in Appendix B.

3) Domain Randomization: We implement a consistent Domain Randomization (DR) configuration during the specialist policy training and generalist policy distillation to facilitate robust sim-to-real transfer. Our DR mainly involves randomizing physical parameters, initial states, and actuator dynamics. Specifically, we vary ground friction and inertial properties to handle environmental diversity, while applying external force perturbations to simulate real-world uncertainties. The full randomization configuration is listed in Table VI.

TABLE VI: Domain Randomization Parameters. ($\mathcal{U}[\cdot]$: uniform distribution)

Domain Randomization	Sampling Distribution	Unit
Static friction	$\mu_\text{static} \sim \mathcal{U}[0.6, 1.0]$	-
Dynamic friction	$\mu_\text{dynamic} \sim \mathcal{U}[0.4, 0.8]$	-
Torso payload mass	$m_\text{payload} \sim \mathcal{U}[-1.0, 5.0]$	kg
Initial joint positions	$\theta_\text{init} \sim \mathcal{U}[0.8, 1.2] \times \theta_\text{nominal}$	-
External push interval	$\Delta t_\text{push} \sim \mathcal{U}[5.0, 8.0]$	s
Push linear velocity	$v_x, v_y \sim \mathcal{U}[-0.5, 0.5], v_z \sim \mathcal{U}[-0.2, 0.2]$	m/s
Push angular velocity	$\omega_r, \omega_p \sim \mathcal{U}[-0.52, 0.52], \omega_y \sim \mathcal{U}[-0.78, 0.78]$	rad/s
Actuator stiffness K_p	$k_p \sim \mathcal{U}[0.8, 1.2] \times K_{p,\text{nominal}}$	-
Actuator damping K_d	$k_d \sim \mathcal{U}[0.8, 1.2] \times K_{d,\text{nominal}}$	-

4) Terrain Curriculum: We adopt an adaptive terrain curriculum inspired by [41] for the training of π_l and π_g, which automatically modulates terrain difficulty based on agent performance. The environment is structured as a 10×20 grid of $8\text{m} \times 8\text{m}$ patches, encompassing five distinct terrain types: slopes, pits, hanging bars, stairs, and boxes. The difficulty is scaled by adjusting geometric parameters: slopes range from 0° to 20°; pits vary in height from 0.05m to 0.3m; hanging bars enforce vertical clearance constraints of 0.67m to 0.72m; stair step heights range from 0m to 0.15m; and climbing boxes vary in height from 0.45m to 0.65m.

5) Training Implementation Details: In this section, we detail the neural network architectures and the training parameters for both the specialist and generalist policies.

a) Specialist Policy Architecture: The specialist policies ($\pi_\text{l}, \pi_\text{r}, \pi_\text{w}$) are trained using the Actor-Critic framework and comprise some or all of the following modules:

- **History Encoder:** An MLP with hidden units $[1024, 512, 128]$, utilized by $\pi_\text{l}, \pi_\text{r}$.
- **Elevation Map Encoder:** A 3-layer MLP $[128, 64, 32]$ employed by π_l and π_w to process local terrain geometry.
- **Privileged Proprioception Encoder:** A 3-layer MLP $[128, 64, 32]$ for π_l to encode privileged states.
- **Actor Head and Critic:** Both networks are modeled as 3-layer MLPs with dimensions $[512, 256, 128]$. For the fall recovery specialist π_r, we adopt a multi-critic structure [22] comprising four independent critics.

b) Generalist Policy Architecture: The generalist policy π_g leverages a Mixture-of-Experts (MoE) architecture, where a gating network modulates the contributions of experts. The experts and the gating network are instantiated as 3-layer MLPs $[512, 256, 128]$. To incorporate exteroception, a CNN-based depth encoder is employed to compress sequential depth images into a 128-dimensional latent representation. In parallel, a history encoder is utilized to process proprioceptive sequences over a 10-step horizon. This temporal context, combined with the latent representation from the depth encoder, is subsequently fed into the MoE to predict the actions.

c) Training Hyperparameters: Specialist policies are trained via PPO [43], while the generalist policy is distilled using supervised behavior cloning. Detailed training hyperparameters are provided in Table VII.

TABLE VII: Optimization Hyperparameters for Specialists and Generalist Policies

Hyperparameter	**Specialist Training (PPO)**	**Generalist Distillation**
Number of environments	4096	4096
Learning rate	1.0×10^{-3} (Adaptive)	1.0×10^{-3} (Fixed)
Num. epochs per iteration	5	8
Num. mini-batches	4	12
Steps per training batch	24	12
Discount factor (γ)	0.99	-
GAE parameter (λ)	0.95	-
PPO clip parameter (ϵ)	0.2	-
Entropy coefficient	0.005	-
Desired KL divergence	0.01	-
Total Training Iterations		
Upright Locomotion (π_l)	30,000	
Recovery (π_r)	10,000	
Whole-Body Coordination (π_w)	50,000	
Generalist Policy (π_g)	30,000	

Algorithm 1 Synergetic Policy Distillation

Require: Set $\mathcal{C} = \{(c_{rec}, \pi_r), (c_{loco}, \pi_l), (c_{coor}, \pi_w)\}$, ratio ρ, decay $\Delta\rho$, threshold ϵ

Initialize storage buffer $\mathcal{D}$ and student policy π_θ
for iteration = $1, 2, \ldots$ **do**
  Clear storage $\mathcal{D}$
  // **Data Collection with SAR and SFI**
  for step $t = 1$ to T in N parallel environments **do**
    Get student state $o_{u,t}$ and privileged states $s_{i,t}$.
    SFI: Inject external forces based on context.
    Determine case index $i \leftarrow f(b_t, I_t)$ and select corresponding specialist π_i.
    Get expert action: $a_t^* \leftarrow \pi_i(s_{i,t})$.
    Get student action: $a_t^u \leftarrow \pi_u(o_{u,t})$.
    SAR: Sample $b_t \sim \text{Bernoulli}(\rho)$:
    $a_{env} \leftarrow b_t a_t^* + (1 - b_t) a_t^u$
    Execute a_{env}, and store $(o_{u,t}, a_t^*)$ in $\mathcal{D}$.
    Termination: if specialist π_i termination criteria met **or** SFI-specific failure (Eq. 5) **then** reset env
  end for
  // **Policy Optimization**
  for epoch = 1 to K **do**
    Sample minibatches from $\mathcal{D}$.
    $\mathcal{L}_{distill} \leftarrow \text{MSE}(\pi_u(o_{u,t}), a_t^*)$.
    Update θ via gradient descent to minimize $\mathcal{L}_{distill}$.
  end for
  // **Specialist Annealing Schedule**
  if $\mathcal{L}_{distill} < \epsilon$ **then**
    $\rho \leftarrow \max(0, \rho - \Delta\rho)$
  end if
end for

6) Training Pseudocode: We present the training procedure for the generalist policy π_{g} in Algorithm 1.

TABLE VIII: Depth Augmentation and Camera Randomization Parameters. ($\mathcal{U}[\cdot]$: uniform distribution)

Camera Randomization	Value / Range	Unit
Gaussian noise	$\sigma_{\text{noise}} = 0.02$	m
Gaussian filter kernel	$k \in \{3 \times 3\}$	pixel
Gaussian filter sigma	$\sigma_{\text{filter}} = 1.0$	-
Camera position	$\Delta x, \Delta y, \Delta z \sim \mathcal{U}[-0.05, 0.05]$	m
Camera rotation	$\theta_{\text{pitch}} \sim \mathcal{U}[-10, 5]$, $\theta_{\text{roll}}, \theta_{\text{yaw}} \sim \mathcal{U}[-1, 1]$	$^\circ$
Camera horizontal FOV	$\Delta\text{FOV}_{\text{h}} \sim \mathcal{U}[-10, 10]$	$^\circ$

B. Adversarial Motion Prior (AMP) Formulation

To facilitate naturalistic motion patterns, we incorporate the Adversarial Motion Prior (AMP) framework [38, 12] into the training of the locomotion (π_{l}) and recovery (π_{r}) specialists. The AMP module produces a style-dependent reward r_t^{style}, by training a discriminator D_ϕ to differentiate between reference trajectories and policy-generated ones.

The discriminator operates on a compact input representation $\tau_t = (s_{t-3:t+1}^{\text{amp}})$, which comprises a 5-step sequence of AMP states. Each individual AMP state $s_t^{\text{amp}} \in \mathbb{R}^{23}$ consists of the robot's joint positions. Following the Least Squares GAN (LSGAN) [34] formulation, the discriminator is optimized by maximizing the following objective function:

$$\begin{aligned}\mathcal{L}_D = \mathbb{E}_{\tau\sim\mathcal{M}}[(D_\phi(\tau) - 1)^2] + \mathbb{E}_{\tau\sim\mathcal{P}}[(D_\phi(\tau) + 1)^2] \\ + \frac{\alpha^d}{2}\mathbb{E}_{\tau\sim\mathcal{M}}[\|\nabla_\phi D_\phi(\tau)\|_2^2],\end{aligned} \tag{9}$$

where $\mathcal{M}$ and $\mathcal{P}$ denote the reference dataset and the on-policy rollout buffer, respectively. Training stability is maintained via a gradient penalty with coefficient α^d applied to the reference data.

The discriminator's output $d = D_\phi(\tau_t)$ is subsequently mapped to a surrogate reward r_t^{style} to guide the policy:

$$r^{\text{style}}(s_t) = w_{\text{style}} \cdot \max\left(0,\, 1 - \frac{1}{4}(d-1)^2\right), \tag{10}$$

where w_{style} is a scaling coefficient. The composite reward used for policy optimization is the summation of task-specific and style-dependent components: $r_t = r_t^{\text{task}} + r_t^{\text{style}}$.

Reference motions for the locomotion specialist (π_{l}) are source from the LAFAN1 dataset [16], retargeted by Unitree. For the recovery specialist (π_{r}), reference motions are sourced from motion capture (MoCap) data and retargeted to the Unitree G1 platform.

C. Case-Adaptive Specialist Selection Details

The Case-Adaptive Specialist Selection (CASS) mechanism selects an appropriate specialist to guide the student policy based on the robot's head height b_t and the environmental context I_t. This mechanism incorporates a mapping function $f(b_t, I_t)$ that identifies the robot's specific case, alongside a set $\mathcal{C} = \{(c_{\text{rec}}, \pi_r), (c_{\text{loco}}, \pi_l), (c_{\text{coor}}, \pi_w)\}$ that maps these cases to their respective specialists. Specifically, if the robot's head height b_t is below $1.1m$, it is categorized as case c_{rec}, where the fall recovery policy π_r provides guidance. If b_t exceeds $1.1m$, the classification depends on the local environmental context I_t. In scenarios where the robot is in proximity to a hanging bar or a high platform, it is assigned to case c_{coor} and guided by π_w; otherwise, it is classified as c_{loco} and supervised by π_l. Once π_w is selected, CASS provides reference motions tailored to different terrains, enabling π_w to generate the appropriate guidance actions. Consequently, c_{coor} includes into several sub-cases c_{coor}^i based on the specific reference motion requirements. The definition of I_t is illustrated in Figure 7. Accordingly, $f(b_t, I_t)$ can be expressed as:

$$f(b_t, I_t) = \begin{cases} c_{\text{rec}}, & \text{if } b_t < 1.1 \\ c_{\text{coor}}^1, & \text{if } b_t \geq 1.1 \text{ and } I_t = 1 \\ c_{\text{coor}}^2, & \text{if } b_t \geq 1.1 \text{ and } I_t = 2 \\ c_{\text{loco}}, & \text{otherwise} \end{cases}, \tag{11}$$

where $I_t = 1$ signifies that the robot is facing a high platform requiring a climbing maneuver, while $I_t = 2$ represents the presence of a hanging obstacle approximately $0.7m$ above the ground that can be navigated by forward rolling.

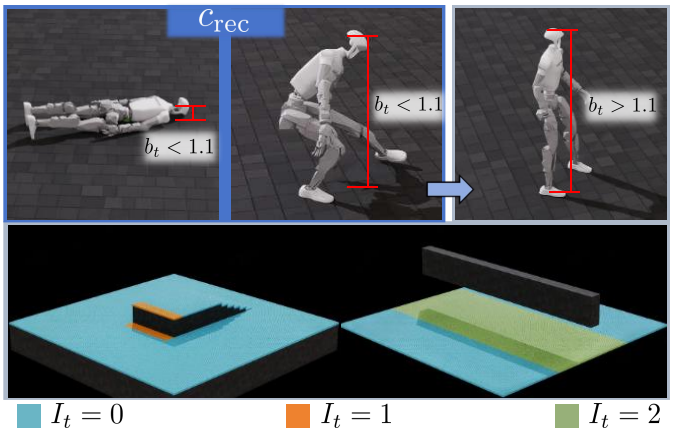


Fig. 7: Illustration of the case-adaptive specialist selection process based on b_t and I_t.

D. Whole-Body Coordination Specialists π_w Details

To facilitate local terrain awareness, the whole-body coordination specialists π_w utilizes a terrain height-scanner centered on the robot's pelvis to get the local terrain height samples (i.e., an elevation map). The sensor captures the local terrain geometry as a grid of $1.6m \times 1.0m$ with a spatial resolution of $0.1m$. Furthermore, we employ a random initialization strategy. In Box terrain, the robot is initialized at a longitudinal distance sampled from $\mathcal{U}[0, 0.55]m$ ralative to the box. For Hanging Bar terrain, the robot is initialized at a distance sampled from $\mathcal{U}[3, 3.5]m$ from the bar. This initialization strategy ensures that π_{w} encounters a diverse set of relative distances to obstacles, thereby improving the ability to handle environmental variability.

E. Depth Alignment for Sim-to-Real Transfer

To bridge the visual sim-to-real gap, we implement a process that matches simulated depth distributions to real-world sensor characteristics. By injecting Gaussian noise, blur, and randomizing camera intrinsic and extrinsic parameters, we approximate the sensor observations of the actual hardware. The specific noise and randomization parameter ranges are detailed in Table VIII.

TABLE IX: Reward Function Definitions for the Whole-Body Coordination Specialist (π_w).

Reward Terms	**Equation**	**Weight**
Body position	$\exp\left(-\left(\frac{1}{\lvert\mathcal{B}_{\text{target}}\rvert}\sum_{b\in\mathcal{B}_{\text{target}}}\lVert\mathbf{p}_b^{\text{des}}-\mathbf{p}_b\rVert^2\right)/0.3^2\right)$	1.0
Body orientation	$\exp\left(-\left(\frac{1}{\lvert\mathcal{B}_{\text{target}}\rvert}\sum_{b\in\mathcal{B}_{\text{target}}}\lVert\log(R_b^{\text{des}}R_b^\top)\rVert^2\right)/0.4^2\right)$	1.0
Body linear velocity	$\exp\left(-\left(\frac{1}{\lvert\mathcal{B}_{\text{target}}\rvert}\sum_{b\in\mathcal{B}_{\text{target}}}\lVert\mathbf{v}_b^{\text{des}}-\mathbf{v}_b\rVert^2\right)/1.0^2\right)$	1.0
Body angular velocity	$\exp\left(-\left(\frac{1}{\lvert\mathcal{B}_{\text{target}}\rvert}\sum_{b\in\mathcal{B}_{\text{target}}}\lVert\boldsymbol{\omega}_b^{\text{des}}-\boldsymbol{\omega}_b\rVert^2\right)/3.14^2\right)$	1.0
Anchor position	$\exp\left(-\lVert\mathbf{p}_{\text{anchor}}^{\text{des}}-\mathbf{p}_{\text{anchor}}\rVert^2/0.3^2\right)$	0.5
Anchor orientation	$\exp\left(-\lVert\log(R_{\text{anchor}}^{\text{des}}R_{\text{anchor}}^\top)\rVert^2/0.4^2\right)$	0.5
Action smoothness	$\lVert\mathbf{a}_t-\mathbf{a}_{t-1}\rVert^2$	-0.1
Joint position limit	$\sum_{j=1}^{N}[\max(l_j-\theta_j,0)+\max(\theta_j-u_j,0)]$	-10.0
Undesired self-contacts	$\sum_{b\notin\mathcal{B}_{\text{ee}}}\mathbf{1}[\lVert f_b^{\text{self}}\rVert>1\text{N}]$	-0.1

TABLE X: Reward Function Definitions for the Locomotion Specialist (π_l).

Reward Terms	**Equation**	**Weight**
Track lin. vel.	$\exp\{-\frac{\lVert\boldsymbol{v}_{\text{lin}}^{\text{cmd}}-\boldsymbol{v}_{\text{lin}}\rVert_2^2}{0.5}\}$	5.0
Track ang. vel.	$\exp\{-\frac{(\boldsymbol{\omega}_{\text{yaw}}^{\text{cmd}}-\boldsymbol{\omega}_{\text{yaw}})^2}{0.5}\}$	5.0
Joint acc.	$\lVert\ddot{\theta}\rVert_2^2$	-5×10^{-7}
Joint vel.	$\lVert\dot{\theta}\rVert_2^2$	-1×10^{-3}
Action rate	$\lVert\boldsymbol{a}_t-\boldsymbol{a}_{t-1}\rVert_2^2$	-0.03
Action smoothness	$\lVert\boldsymbol{a}_t-2\boldsymbol{a}_{t-1}+\boldsymbol{a}_{t-2}\rVert_2^2$	-0.05
Angular vel. (x,y)	$\lVert\boldsymbol{\omega}_{xy}\rVert_2^2$	-0.05
Orientation	$\lVert\boldsymbol{g}_{xy}^{\text{torso}}\rVert_2^2$	-1.5
Joint power	$\lvert\tau\rvert\lvert\dot{\theta}\rvert^\top$	-2.5×10^{-5}
Feet stumble	$\mathbb{I}(\exists i, \lvert\boldsymbol{F}_i^{xy}\rvert\geq3\lvert F_i^z\rvert)$	-1.0
Torques	$\sum_{\text{all joints}}\tau_i^2$	-1×10^{-5}
Joint deviation	$\sum_{k\in\{\text{arm, waist, hip}\}}w_k\sum_{j\in\mathcal{J}_k}\lvert\theta_j-\theta_j^{\text{def}}\rvert$	-0.5
Joint pos. limits	$\sum_{\text{all joints}}\boldsymbol{out}_i$	-2.0
Joint vel. limits	$\text{ReLU}(\dot{\theta}-\dot{\theta}^{\max})$	-1.0
Torque limits	$\text{ReLU}(\tau-\tau^{\max})$	-1.0
Feet lateral distance	$\lvert(y^b_{\text{left feet}}-y^b_{\text{right feet}})-0.2\rvert$	0.5
Feet slippage	$\sum_{\text{feet}}\lVert\boldsymbol{v}_i^{\text{foot}}\rVert\cdot\mathbb{I}_{\text{contact}}$	-0.25
Collision	$n_{\text{collision}}$	-15.0
Feet air time	$\sum_{\text{foot}}(t_i^{\text{air}}-0.5)\cdot\mathbb{I}(\text{first contact}_i)$	1.0
Stuck	$(\lVert\boldsymbol{v}\rVert_2\leq0.1)\cdot(\lVert\boldsymbol{c}^v\rVert_2\geq0.2)$	-1.0
Feet clearance	$\sum_{\text{foot}}((z^i-h^{\text{target}})^2\cdot\lVert\boldsymbol{v}_{xy}^i\rVert)$	2.0
Alive	1	2
AMP reward	$\max\left[0,1-\frac{1}{4}(D_\phi(\tau)-1)^2\right]$	3.0

TABLE XI: Reward Function Definitions for the Recovery Specialist (π_r). The f_{tol} adopts a Gaussian-style formulation, as detailed in [20].

Reward Terms	**Equation**	**Weight**
Task Rewards		$w^{\text{task}}=1.0$
Orientation	$f_{\text{tol}}(-\theta^z_{\text{base}},[0.99,\infty],1,0.05)$	1.0
Head height	$f_{\text{tol}}(h_{\text{head}},[1,\infty],1,0.1)$	1.0
Style Rewards		$w^{\text{style}}=1.0$
Hip joint deviation	$\sum_{\text{hips}}\mathbb{I}(\max\lvert\theta_i\rvert>0.9\vee\min\lvert\theta_i\rvert<0.8)$	-10.0
Knee deviation	$\sum_{\text{knees}}\mathbb{I}(\max\lvert\theta_i\rvert>2.85\vee\min\lvert\theta_i\rvert<-0.06)$	-0.25
Shoulder roll dev.	$\mathbb{I}(\theta_{\text{left}}<-0.02\vee\theta_{\text{right}}>0.02)$	-2.5
Thigh orientation	$f_{\text{tol}}(\frac{1}{2}\sum_{\text{thighs}}(\theta^z_{\text{thigh}}),[0.8,\infty],1,0.1)$	10.0
Feet distance	$\mathbb{I}(\lVert\mathbf{p}^{xy}_{\text{left\_f}}-\mathbf{p}^{xy}_{\text{right\_f}}\rVert^2>0.9)$	-10.0
Angular vel. (x,y)	$\exp(-2\lVert\boldsymbol{\omega}_{xy}\rVert_2^2)\cdot\mathbb{I}(h_{\text{base}}>h_{\text{stage1}})$	25.0
Foot displacement	$\exp(\text{clip}(-2\lVert\mathbf{q}^{xy}_{\text{base}}-\mathbf{q}^{xy}_{\text{foot}}\rVert^2,0.3,\infty))$	2.5
AMP reward	$\max\left[0,1-\frac{1}{4}(D_\phi(\tau)-1)^2\right]$	80.0
Regularization Rewards		$w^{\text{regu}}=1.0$
Joint acc.	$\lVert\ddot{\theta}\rVert_2^2$	-2.5e-7
Joint vel.	$\lVert\dot{\theta}\rVert_2^2$	-1e-3
Action rate	$\lVert\mathbf{a}_t-\mathbf{a}_{t-1}\rVert_2^2$	-0.01
Action smoothness	$\lVert\mathbf{a}_t-2\mathbf{a}_{t-1}+\mathbf{a}_{t-2}\rVert_2^2$	-0.05
Torques	$\sum\tau_i^2$	-1×10^{-5}
Joint power	$\lvert\tau\rvert\lvert\dot{\theta}\rvert^\top$	-2.5×10^{-5}
Joint pos. limits	$\sum\text{ReLU}(\lvert\theta_i\rvert-\theta_i^{\max})$	-2.0
Joint vel. limits	$\text{ReLU}(\dot{\theta}-\dot{\theta}^{\max})$	-1.0
Post-Task Rewards (Conditioned on $h_{base}>h_{stage3}$)		$w^{\text{task}}=1.0$
Tracking errors	$\exp(-2\lVert\boldsymbol{\omega}_{xy}\rVert_2^2),\exp(-5\lVert\mathbf{v}_{xy}\rVert_2^2),\exp(-5\lVert\mathbf{g}_{xy}\rVert_2^2)$	10.0
Base height	$\exp(-20\lvert h_{\text{base}}-0.75\rvert)$	10.0
Target joint dev.	$\exp(-0.1\sum(\theta_i-\theta_i^{\text{def}})^2)$	10.0
Target feet dist.	$f_{\text{tol}}(y_{\text{L}}-y_{\text{R}},[0.3,0.4],0.1,0.05)$	-5.0