Robotics: Science and Systems 2026
Sydney, Australia, July 13–17, 2026

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

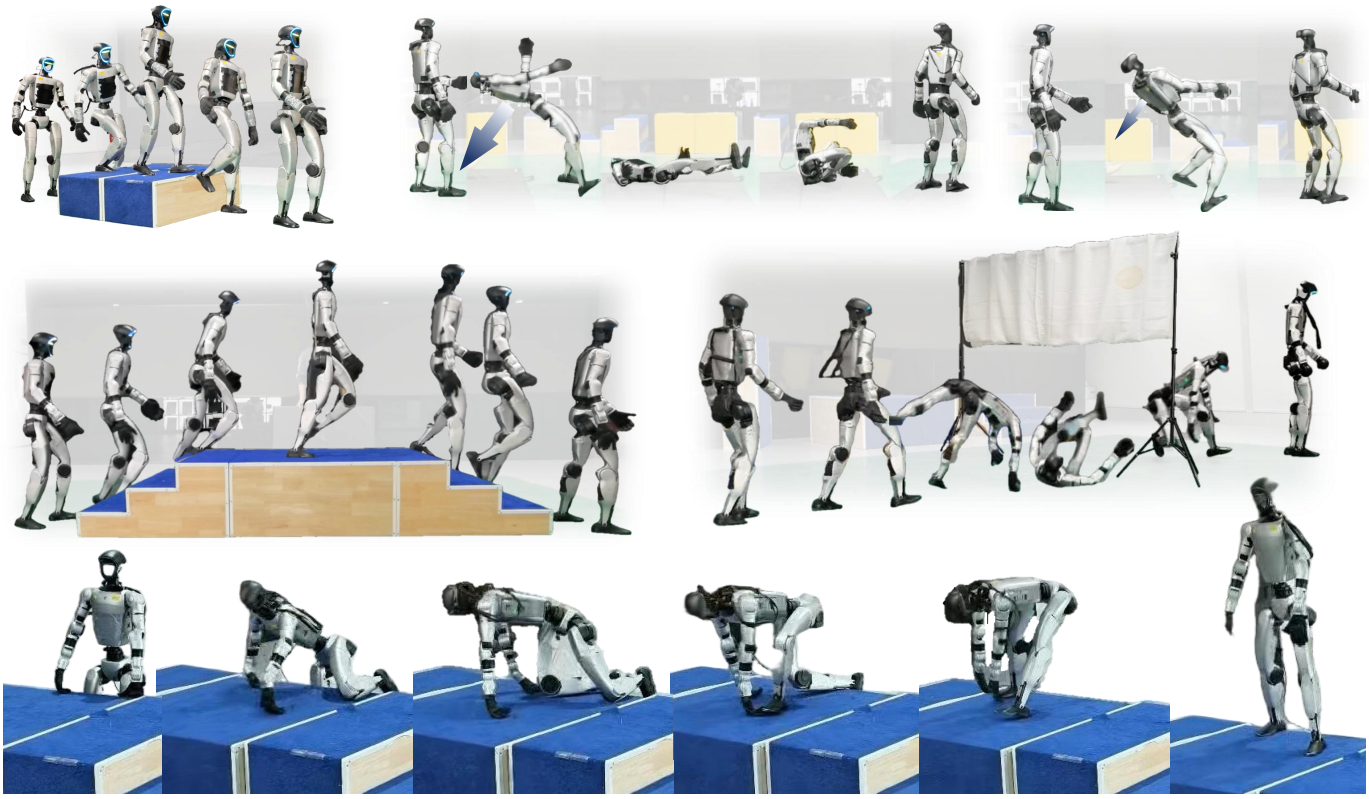


Fig. 1: X-Loco achieves vision-based generalist humanoid locomotion control. Relying solely on velocity commands without reference motions, X-Loco leverages proprioception and exteroception to traverse complex terrains, performing behaviors such as climbing up and down stairs, fall recovery, box climbing, 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. X-Loco demonstrates vision-based humanoid locomotion that first jointly integrates upright locomotion, whole-body coordination and fall recovery, while operating solely under velocity commands without relying on reference motions during the inference time. 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 [55, 10, 39, 50, 4, 18, 2]. As a fundamental capability, locomotion control has witnessed rapid progress facilitated by re-

inforcement learning (RL) and high-fidelity physical simulators [52, 46, 12, 36, 33, 69, 22, 18]. Recent advancements have demonstrated remarkable results, ranging from traversing challenging terrains [57, 29, 49, 13] and executing robust fall recovery [7, 19, 61] to agilely tracking highly dynamic motions [18, 28, 65, 58, 14, 73].

Despite these advancements, a key limitation of existing humanoid locomotion control methods lies in their predominant focus on isolated or specific categories of locomotion skills, such as upright locomotion [57, 29, 49, 76] or fall recovery [22, 61, 30, 51]. This functional fragmentation renders these humanoid locomotion controllers incapable of handling complex scenarios, such as autonomously resuming locomotion after a fall. Furthermore, these methods fail to achieve whole-body coordination skills interacting with terrains, which prevents the robot from traversing challenging terrains. While motion tracking-base methods [14, 66, 71, 65, 64] and teleoperation paradigms [16, 70, 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. Hierarchical architectures [59, 60] comprising a high-level reference generator and a low-level tracker struggle to respond agilely to complex real-world scenarios or unexpected events in real time and the tight coupling between these two systems complicates the training process. Overall, existing methods are unable to achieve a generalist policy that enables humanoid robots to autonomously perform vision-based upright locomotion, fall recovery and whole-body coordinated motor skills according to environmental conditions.

Developing a vision-based generalist locomotion controller capable of integrating locomotion skills ranging from upright locomotion to fall recovery, alongside whole-body coordination behaviors like box climbing, remains an open challenge for humanoid robots. This necessitates the robot's ability to autonomously execute diverse motor behaviors in real-time to negotiate with complex environments, independent of human teleoperation or reference motions. To realize such a versatile controller, several challenges must be addressed: First, achieving whole-body coordination skills [64, 14] for terrain interaction without reference motions introduces substantial complexity, as the absence of guidance leads to inefficient exploration coupled with the requirement for precise environmental perception. Second, directly and concurrently learning diverse skills including terrain traversal, fall recovery, and whole-body coordination significantly exacerbates exploration difficulty and introduces challenges such as gradient interference. Lastly, beyond mastering diverse skills, the controller needs to leverage visual perception to understand the environment and generate appropriate, real-time responses to conditions.

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 [12, 40, 72, 12], visual locomotion across complex terrains [57, 49, 29] and multiple locomotion skill learning [76, 62]. By decoupling the control of the upper and lower limbs [5, 48, 62], 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, 7]. FIRM [61] has explored the

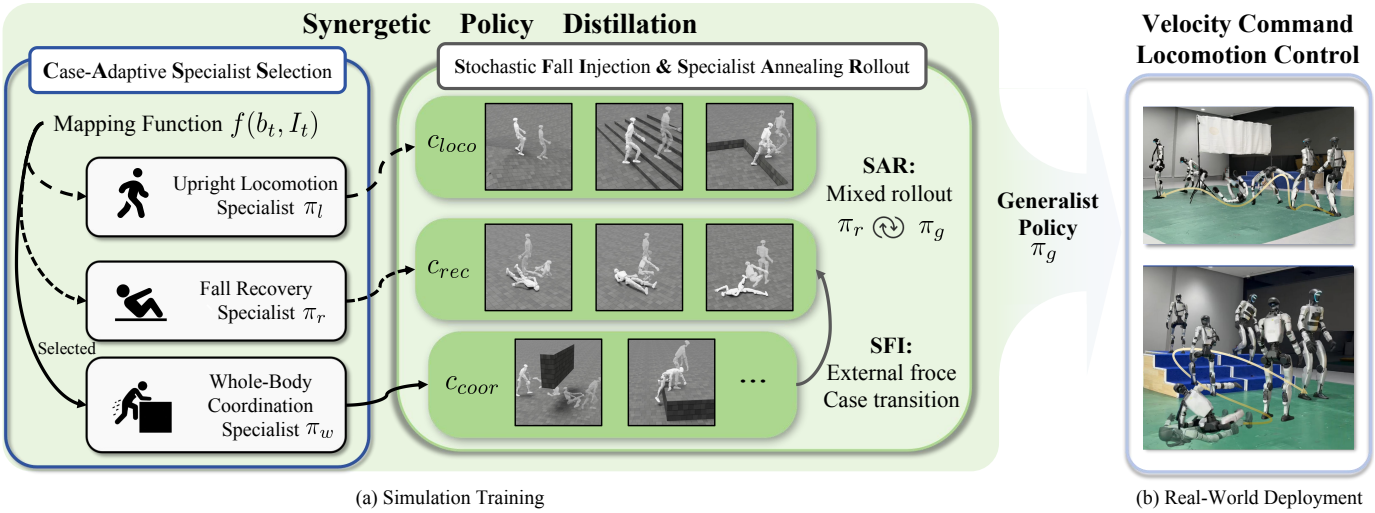


Fig. 2: **Overview of X-Loco**. (a) We first train three privileged specialist policies across three distinct locomotive skills. Subsequently, we employ Synergetic Policy Distillation to efficiently distill these skills into a visual generalist policy π_g. (b) X-Loco can perform diverse locomotion skills in the real world.

synergy between fall-safety and fall-recovery and AHC [74] 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, 68, 17] and motion imitation [37, 54, 31], humanoid robots can perform highly dynamic behaviors [18, 58, 73, 28] and universal motion tracking [14, 71, 65, 8] given reference motions. Exteroceptive information is integrated into a motion tracking framework to enable complex loco-manipulation [66] 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, 67]. Mixture-of-Experts (MoE) [6, 23] has been shown to mitigate gradient conflict in multi-skill learning and effectively facilitate the acquisition of diverse locomotion gaits [57, 21]. Methods such as MELA [63] and MTAC [47] utilize a hierarchical structure to decouple individual motor skills, which are subsequently coordinated by a high-level module. OGMP [41] leveraging hybrid automatons enables a single policy to master loco-manipulation tasks that require multiple skills through preference-reward-guided optimization. Locomotion skills with minor morphological variations can be acquired by one-stage framework leveraging reward function design [62, 35] and curriculum learning [9, 56].

In legged locomotion, policy distillation is widely utilized for privileged information learning [9, 16, 8] and sim-to-real transfer [66, 27]. Policy distillation enables dynamic parkour behaviors by training specialized teachers for different terrains and distilling their expertise into a single policy [75, 76, 20, 44]. 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) [46] and Generalized Advantage Estimation (GAE) [45] 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, 53] 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 based on a blind motion tracking paradigm [28] which can achieve high-fidelity motion tracking without AMP-based reward. 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. Specifically, CASS serves as the high-level logic switcher that defines the learning objective by selecting the target specialist. Once the specialist is designated, SAR governs the rollout data distribution by adaptively mixing specialist guidance with student exploration to bridge the gap between specialist behavior and non-privileged student execution. To enable the policy to learn recovery maneuvers across various hazardous states rather than merely initiating a simple stand-up routine from the ground at the start of an episode, SFI actively pushes the robot from stable locomotion into recovery states. This tripartite synergy ensures that the student policy not only mimics specialist actions but also masters the temporal transitions and robustness required for complex real-world deployments, as detailed in Algorithm 1 of the Appendix.

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.

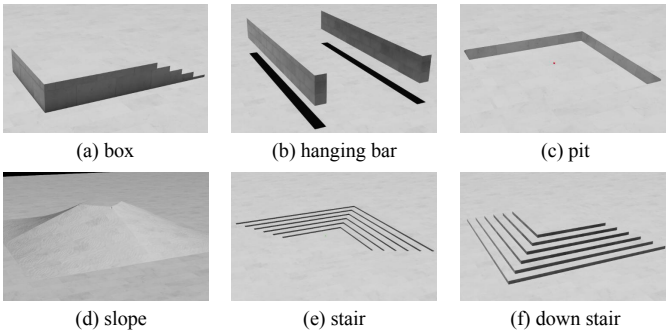
2) Specialist Annealing Rollout: Only with the CASS mechanism, training a single policy to master multiple specialist domains simultaneously is challenging. This difficulty stems from two primary issues: First, an untrained policy frequently generates trajectories that are out-of-distribution (OOD) relative 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:

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

where $\xi_t \sim \text{Bernoulli}(\rho)$ 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.



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

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

Fig. 3: Terrains used for training and evaluation.

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 other parallel environments. Ray-caster camera based on NVIDIA Warp [32] achieves a rendering speed 30 times faster than the default simulator camera, but it struggle to concurrently handle both dynamic and static meshes. Our approach further 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, maintaining a rendering speed 10 times faster than the default simulator camera.

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:

- **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 [57]:** This baseline employs a two-stage pipeline training a vision-based policy to achieve robust locomotion across complex terrains.
- **AHC [74]:** This baseline implements an adaptive blind control policy that unifies locomotion with autonomous fall recovery behaviors.
- **PPO [46]:** 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: Based on the terrain types, Upright Locomotion and Whole-Body Coordination (WBC) tasks are categorized into several sub-tasks. While the specialists define the performance upper bound for each specific task, X-Loco is capable of executing all tasks using a single generalist policy, which is a capability that other baselines lack. We provide a detailed analysis of the performance of the specialists, X-Loco, and the baselines across different tasks based on Table I in the following.

a) Specialist vs. Baselines: 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 $\boldsymbol{h}_t$, rather than blind tracking reference motions that disregards terrain interactions. 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 and provides a solid foundation for distilling these capabilities into a generalist policy.

b) X-Loco vs. Baselines: X-Loco 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 relies on pre-defined reference motions, and MoRE shows competitive results in locomotion but fail completely in **WBC** and **Recovery** tasks, X-Loco successfully masters all three tasks. Although AHC with fall recovery capability can perform two types of tasks, it remains a blind approach, which results in significantly inferior terrain traversal performance compared to X-Loco. Notably, directly incorporating perception without any adjustments to the framework of AHC leads to training collapse. In the **Upright Locomotion** task, X-Loco achieves

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	Upright 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 [57]	**0.992** ± .008	7.863 ± .018	0.844 ± .009	6.833 ± .114	0.926 ± .009	7.387 ± .013	0.921 ± .009	-	-	-	-
PPO [46]	0.823 ± .021	6.674 ± .152	0.793 ± .018	6.396 ± .184	0.781 ± .025	6.565 ± .141	0.799 ± .021	-	-	-	-
AHC [74]	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	-	-

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 during the inference time, 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 relatively 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. We provide a further analysis of X-Loco regarding failure-prone cases in the following. Furthermore, a manually designed selection strategy for specific scenarios could enable terrain traversal by switching specialists, but it would introduce nontrivial design complexity and suffer from limit scalability. These results indicate that our framework can effectively distill and integrate diverse expert knowledge into a single policy, 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.

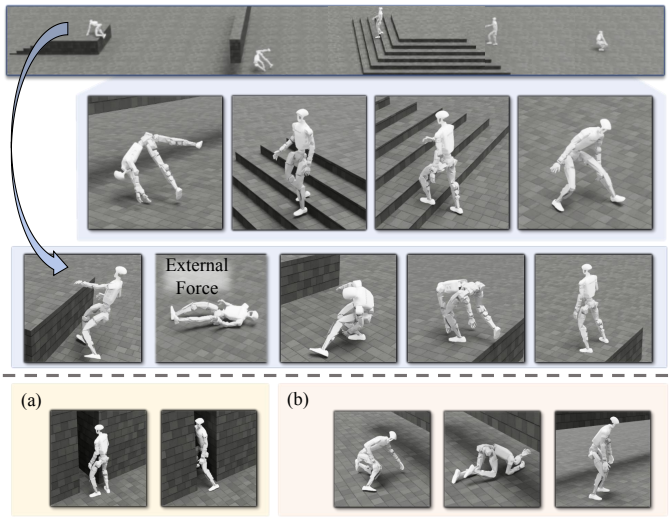


Fig. 4: **Top**: Testing the generalist policy on hybrid, challenging terrains. **Bottom**: Extensibility of the X-Loco to include vision-guided (a) lateral sidling and (b) kneeling crawling.

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**

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** tasks. This confirms that the MLP backbone lacks the capacity

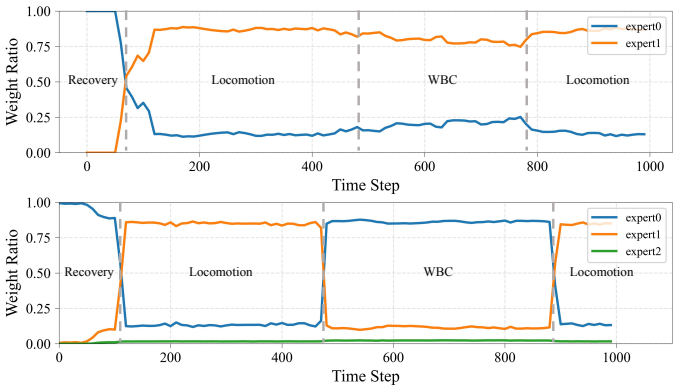


Fig. 5: MoE experts utilization analysis.

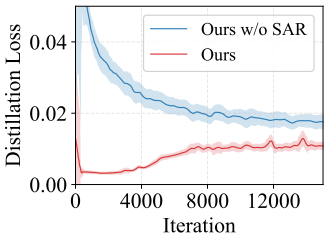


Fig. 6: Distillation loss curves of ablation on SAR.

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

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 found that MoE-2 demonstrates more effective specialization as shown in Fig. 5, with one expert dedicated to fall recovery while both experts are jointly activated for other task. In contrast, one expert in the MoE-3 variant remains inactive, leading to sparse updates and sub-optimal utilization.

c) Ablations on SAR & SFI: As shown in Fig. 6, 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.

TABLE V: Success rates of sub-phases in WBC tasks.

Box Climbing			**Forward Rolling**		
Sub-p1	Sub-p2	Sub-p3	Sub-p1	Sub-p2	Sub-p3
100%	100%	76.5%	100%	82.4%	76.5%

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 subsequently 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.

3) Failure Case Analysis: We provide a further analysis of the failure cases encountered during WBC tasks. We partitioned the Box Climbing task into three distinct sub-phases: hands-propping, single-leg kneeling, and kneeling-to-standing. Similarly, Forward Rolling was divided into bending, rolling, and standing phases. As shown in Table V, failures are concentrated (i) during the transition from single leg kneeling to standing during box climbing, and (ii) in the late stage of forward rolling where an explosive leg thrust is required. These failures stem from the difficulty in generating the precise peak joint forces across the specific joints required for such complex, high-torque state transitions, particularly in the absence of reference motions.

D. Real-World Deployment

a) Performance on Hybird Terrains: To validate the robustness and versatility of X-Loco in the real world, we

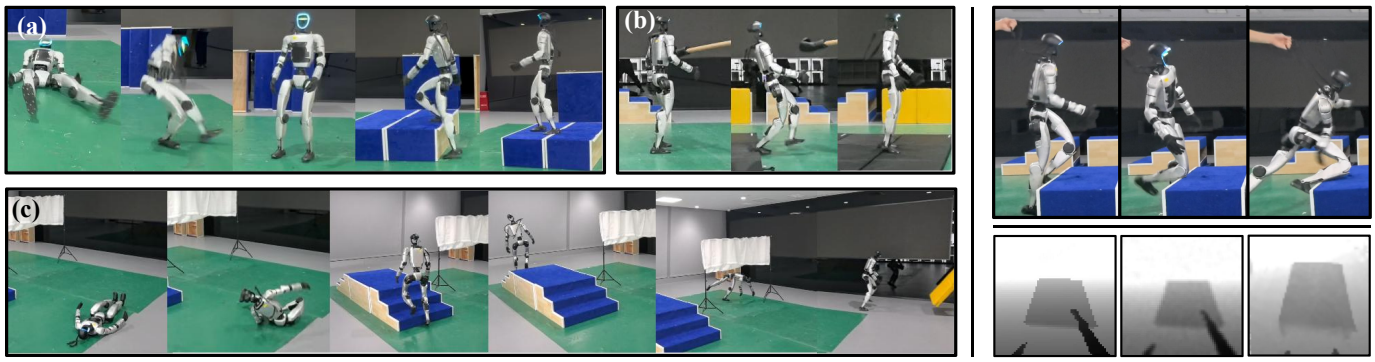


Fig. 7: **Left**: (a) fall recovery and platform traversal; (b) resilience to external disturbances; (c) a continuous sequence of recovery, stair climbing, and rolling under an overhead bar. **Right**: (Top) a failure case where the robot trips due to lack of camera randomization. (Bottom, from left to right) original depth in simulation, noisy depth, and processed real-world depth.

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. 7 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. More real-world performance analysis is provided in Appendix F.

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. 7. 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. 7, 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., 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 online RL, allowing the policy to explore and generalize beyond specialist demonstrations. Additionally, employing a hierarchical framework that includes a learning-based selection module and several specialized policies is another promising approach to further enhance the locomotive performance of humanoid robots.

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] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

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

[20] David Hoeller, Nikita Rudin, Dhionis Sako, and Marco Hutter. Anymal parkour: Learning agile navigation for quadrupedal robots. *Science Robotics*, 9(88):eadi7566, 2024.

[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. Exbody2: 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. Be-

yondmimic: 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] Prashanth Ravichandar, Lokesh Krishna, Nikhil Sobanbabu, and Quan Nguyen. Preferenced oracle guided multi-mode policies for dynamic bipedal loco-manipulation. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6600–6606. IEEE, 2025.

[42] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. 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*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

[43] 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.

[44] Nikita Rudin, Junzhe He, Joshua Aurand, and Marco Hutter. Parkour in the wild: Learning a general and extensible agile locomotion policy using multi-expert distillation and rl fine-tuning. *arXiv preprint arXiv:2505.11164*, 2025.

[45] 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.

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

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

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] Richard S Sutton, Andrew G Barto, et al. *Introduction*

to reinforcement learning, volume 135. MIT press Cambridge, 1998.

[53] Annan Tang, Takuma Hiraoka, Naoki Hiraoka, Fan Shi, Kento Kawaharazuka, Kunio Kojima, Kei Okada, and Masayuki Inaba. Humanmimic: Learning natural locomotion and transitions for humanoid robot via wasserstein adversarial imitation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13107–13114. IEEE, 2024.

[54] 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.

[55] 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.

[56] 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.

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

[58] 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.

[59] Weiji Xie, Jiakun Zheng, Jinrui Han, Jiyuan Shi, Weinan Zhang, Chenjia Bai, and Xuelong Li. Textop: Real-time interactive text-driven humanoid robot motion generation and control. *arXiv preprint arXiv:2602.07439*, 2026.

[60] Michael Xu, Yi Shi, KangKang Yin, and Xue Bin Peng. Parc: Physics-based augmentation with reinforcement learning for character controllers. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025.

[61] 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.

[62] 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.

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

[64] 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.

[65] 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.

[66] 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.

[67] 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.

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

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

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