Learning Adaptive High-Platform Traversal for Humanoid Robots

Author Names Omitted for Anonymous Review. Paper-ID 1122

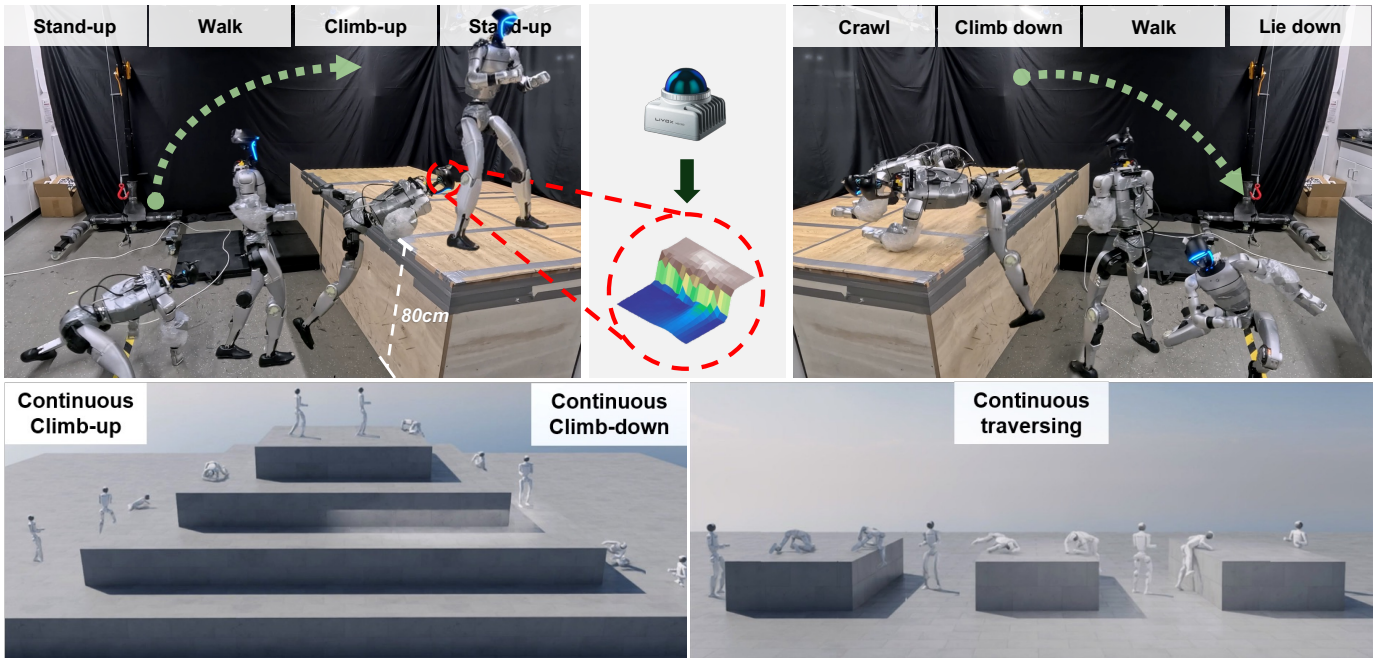


Fig. 1: The robot adaptively traverses high platforms of up to 80 cm (114% of leg length) by leveraging diverse full-body behaviors, including climb-up, climb-down, stand-up, lie-down. Enabled by LiDAR-based elevation mapping, the policy exhibits context-aware whole-body coordination, allowing continuous and robust traversal across challenging terrain in both simulation and the real world.

***Abstract*—Restricting humanoid robots to feet-based locomotion limits traversal in environments with tall vertical ledges. Existing learning systems can jump onto elevated structures, but achievable heights remain limited and high-impact dynamics make scaling to ledges exceeding leg length risky for real robots. A safer alternative is full-body climbing, which requires coordinating arms, torso, and legs through contact-rich, goal-reaching maneuvers. However, learning these behaviors without reference motions is difficult: phase-invariant tracking rewards are ill-defined, sparse terminal feedback impedes exploration, and strong safety regularization is necessary to prevent hardware damage. In this work, we present a sim-to-real framework for end-to-end high-platform traversal that learns full-body maneuvers without reference trajectories. Our key contribution is a ratchet-style progress reward that maintains a best-so-far task-space frontier and rewards only strict improvement beyond this historical best. This yields dense, task-aligned supervision while remaining velocity-free, enabling contact-aware exploration under strong physical constraints. Using this reward, we train LiDAR-based primitives for four maneuvers (climb-up, climb-down, stand-up, and lie-down) and distill them together with walking and crawling into a single perceptive policy that autonomously selects and transitions between skills from local elevation maps. Experiments on a 29-DoF Unitree G1 humanoid demonstrate robust zero-shot sim-to-real transfer, enabling reliable traversal of platforms up to 0.8m (114% of leg length) across variations in height and approach pose.**

I. Introduction

Locomotion is a fundamental capability for humanoid robots, yet has long remained challenging. Recent advances in deep reinforcement learning (DRL) have enabled robust, feet-based locomotion over uneven terrain [1, 2]. To further expand traversable terrain, prior systems have also learned whole-body jumping to get onto and off elevated structures [3, 4, 5]. However, jumping-based solutions typically achieve limited height (often below 63% of leg length), and scaling to substantially higher platforms (e.g., ledges or tables exceeding 100% of leg length) requires large impulsive torques and induces high-impact dynamics, which can exceed actuator limits and pose unacceptable risk in real-world deployment.

For such extreme heights, a more reliable alternative is full-body climbing, which coordinates arms, torso, and legs to create distributed supports to ascend or descend in a controlled manner. Building on climbing, complete high-platform traver-

sal involves multiple behaviors: climb-up and climb-down at vertical edges, walking or crawling on the platform, stand-up and lie-down for posture transition between prone and upright configurations. Despite its promise, learning and executing high-platform traversal presents two key challenges.

First, the four full-body maneuvers (climb-up, climb-down, stand-up, and lie-down) are difficult to learn with DRL. In contrast to cyclic, command-conditioned locomotion (e.g., walking, running, or crawling), where tracking objectives such as base velocity and periodic contacts provide dense supervision, these maneuvers are contact-rich and goal-reaching. Success is defined by satisfying terminal conditions through staged contact transitions and whole-body reconfiguration (e.g., moving the lower body and center of mass above the platform during climb-up). Their contact patterns and motion speeds vary across phases and depend on perceived geometry, making dense tracking-style rewards ill-defined; while naive formulations often yield brittle motion references or fail due to sparse reward signals and unstable exploration under safety constraints.

Second, complete traversal is a long-horizon *sequencing* problem. It requires the robot to (i) recognize which behavior is appropriate from local terrain observations and user commands, (ii) switch policies at the correct time and contact configuration, and (iii) remain stable during transitions where the support set and action distribution change abruptly (e.g., from climbing to walking, or from prone crawling to standing). These discontinuities make end-to-end learning with a single monolithic policy difficult: naive training often collapses to a subset of behaviors, while explicit switching heuristics are brittle and sensitive to small geometric variations.

To address these challenges, we present a learning framework for **adaptive high-platform traversal** that combines (i) reward shaping for contact-rich goal-reaching maneuvers and (ii) skill integration into a single perceptive controller. Our key idea is a two-stage pipeline. First, we learn a library of six terrain-conditioned skills: four goal-reaching full-body maneuvers (*climb-up, climb-down, stand-up, lie-down*) and two standard locomotion skills (walking, crawling). To make the goal-reaching maneuvers learnable without motion references, we introduce a generalized *ratchet progress* reward that maintains a self-updating best-so-far task state and penalizes the agent unless it strictly surpasses its historical best. This yields dense, task-aligned supervision while remaining speed-free, enabling exploration under strong safety regularization (e.g., contact-force and torque penalties) and preventing "retrace" exploitation. To make the perception more reliable, we leverage a LiDAR-based elevation mapping pipeline and mitigate its sim-to-real gap through both simulated signal imnoising and real signal processing, which enables the policy to tolerate localization drift, self-occlusion, and missing/ghost measurements that frequently arise during climbing and posture transitions. Second, we distill the six specialized teachers into a *single* student policy via behavior clone (BC) and DAgger, trained on a mixture of skill-focused and transition-focused environments. The resulting unified policy internalizes both skill selection and transition timing, eliminating brittle hand-designed switching rules.

We evaluate our system on a 29-DoF Unitree G1 humanoid in simulation and on hardware. In simulation, the distilled policy completes long-horizon traversal courses that chain acyclic maneuvers with periodic gaits under degraded perception and perturbations, achieving a 95.4% success rate over 1,000 trials. On hardware, the policy transfers zero-shot and executes continuous traversal loops on a $0.8\,\mathrm{m}$ table ($\approx 114.3\%$ of leg length), demonstrating robust behavior selection and smooth transitions across skills. We further show that the proposed ratchet progress reward is critical for acquiring adaptive climb-up behavior: common alternatives (velocity tracking, distance shaping, curiosity with sparse success, or incremental progress) either fail to solve the task or succeed only by producing unsafe, high-impact strategies.

In summary, our contributions are: (i) a unified RL training formulation for contact-rich humanoid maneuvers with LiDAR perception, (ii) a *ratchet progress* reward that provides dense supervision without prescribing a trajectory and is compatible with strong safety regularization, (iii) a teacher-student distillation pipeline that integrates six heterogeneous skills into a single traversal policy with robust, learned transitions, and (iv) extensive simulation and real-world results demonstrating adaptive traversal of extremely high platforms with zero-shot sim-to-real transfer.

II. Related Work

changyi:Tsingyuan: Table 1 for related works removed now for compiling Appendix.Kaiak

We review prior work in three aspects most relevant to our system, and summarize key distinctions in Tab. **??**.

A. Learning Feet-Based Locomotion

Deep reinforcement learning (DRL) has substantially improved the robustness and agility of legged locomotion, with early successes on quadrupeds [6, 7, 8, 9] and recent progress on humanoids. Humanoids with learned controllers can walk and run in uneven terrain [10, 11, 2], traverse sparse footholds [12, 13, 14], and step or jump into elevated structures [3, 4, 5]. However, these methods primarily rely on *feet-only* contacts, which requires large impulsive torques to reach high platforms, resulting in limited height (typically below 63% of leg length). In contrast, our system exploits whole-body, multi-contact coordination to distribute load and traverse platforms exceeding 114% of leg length.

B. Learning Humanoid Full-Body Maneuvers

Recent work has started to learn individual full-body maneuvers such as stand-up [15, 16, 17]. However, these methods often use task rewards that conflict with safety regularization and therefore require multi-stage pipelines or rely on heavy task-specific engineering (e.g., virtual-force shaping and carefully tuned curriculum for regularization and action rescaling), which has largely limited progress to relatively simple behaviors. In contrast, our generalized *ratchet progress*

reward supports single-stage RL training for multiple contact-rich, goal-reaching maneuvers, while jointly optimizing task completion and safety regularization.

A complementary line of work learns human-like behaviors (e.g., dancing, walking, crawling, and jumping) by training policies to track human motions [18, 19, 20, 21, 22, 23, 24] using dense imitation rewards [25]. Building on these foundations, OmniRetarget [26] enables full-body climbing by preserving robot-scene contact relationships during retargeting. However, motion-tracking approaches fundamentally rely on prerecorded trajectories and therefore require close alignment between the reference motion, the environment geometry, and the robot's initial state. This strongly limits adaptation to unseen terrain and perturbed initial conditions common in real deployment. Motion generation models [27, 28, 29] could in principle provide adaptive references, but generated climbing motions are often physics-infeasible and do not explicitly reason about multi-contact feasibility and deployment constraints (e.g. torque limit, contact force, generation speed, perception gap) ; consequently, they are typically validated only in simulation [30]. In contrast, our policy is perceptive, reference-free, and deployable: it learns terrain-conditioned strategies that generalize across platform heights and initial poses, enabling autonomous traversal in diverse real-world environments.

C. Policy Distillation for Legged Robots

Teacher-student distillation is widely used to train deployable policies for legged robots [31, 32, 33, 34, 35, 36]. A common paradigm trains a teacher with privileged simulation information and distills it into a student that relies only on onboard observations (e.g., depth images or tactile signals) for deployment. More recently, multi-expert distillation has been used to integrate terrain-conditioned skills into a single quadrupedal policy [37, 38], typically using DAgger-style [39] data aggregation, which we also adopt. However, existing multi-skill distillation has largely focused on quadrupeds, where skills share similar feet-contact modes and transitions occur near a nominal walking posture. In contrast, our teacher set spans heterogeneous humanoid behaviors, including full-body maneuvers and locomotion skills with substantially different state and action distributions (e.g., climbing, walking, crawling, and posture transitions). This substantially increases the difficulty of both RL and distillation training: the teacher skills must be trained with compatible terminal-state distributions to enable safe and smooth concatenation, and appropriate teacher actions should be provided conditioned not only on terrain geometry and user commands, but also on the robot state and transition progress.

III. High-Platform Traversal Policy Learning

Our goal is to learn a perceptive humanoid policy that can robustly traverse *extremely high* platforms in the real world. As introduced in Sec. I, such traversal requires multiple terrain-conditioned behaviors: four full-body maneuvers (*climb-up, climb-down, stand-up, lie-down*) and two standard locomotion skills (walking, crawling). To handle this diversity, we adopt a two-stage learning pipeline [35, 36]. First, we develop a unified RL training framework to learn the four full-body maneuver policies with LiDAR perception (Sec. III-B). To enable efficient learning of these contact-rich goal-reaching tasks, we introduce a generalized *ratchet progress* reward that provides dense supervision while supporting exploration under strong safety regularization (Sec. III-A). Second, we distill all six policies into a single policy that autonomously selects and transitions between behaviors based on perception, enabling end-to-end high-platform traversal (Sec. III-C).

A. Ratchet Progress Reward for Humanoid Maneuvers

1) Task Definition:

We model the full-body, contact-rich humanoid maneuver as a goal-reaching task, where success is defined by satisfying a terminal condition rather than tracking a reference trajectory. Let s_t denote the robot state at timestep t. We define a *task state* $x_t = \phi(s_t)$, where $\phi(\cdot)$ extracts a minimal set of variables needed to evaluate task completion. For each maneuver, we specify a target task state x^g and declare success when $x_t \geq x^g$, where $\geq$ denotes the ordering induced by the chosen task metric.

To instantiate the four maneuver objectives (Table I), we use the following notation. p_{CoM}, p_{head}, and p_{LB} denote the positions of the full-body center of mass, head, and lower body, respectively. We use h and x to denote environment- and pose-dependent thresholds, such as the platform edge height h_{edge} or the nominal standing head height h_{head}^{stand}. For standing stability, we define a balance margin $d_{bal} = \|p_{CoM} - \bar{p}_{feet}\|$ where $\bar{p}_{feet}$ is the geometric center of the feet. These definitions yield concise, task-specific terminal conditions while keeping x_t low-dimensional and easy to compute online.

TABLE I: Task Definition of Four Goal-Reaching Maneuvers

Task	Task State (x_t)	Target Task State (x^g)
Climb-up	p_{CoM}, p_{LB}	$p_{LB}^{(z)} > h_{edge} \wedge p_{CoM}^{(x)} > x_{edge}$
Climb-down	p_{CoM}, p_{LB}	$p_{LB}^{(z)} < h_{LB}^{stand} \wedge p_{CoM}^{(x)} < x_{edge}$
Stand-up	$p_{head}^{(z)}, d_{bal}$	$p_{head}^{(z)} > h_{head}^{stand} \wedge d_{bal} < \delta$
Lie-down	p_{CoM}, p_{head}	$p_{CoM}^{(z)} < h_{CoM}^{prone} \wedge p_{head}^{(z)} < h_{head}^{prone}$

2) Ratchet Progress Reward:

These goal-reaching maneuvers do not admit a phase-invariant predefined reference, such as a consistent velocity or contact pattern for command-conditioned locomotion. To provide a meaningful reference at every timestep without prescribing a motion template, we introduce a *self-updating task-space reference* that records the best progress achieved so far along the trajectory. This best-so-far task state at timestep t is defined as:

$$x_t^* = \max(x_0, x_1, \ldots, x_{t-1}) \tag{1}$$

which can be updated online via $x_t^* = \max(x_{t-1}^*, x_{t-1})$, with

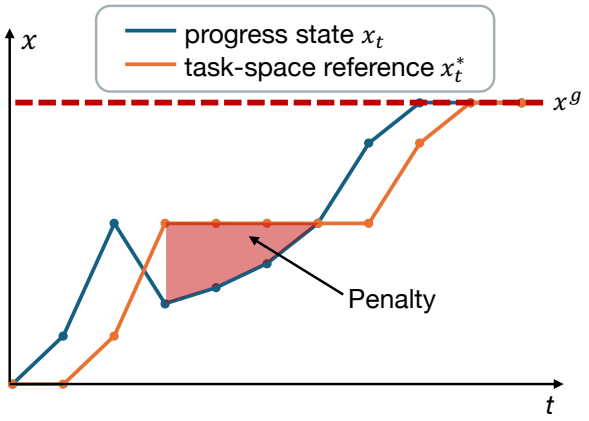


Fig. 2: Illustration of the evolution of x_t^* and the corresponding penalty computation.

$x_0^* = x_0$. Intuitively, x_t^* tracks the current frontier of task-space progress demonstrated by the agent.

Based on the best-so-far reference, we define a binary *ratchet-style* task reward, which is illustrated in fig 2:

$$r_t = \begin{cases} 0, & \text{if } x_t > x_t^*, \\ -1, & \text{otherwise.} \end{cases} \tag{2}$$

That is, the agent receives no penalty only when it *strictly surpasses* its historical best, and is penalized otherwise. Although simple, the above construction is tailored to contact-rich, goal-reaching maneuvers, with three key properties.

- **Dense task-aligned supervision.** The reward is evaluated at every timestep to penalize any failure to exceed the best-so-far progress. This provides a dense signal that keeps exploration within task-relevant behaviors, which is essential for contact-rich maneuvers where terminal-only rewards are too sparse to discover feasible contact sequences.
- **Speed-free progress enables exploration and deployment.** Because the reward depends only on *whether* progress improves but not *how much*, it does not encourage rushing in task space. This supports (i) *patient, contact-aware exploration*, allowing the robot to hold intermediate supports until necessary contacts become stable (e.g., during climb-up, keep one leg grounded until the other stably lands on the platform; during stand-up, hold torso ascent until limbs become load-bearing), and (ii) *effective regularization* of impact/torque/force can enforce safe motions without having to counteract a speed-driven task incentive.
- **History dependence prevents retracing exploits.** Incremental criteria such as $(x_t > x_{t-1})$ can be gamed by oscillating backward and forward. In contrast, our historical best criteria $(x_t > x_t^*)$ ensures optimization with genuine advancement toward the goal.

B. Learning Perceptive Full-Body Maneuvers

In this section, we describe our RL framework for learning the four full-body maneuver skills. The two standard locomotion skills (walking and crawling) are trained using a conventional velocity-tracking formulation and are therefore omitted for brevity.

1) RL Training Environment:

State, Observation, and Action. We train each maneuver as a single-skill policy in a Markov Decision Process (MDP). The observation space includes robot proprioception $s_t^{\text{proprio}} \in \mathbb{R}^{64}$ (gravity vector, base angular velocity, and joint positions/velocities), the previous action $a_{t-1} \in \mathbb{R}^{29}$, the task state x_t, and optionally a local elevation map $m_t \in \mathbb{R}^{441}$ at $0.05\,\text{m}$ resolution covering a $1 \times 1\,\text{m}^2$ area. All policies take a 6-step history of $(s_t^{\text{proprio}}, a_{t-1})$ to capture short-term dynamics. Because the climbing skills must perceive the platform geometry, the *climb-up* and *climb-down* policies additionally take m_t as input. All policies output target joint positions $a_t \in \mathbb{R}^{29}$, which are tracked by a low-level PD controller.

Our task reward depends on the best-so-far task state x_t^*, which evolves along the trajectory and is not contained in the instantaneous robot state. This introduces history dependence that can degrade value estimation if the critic observes only s_t. To mitigate this, we augment the critic input with x_t^*, enabling more accurate value prediction under the ratchet reward.

Simulation Environment. To learn robust behaviors, we extensively randomize the terrain configuration and initial conditions. The platform height is sampled from $[0.55\,\text{m},\ 0.85\,\text{m}]$. For *climb-up*, the initial distance from the robot base to the vertical surface and the initial yaw angle are sampled from $[0.15\,\text{m},\ 0.35\,\text{m}]$ and $[-60°,\ 60°]$; for *climb-down*, they are sampled from $[0.30\,\text{m},\ 0.45\,\text{m}]$ and $[-75°,\ 75°]$. To improve sim-to-real transfer, we apply a comprehensive suite of domain randomization following [24], including perturbations to base mass and CoM location, surface friction and restitution, joint defaults, initial base/joint states, and periodic impulsive velocity pushes to the base. We further apply symmetric augmentation [40] by mirroring states and actions, which reduces handedness bias and improves generalization across approach angles.

Initial Posture Sampling. Since single-skill policies are executed sequentially during distillation, their initial-state distributions must encompass the terminal states generated by preceding skills. As illustrated in Fig. 3, transitions predominantly occur around two canonical postures: standing (start of walking, climb-up, and lie-down; end of walking, climb-down, and stand-up) and prone (start of crawling, climb-down, and stand-up; end of crawling, climb-up, and lie-down). We define nominal joint configurations for these postures as q_{stand} and q_{prone}, respectively.

For each skill, initial joint angles are sampled by perturbing the corresponding nominal starting posture, ensuring that training begins from physically plausible states that are compatible with upstream transitions. To enable seamless switch-out between skills, we additionally shape the terminal behavior of full-body maneuvers toward the nominal ending posture using a terminal-pose reward (Sec. III-B2). If the reachable terminal-state distribution of a skill is not fully contained within the initial-state distribution of its successor, we subsequently retrain the successor skill while augmenting

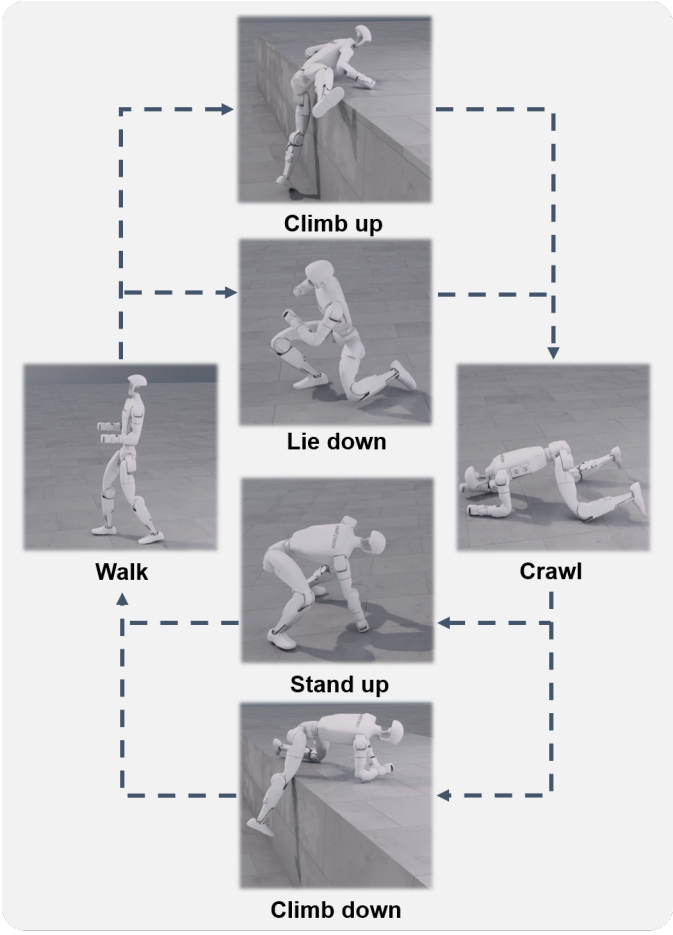


Fig. 3: Skill transitions available to the distilled policy. Walking and crawling serve as both essential moving skill and connector skills that links all full-body maneuvers.

its initial-state distribution to cover all possible terminal configurations produced by its predecessors.

2) Reward Design:

We define the total reward as the sum of five components:

$$r = r_{\text{alive}} + r_{\text{reg}} + r_{\text{force}} + r_{\text{task}} + r_{\text{tp}} \tag{3}$$

The first three terms r_{alive}, r_{reg}, and r_{force}, are shared across all skills. r_{alive} is a constant survival bonus that discourages early termination. r_{reg} aggregates standard regularization penalties that promote smooth and energy-efficient motions (e.g., action rate, joint velocity/acceleration/torque penalties).

Because full-body maneuvers involve frequent terrain contacts beyond the feet, limiting impact is critical for safe deployment. We therefore include a contact-force penalty r_{force} that grows rapidly once contact forces exceed a safe threshold:

$$r_{\text{force}} = -\left(\exp\left(\alpha \cdot \max(0, F_t - F_{limit})\right) - 1\right) \tag{4}$$

where F_t is the maximum contact force at timestep t, F_{limit} is a safety threshold, and $\alpha > 0$ controls the penalty scale. Specially, we set $F_{limit} = 0$ for the head link, since even light head contact can destabilize the head-mounted LiDAR and severely degrade perception.

The remaining two terms, r_{task} and r_{tp}, are task-specific but require only minimal specification. r_{task} is the ratchet progress reward introduced in Sec. III-A, which drives goal completion. r_{tp} encourages a desired terminal posture to facilitate reliable behavior transitions. It is activated *only* after the goal is reached and within the final second of an episode:

$$r_{\text{tp}} = \mathbb{I}_{(t>H-1s)} \cdot \mathbb{I}_{(\text{goal reached})} \cdot \exp(-\beta \|q_t - q_{tar}\|^2) \tag{5}$$

where $\mathbb{I}$ is the indicator function, H is the episode duration, q_t and q_{tar} denote the current and desired terminal joint angles of the robot, and $\beta > 0$ is a scale parameter.

3) Robust Perception via Elevation Mapping: Effective full-body adaptation requires high-fidelity spatial awareness and high perception accuracy. To achieve this, we implement a LiDAR-based elevation mapping pipeline based on [1], utilizing [41] for real-time localization and [42] for centralized map generation. However, dynamic maneuvers inevitably induce localization drift and sensor occlusion: contact-induced jitters and rapid accelerations accumulate odometry drift, the robot's own limbs create spurious "ghost" point batches, uncommon robot poses often result in severe FOV limitations, and the inherent probabilistic nature of elevation mapping introduces uncertainty. To bridge the resulting perception gap, we employ a dual strategy as shown in Fig. 4:

- **Real-World Post-processing:** We mitigate raw sensor degradation through a spatial outlier filter to block high-variance noise batches and an inpainting algorithm from [43] to fill occluded "NaN" regions, ensuring the policy receives a structurally continuous terrain representation.
- **Simulation Artifact Modeling:** We explicitly model three types of perceptual artifacts during training to enforce robustness against imperfect perception: per-cell Gaussian noise to mimic the probabilistic noise of mapping, coordinate offsets to model localization drift, and injected outlier clusters to simulate spurious "ghost" artifacts and points.

C. Policy Distillation for Skill Integration

To synthesize a unified, context-aware controller from our diverse skill library, we employ a two-stage teacher-student distillation pipeline inspired by [36]. The student policy is first pre-trained via Behavior Cloning (BC) [44] using a dataset of state-action pairs collected from specialized teacher policies. This is followed by refinement using DAgger [39] to mitigate the distributional shift and expand data coverage by incorporating student-generated trajectories into the training loop. We dynamically select the appropriate teacher based on the robot's state, privileged information, and commanded velocity, utilizing the Mean Squared Error (MSE) between the student and teacher actions as the primary supervision signal.

Leveraging massively parallel simulation environments, we categorize training environments to specifically target either a core skill (e.g., walking or crawling) or a critical skill

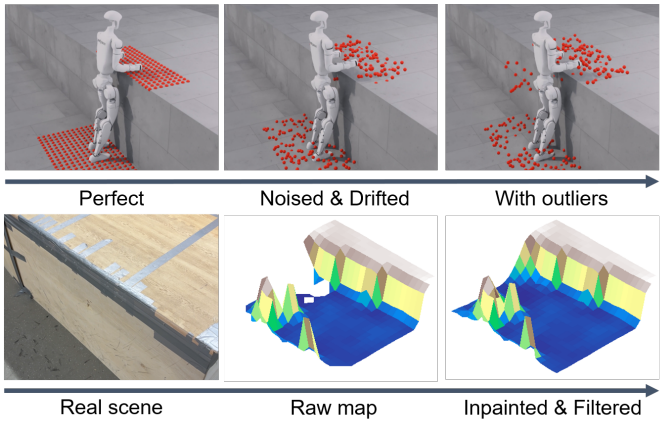


Fig. 4: The sim-to-real gap in LiDAR mapping is addressed through a dual approach that combines artifact modeling in simulation with real-world post-processing.

transition. Compared to sequential rollouts on predefined routes — where data collection for downstream skills is often bottlenecked by the failure rates of preceding maneuvers — this distributed data collection strategy alleviates uneven data distribution and facilitates more generalized distillation. By decoupling skills from a fixed route, the student policy is encouraged to learn fundamental transition rules rather than overfitting to a specific sequence of states.

To enhance the robustness of the student policy and narrow the sim-to-real gap, we incorporate domain randomization, elevation map degradation, and Gaussian action noise during the distillation process. The resulting distilled policy proactively adapts its locomotor modality to the perceived terrain geometry , achieving seamless and robust transitions across all six diverse teacher skills.

IV. Experiments

In this section, we present a series of qualitative and quantitative evaluations to address the following questions:

1) Does the proposed system enable long-sequence traversal across extreme high platform in face of noisy perception and the highly disparate action distributions across diverse skills (Sec. IV-A)?
2) Do the acquired policies exhibit adaptability and robustness when interacting with varied environments (Sec. IV-B)?
3) How does the proposed progress reward formulation effectively facilitate the acquisition of full-body maneuvers (Sec. IV-C)?

We conduct all experiments with a 29-DoF Unitree G1 humanoid robot in both simulation and on hardware. Simulation environments are implemented in NVIDIA Isaac Sim. For real-world deployment, terrain perception is handled by an Intel Core i7 CPU, which processes data from a Livox MID-360 LiDAR to generate real-time elevation maps.

A. Integrated System Performance for Platform Traversal

The system enables continuous high-agility traversal; We first designed three challenging simulation courses that require sequential execution of the robot's full locomotor repertoire, as illustrated in Fig. 1. These courses—Continuous Traversing, Ascending, and Descending—chain multiple acyclic maneuvers (e.g., climbing, standing-up) and periodic gaits (e.g., walking) into single, cohesive sequences.

The robot is consistently commanded via a simple set of velocity commands. To evaluate the performance of the system performance, we induce environmental perturbations and artifacted perception to simulate realistic constraints, including LiDAR degradation and drift of state estimation. Under these conditions, the robot demonstrates sustained robustness, autonomously adjusting its configuration to overcome varying terrain geometries without losing balance. Following a set of predefined command sequences, the policy achieves a success rate of 95.4% over 1,000 trials.

Zero-shot transfer to long-sequence real-world deployment. As illustrated in Fig. 1, we further validate the policy in the real-world via zero-shot sim-to-real transfer. We tasked the robot with a continuous climb-up, climb-down, stand-up and lie-down sequence with a 0.8m table, as we are unable to deploy other routes displayed above due to the facility constraints. Our policy enables the robot to successfully complete two consecutive full-loop traversals without failure. Notably, the policy adopts context-aware motor strategies, such as utilizing different lead legs based on the approach angle, highlighting its adaptability during dynamic deployment in the real-world.

B. Adaptability and Robustness to Varying Environments

Single policies exhibits strong robustness and adaptability. To rigorously quantify the robustness of each full-body maneuver, we evaluate the proposed method across a diverse set of challenging scenarios in simulation. To ensure statistical significance, each task is assessed over 1,000 independent trials, while maintaining identical terrain distributions, domain randomizations, and pose initializations as those used during training. Success Rate (SR) is adopted as the primary performance metric. As summarized in Table II, the specialized teacher policies achieve near-perfect success rates across all simulated tasks. In addition, the recorded maximum contact forces (M.C.F.) remain within safe limits for all skills, validating the effectiveness of the proposed contact-force regularization under varied terrain geometries and initial conditions.

We further validate the system on hardware by subjecting the climb-up policy to nine distinct real-world configurations, varying platform heights from 0.6,m to 0.8,m (approximately 114.3% of leg length) and approach angles from $-\pi/4$ to $\pi/4$. Across these experiments, the policy exhibits consistent reliability, including in extreme cases such as a 65° approach angle, which lies entirely outside the training distribution. In such scenarios, rather than executing a naive forward reach, the robot displays coordinated whole-body behavior: it reorients

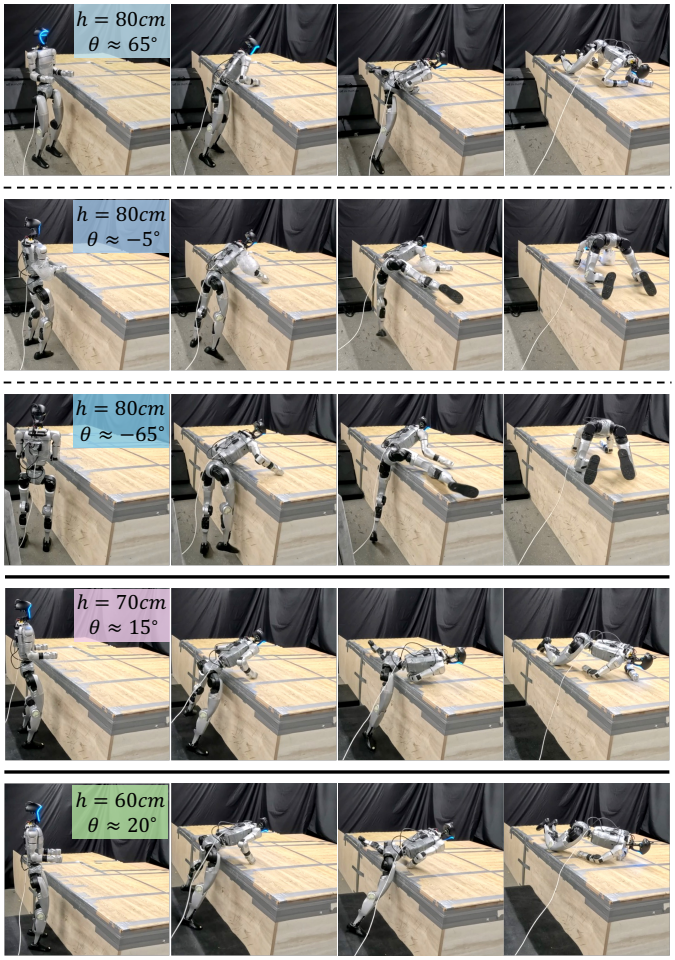


Fig. 5: Robust adaptation of the climb-up policy to geometric and perceptual variations. Our policy adapts to diverse platform heights and approach angles despite noisy LiDAR perception. The system remains effective in extreme edge cases—such as a 0.8,m platform (≈114.3% of leg length) and a steep 65° inclination—demonstrating strong zero-shot transfer ability to out-of-distribution (OOD) scenarios.

its torso to face the platform and leverages full-body motion to initiate the climb. Representative trials are shown in Fig. 5. We further evaluate the real-world performance of additional full-body maneuvers, including climb-down, stand-up, and lie-down; as shown in Fig. 6, all achieve a 5/5 success rate. Collectively, these results underscore the robustness and strong sim-to-real transfer capability of policies trained using the proposed progress-based reward formulation.

Symmetry augmentation facilitates the development of balanced behaviors. By incorporating symmetry augmentation during training, the policy converges to an emergent balanced strategy rather than a biased "handedness". During climb-up maneuvers, the lead leg is dynamically selected based on the robot's relative heading to the platform. In practice, this balanced motion is critical, as biases significantly restrict a robot's feasible workspace and degrade climbing

TABLE II: Comparative Success Rates of Single Skill in Sim and Real. **SR:** Success Rate; **S/T:** Success / Trials; **M.C.F.:** Max Contact Force.

	Simulation		Real World			
Task	**SR (%)**	**M.C.F. (N)**	H **(m)**	A (°)	**S/T**	**SR (%)**
Climb-up (w/ all)	98.8	638 ± 479	0.6	$[-45, -15]$	5/5	97.8
			0.6	$[-15, +15]$	5/5	
			0.6	$[+15, +45]$	5/5	
			0.7	$[-45, -15]$	4/5	
			0.7	$[-15, +15]$	5/5	
			0.7	$[+15, +45]$	5/5	
			0.8	$[-45, -15]$	5/5	
			0.8	$[-15, +15]$	5/5	
			0.8	$[+15, +45]$	5/5	
Climb-up (w/o drift&outlier) (w/ post-process)	-	-	0.8	$[-45, +45]$	0/5	0.0
Climb-up (w/ drift&outlier) (w/o post-process)	-	-	0.8	$[-45, +45]$	3/5	60.0
Climb-down	99.9	754 ± 241	0.8	$[-45, +45]$	5/5	100.0
Stand-up	99.5	632 ± 222	–	–	5/5	100.0
Lie-down	100.0	576 ± 125	–	–	5/5	100.0

performance.

Bridging the elevation mapping sim-to-real gap is essential for policy performance. To isolate the impact of our perception pipeline, we conducted an ablation study by selectively omitting specific mapping processing stages. As shown in Table II, while the full system maintains nearly 100% success, a policy trained without simulated drift and outlier modeling fails completely on hardware. Furthermore, deploying the robust policy without real-time filtering and inpainting degrades the success rate to 60%. These findings highlight that modeling sensor artifacts during training is necessary to tolerate mapping uncertainty decrease OOD cases by expanding the data distribution. Real-time map reconstruction is vital to prevent the policy from encountering erratic data, such as outlier batches or NaN holes, that may leads to catastrophic failure.

C. Skill Acquisition via Progress Rewards

To evaluate the efficacy of progress-based rewards in learning adaptive full-body maneuvers, we benchmark the training of the Climb-up skill against several baseline formulations. In each baseline, the progress-based reward is replaced by alternative task rewards while maintaining identical weights and hyperparameters. Detailed configurations are provided in the Supplementary Materials.

1) Velocity: Employs a standard velocity-tracking objective aimed at matching a torso velocity command[45, 1, 14]. This command is defined in the world frame and oriented forward toward the platform.

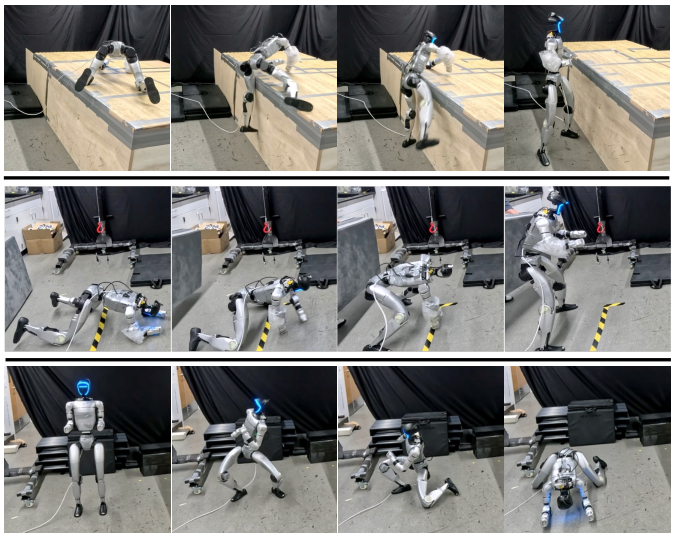


Fig. 6: Other three full-body maneuvers deployed in real robots: Climb-down, Stand-up and Lie-down.

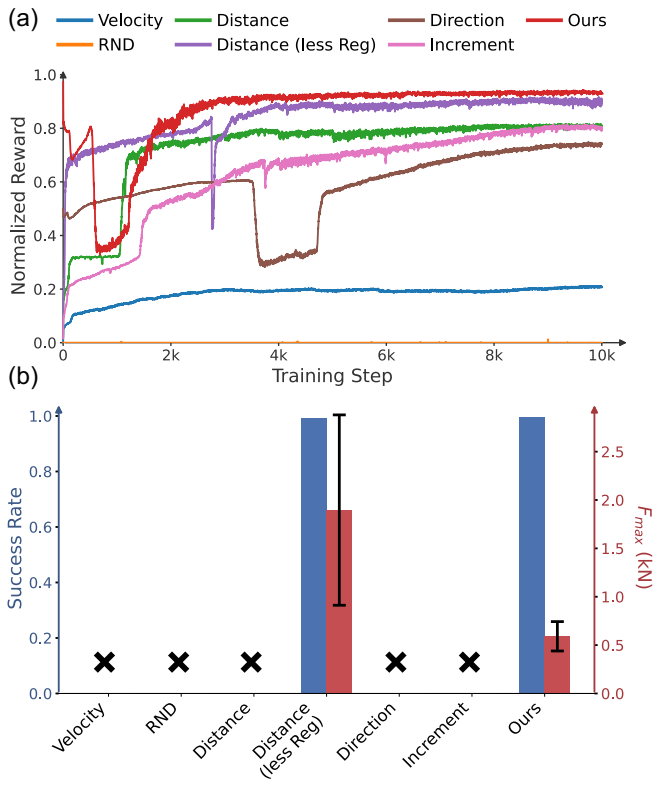


Fig. 7: Comparative analysis of reward formulations. (a) Learning curves for normalized task rewards. (b) Success rate (blue) and maximum contact force (red) of the trained policies.

2) RND (Random Network Distillation [46]): Utilizes intrinsic rewards generated via RND to incentivize exploration, combined with a sparse task reward upon task completion. [47, 48].
3) Distance: Penalizes the distance to the target, encouraging the agent to minimize this gap at every timestep.
4) Distance (less Reg.): Follows the same distance-minimization objective as above but with significantly lower regularization penalties.
5) Direction: Rewards any base velocity in the direction towards the goal, while penalizing small velocity to prevent stalling [49, 50].
6) Increment: Rewards the difference between the previous and the current distance of the system state from the goal state [51].

changyi:consider adding a table to list the formulations of baseline rewards in a concise format

We report the learning curves for task rewards, alongside the success rates and maximum contact forces for each formulation. To ensure statistical significance and assess robustness, each policy is evaluated over 1,000 independent trials across randomized environment configurations. These results are summarized in Fig 7.

The baseline 1) failed 'Climb-up' entirely and is stuck at the edge. The task rewards (velocity tracking) can not be optimized either. Because it over-constrains the motion to a fixed speed, the robot cannot discover the adaptive velocity needed to clear the edge. For other baselines, the task rewards are optimized better, yet the task performance is still poor.

The baseline 2) (RND + sparse reward) fails as a result of its inability to provide structured guidance for high-precision, multi-step coordination. While curiosity-driven learning is designed to discover diverse behaviors, it lacks a directional gradient, often leading the agent to exhaust its curiosity by exploring novel but physically irrelevant states—such as flailing in the lower state-space—without ever bridging the gap to the final goal. Consequently, in the absence of a dense reward like our progress-based reward signal, the agent is unlikely to randomly execute the specific sequence required to trigger the sparse success reward within the training budget.

The distance-minimization objective induces a significant velocity bias, as the agent seeks to maximize cumulative returns by approaching the target as rapidly as possible. This objective conflicts directly with contact force regularization, preventing baseline 3) from discovering a policy that balances task success with physical safety. In contrast, while reducing regularization in baseline 4) allows the policy to successfully reach the goal, it results in an aggressive "jumping" behavior toward the goal. As illustrated in Fig 7(b), this lack of constraint leads to excessively large contact forces.

Baselines 5) and 6) both fail by becoming trapped in a local optimum at the platform's edge, where the agent exhibits a repetitive "back-and-forth" motion—abruptly retreating only to slowly advance again. This failure mode occurs because, unlike our history-dependent progress-based reward , these formulations rely on instantaneous velocity direction. By rewarding any forward progress toward the goal, the reward structure incentivizes the agent to maximize the duration of its forward-moving phases; consequently, the agent learns to periodically reset its position to move forward indefinitely, accumulating reward without ever attempting the difficult

maneuver required to actually summit the platform.

V. Conclusion

We presented a sim-to-real framework for end-to-end high-platform traversal with humanoid robots, targeting extreme ledges where jumping becomes unsafe and actuator-limited. Our approach decomposes traversal into four contact-rich full-body maneuvers (climb-up, climb-down, stand-up, and lie-down) and two cyclic locomotion skills (walking and crawling), and unifies them through distillation into a single perceptive policy that autonomously selects and transitions between behaviors from local elevation maps.

A central contribution is the ratchet-style progress reward, which provides dense, task-aligned supervision for goal-reaching maneuvers without requiring reference trajectories or velocity tracking. By rewarding only strict improvement beyond a best-so-far task-space frontier, the reward remains velocity-free, prevents retracing exploits, and supports contact-aware exploration under strong safety regularization. We further showed that bridging the sim-to-real gap of elevation map is essential for reliable deployment. Experiments in simulation and on a 29-DoF Unitree G1 humanoid demonstrate robust long-horizon traversal and zero-shot real-world transfer, enabling reliable traversal of platforms up to 0.8m (114% of leg length) across variations in height and approach pose.

References

[1] J. Long, J. Ren, M. Shi, Z. Wang, T. Huang, P. Luo, and J. Pang, "Learning humanoid locomotion with perceptive internal model," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 9997–10 003.

[2] Y. Xue, W. Dong, M. Liu, W. Zhang, and J. Pang, "A unified and general humanoid whole-body controller for fine-grained locomotion," *arXiv e-prints*, pp. arXiv–2502, 2025.

[3] Z. Zhuang, S. Yao, and H. Zhao, "Humanoid parkour learning," *arXiv preprint arXiv:2406.10759*, 2024.

[4] Q. Zhang, P. Cui, D. Yan, J. Sun, Y. Duan, G. Han, W. Zhao, W. Zhang, Y. Guo, A. Zhang *et al.*, "Whole-body humanoid robot locomotion with human reference," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 11 225–11 231.

[5] Q. Ben, B. Xu, K. Li, F. Jia, W. Zhang, J. Wang, J. Wang, D. Lin, and J. Pang, "Gallant: Voxel grid-based humanoid locomotion and local-navigation across 3d constrained terrains," *arXiv preprint arXiv:2511.14625*, 2025.

[6] J. Tan, T. Zhang, E. Coumans, A. Iscen, Y. Bai, D. Hafner, S. Bohez, and V. Vanhoucke, "Sim-to-real: Learning agile locomotion for quadruped robots," *arXiv preprint arXiv:1804.10332*, 2018.

[7] J. Hwangbo, J. Lee, A. Dosovitskiy, D. Bellicoso, V. Tsounis, V. Koltun, and M. Hutter, "Learning agile and dynamic motor skills for legged robots," *Science Robotics*, vol. 4, no. 26, p. eaau5872, 2019.

[8] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Science robotics*, vol. 5, no. 47, p. eabc5986, 2020.

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

[10] I. Radosavovic, T. Xiao, B. Zhang, T. Darrell, J. Malik, and K. Sreenath, "Real-world humanoid locomotion with reinforcement learning," *Science Robotics*, vol. 9, no. 89, p. eadi9579, 2024.

[11] X. Gu, Y.-J. Wang, X. Zhu, C. Shi, Y. Guo, Y. Liu, and J. Chen, "Advancing humanoid locomotion: Mastering challenging terrains with denoising world model learning," *arXiv preprint arXiv:2408.14472*, 2024.

[12] H. Wang, Z. Wang, J. Ren, Q. Ben, T. Huang, W. Zhang, and J. Pang, "Beamdojo: Learning agile humanoid locomotion on sparse footholds," *arXiv preprint arXiv:2502.10363*, 2025.

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

[14] J. He, C. Zhang, F. Jenelten, R. Grandia, M. Bächer,

and M. Hutter, "Attention-based map encoding for learning generalized legged locomotion," *Science Robotics*, vol. 10, no. 105, p. eadv3604, 2025.

[15] P. Chen, Y. Wang, C. Luo, W. Cai, and M. Zhao, "Hifar: Multi-stage curriculum learning for high-dynamics humanoid fall recovery," *arXiv preprint arXiv:2502.20061*, 2025.

[16] X. He, R. Dong, Z. Chen, and S. Gupta, "Learning getting-up policies for real-world humanoid robots," *arXiv preprint arXiv:2502.12152*, 2025.

[17] T. Huang, J. Ren, H. Wang, Z. Wang, Q. Ben, M. Wen, X. Chen, J. Li, and J. Pang, "Learning humanoid standing-up control across diverse postures," *arXiv preprint arXiv:2502.08378*, 2025.

[18] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, "Expressive whole-body control for humanoid robots," *arXiv preprint arXiv:2402.16796*, 2024.

[19] M. Ji, X. Peng, F. Liu, J. Li, G. Yang, X. Cheng, and X. Wang, "Exbody2: Advanced expressive humanoid whole-body control," *arXiv preprint arXiv:2412.13196*, 2024.

[20] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "Humanplus: Humanoid shadowing and imitation from humans," *arXiv preprint arXiv:2406.10454*, 2024.

[21] T. He, W. Xiao, T. Lin, Z. Luo, Z. Xu, Z. Jiang, J. Kautz, C. Liu, G. Shi, X. Wang *et al.*, "Hover: Versatile neural whole-body controller for humanoid robots," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 9989–9996.

[22] Y. Ze, Z. Chen, J. P. Araújo, Z.-a. Cao, X. B. Peng, J. Wu, and C. K. Liu, "Twist: Teleoperated whole-body imitation system," *arXiv preprint arXiv:2505.02833*, 2025.

[23] Z. Chen, M. Ji, X. Cheng, X. Peng, X. B. Peng, and X. Wang, "Gmt: General motion tracking for humanoid whole-body control," *arXiv preprint arXiv:2506.14770*, 2025.

[24] Q. Liao, T. E. Truong, X. Huang, Y. Gao, G. Tevet, K. Sreenath, and C. K. Liu, "Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion," *arXiv preprint arXiv:2508.08241*, 2025.

[25] X. B. Peng, P. Abbeel, S. Levine, and M. Van de Panne, "Deepmimic: Example-guided deep reinforcement learning of physics-based character skills," *ACM Transactions On Graphics (TOG)*, vol. 37, no. 4, pp. 1–14, 2018.

[26] L. Yang, X. Huang, Z. Wu, A. Kanazawa, P. Abbeel, C. Sferrazza, C. K. Liu, R. Duan, and G. Shi, "Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction," *arXiv preprint arXiv:2509.26633*, 2025.

[27] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano, "Human motion diffusion model," *arXiv preprint arXiv:2209.14916*, 2022.

[28] Z. Jiang, Y. Xie, J. Li, Y. Yuan, Y. Zhu, and Y. Zhu, "Harmon: Whole-body motion generation of humanoid robots from language descriptions," *arXiv preprint arXiv:2410.12773*, 2024.

[29] J. Li, J. Cao, H. Zhang, D. Rempe, J. Kautz, U. Iqbal, and Y. Yuan, "Genmo: A generalist model for human motion," *arXiv preprint arXiv:2505.01425*, 2025.

[30] M. Xu, Y. Shi, K. Yin, and X. B. 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*, 2025, pp. 1–11.

[31] A. Kumar, Z. Fu, D. Pathak, and J. Malik, "Rma: Rapid motor adaptation for legged robots," *arXiv preprint arXiv:2107.04034*, 2021.

[32] T. Miki, J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning robust perceptive locomotion for quadrupedal robots in the wild," *Science robotics*, vol. 7, no. 62, p. eabk2822, 2022.

[33] Z. Fu, X. Cheng, and D. Pathak, "Deep whole-body control: learning a unified policy for manipulation and locomotion," in *Conference on Robot Learning*. PMLR, 2023, pp. 138–149.

[34] X. Cheng, K. Shi, A. Agarwal, and D. Pathak, "Extreme parkour with legged robots," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 11 443–11 450.

[35] Y. Yang, G. Shi, C. Lin, X. Meng, R. Scalise, M. G. Castro, W. Yu, T. Zhang, D. Zhao, J. Tan *et al.*, "Agile continuous jumping in discontinuous terrains," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 10 245–10 252.

[36] C. Lin, Y. R. Song, B. Huo, M. Yu, Y. Wang, S. Liu, Y. Yang, W. Yu, T. Zhang, J. Tan *et al.*, "Locotouch: Learning dexterous quadrupedal transport with tactile sensing," *arXiv preprint arXiv:2505.23175*, 2025.

[37] Z. Zhuang, Z. Fu, J. Wang, C. Atkeson, S. Schwertfeger, C. Finn, and H. Zhao, "Robot parkour learning," *arXiv preprint arXiv:2309.05665*, 2023.

[38] N. Rudin, J. He, J. Aurand, and M. 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.

[39] S. Ross, G. Gordon, and D. 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*. JMLR Workshop and Conference Proceedings, 2011, pp. 627–635.

[40] M. Mittal, N. Rudin, V. Klemm, A. Allshire, and M. Hutter, "Symmetry considerations for learning task symmetric robot policies," 2024. [Online]. Available: https://arxiv.org/abs/2403.04359

[41] W. Xu, Y. Cai, D. He, J. Lin, and F. Zhang, "Fast-lio2: Fast direct lidar-inertial odometry," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2053–2073, 2022.

[42] P. Fankhauser, M. Bloesch, C. Gehring, M. Hutter, and R. Siegwart, "Robot-centric elevation mapping with uncertainty estimates," in *Mobile Service Robotics*. World Scientific, 2014, pp. 433–440.

[43] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. " O'Reilly Media, Inc.", 2008.
[44] F. Torabi, G. Warnell, and P. Stone, "Behavioral cloning from observation," *arXiv preprint arXiv:1805.01954*, 2018.
[45] Z. Zhuang, S. Yao, and H. Zhao, "Humanoid parkour learning," 2024. [Online]. Available: https://arxiv.org/abs/2406.10759
[46] Y. Burda, H. Edwards, A. Storkey, and O. Klimov, "Exploration by random network distillation," 2018. [Online]. Available: https://arxiv.org/abs/1810.12894
[47] C. Schwarke, V. Klemm, M. v. d. Boon, M. Bjelonic, and M. Hutter, "Curiosity-driven learning of joint locomotion and manipulation tasks," in *Proceedings of The 7th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Tan, M. Toussaint, and K. Darvish, Eds., vol. 229. PMLR, 06–09 Nov 2023, pp. 2594–2610. [Online]. Available: https://proceedings.mlr.press/v229/schwarke23a.html
[48] C. Zhang, W. Xiao, T. He, and G. Shi, "Wococo: Learning whole-body humanoid control with sequential contacts," 2024. [Online]. Available: https://arxiv.org/abs/2406.06005
[49] D. Hoeller, N. Rudin, D. Sako, and M. Hutter, "Anymal parkour: Learning agile navigation for quadrupedal robots," *Science Robotics*, vol. 9, no. 88, p. eadi7566, 2024.
[50] X. Cheng, K. Shi, A. Agarwal, and D. Pathak, "Extreme parkour with legged robots," 2023. [Online]. Available: https://arxiv.org/abs/2309.14341
[51] OpenAI, I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, J. Schneider, N. Tezak, J. Tworek, P. Welinder, L. Weng, Q. Yuan, W. Zaremba, and L. Zhang, "Solving rubik's cube with a robot hand," 2019. [Online]. Available: https://arxiv.org/abs/1910.07113

VI. Appendix

A. Effectiveness of Progress-based Reward

We provide additional analysis to further illustrate the effect of our proposed progress-based reward. As Fig. 7 shows, among the baselines, only the distance-based reward with reduced regularization can reach high platforms with high success rate. However, this does not imply that it has learned *climbing*. As shown in Fig. 8, the baseline primarily succeeds via "full-body jumping": fast, impulsive motions with minimal sustained contact on the platform. This strategy produces extremely large contact forces and joint torques, making it physically infeasible and unsafe for real hardware deployment. In contrast, our progress-based reward leads to a qualitatively different solution that exhibits sustained whole-body coordination. The learned policy maintains stability by actively modulating contact locations and distributing interaction forces across multiple body segments throughout the maneuver. This continuous multi-point support enables the robot to complete climb-up while substantially reducing peak contact forces and average joint torques, yielding a motion profile that is compatible with real-world execution.

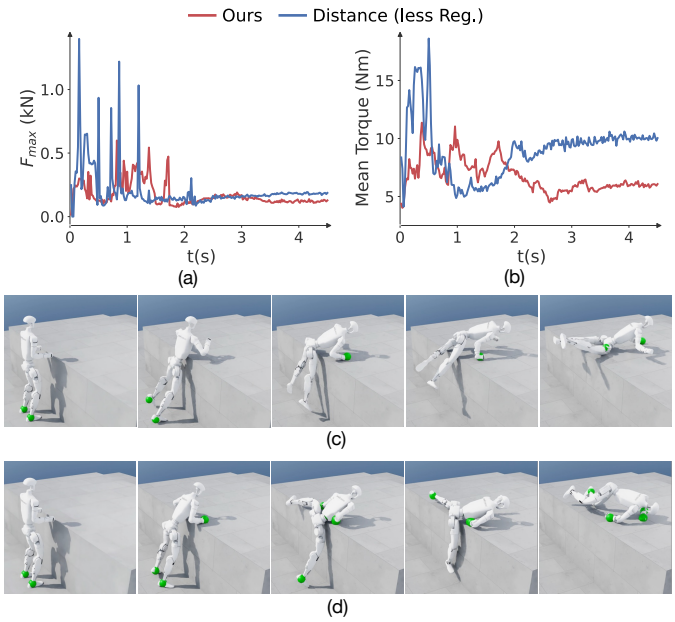


Fig. 8: Comparison against the baseline. (a) Max contact force over body parts w.r.t. time. (b) Mean joint torque w.r.t time. (c)(d) Keyframes of policy trained separately with distance-based reward / our proposed reward. The actual contact points at each timestep are visualized with the green sphere.

We further deploy our policy on hardware and record the torso horizontal displacement (x-direction) using a motion capture system (MoCap). We do not deploy the baseline policy due to the high risk associated with its excessive contact forces and impulsive movements. As illustrated in Fig. 9, the measured trajectory reveals two characteristic properties of functional climb-up behavior: monotone task progress with contact-induced holds. After the initial approach and hand placement (0–$0.7\,\mathrm{s}$), the trajectory exhibits a pronounced plateau centered around $t \approx 1.0\,\mathrm{s}$. This stagnation phase is functionally necessary: the torso remains near the platform edge while the robot lifts and securely places the lead leg. Once this contact is established, the torso resumes forward progression ($t > 1.2\,\mathrm{s}$), driven by coordinated forces from the hands and the newly established foothold. The emergence of this deliberate pause highlights the event-driven nature of contact-rich maneuvers and indicates that our reward formulation learns to prioritize kinematic feasibility and stability, rather than simply minimizing distance to the goal.

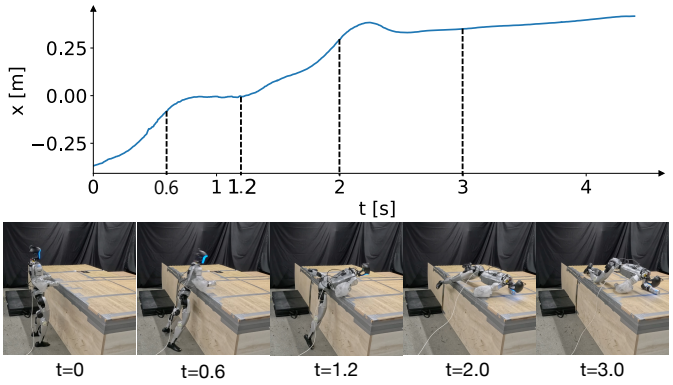


Fig. 9: Trajectory of the robot's torso relative to the platform edge. The top plot shows the horizontal displacement $x(t)$ over time, while the bottom sequence illustrates the corresponding climbing up motion at key timestamps.

We additionally summarize the task reward formulations used by baseline methods in Tab. III.

Note that for Velocity, $v_{cmd} \in [0.5, 1.0]$ m/s in the world frame. p_{goal} is a target 0.5 m inside the platform directly in front of the robot's initial position. Dist. (less Reg.) uses the same task rewards as Distance but reduces the weight of contact force penalties by a factor of 10.

B. Extended Experiments on Context-Aware Traversal

As shown in Fig. 10, we additionally demonstrate two new routes for high-platform traversal in the real world to further validate the context-aware capability of our proposed system. In both routes, the robot autonomously transitions between full-body maneuvers by perceiving the environmental geometry. For instance, when the robot is commanded to walk towards the platform, it perceives the obstacle and automatically triggers the climb-up skill to ascend. Similarly, when commanded to move towards the edge, the system perceives the drop and autonomously initiates the climb-down sequence to descend and reach a stable standing posture on the ground.

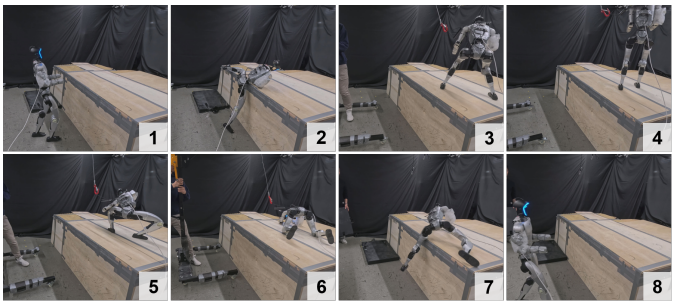
Route (a) includes the sequential execution of all six full-body maneuvers in the following order: walk (on the ground), climb-up, crawl, stand-up, walk (on the platform), lie-down, crawl, climb-down, and walk (on the ground). Our system

TABLE III: Formulations of the baseline task rewards.

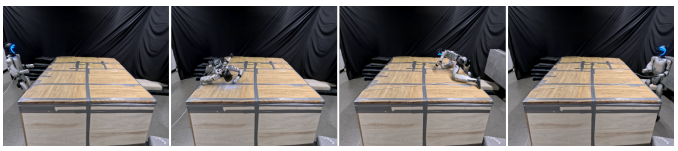
Baseline	Reward	Formulation	Weight
Velocity	Lin. Vel. Tracking	$\exp(-\lVert v_{base,t}^{(xy)} - v_{cmd}^{(xy)} \rVert^2 / 0.5^2)$	4
	Ang. Vel. Tracking	$\exp(-(\omega_{base,t}^{(z)} - \omega_{cmd}^{(z)})^2 / 0.5^2)$	4
RND	Sparse Success	$\mathbf{1}_{\{p_{LB,t}^{(z)} > h_{edge} \wedge p_{CoM,t}^{(x)} > x_{edge}\}}$	8
Distance	L2 Tracking	$(1 + \lVert p_{base,t}^{(xy)} - p_{goal}^{(xy)} \rVert^2)^{-1}$	8
Dist. (less Reg.)	L2 Tracking	$(1 + \lVert p_{base,t}^{(xy)} - p_{goal}^{(xy)} \rVert^2)^{-1}$	8
Direction	Cosine Sim.	$\cos(\theta(v_{base,t}, p_{goal} - p_{base,t}))$	8
Increment	Height Incr.	$\mathbf{1}_{\{p_{LB,t}^{(z)} > p_{LB,t-1}^{(z)}\}} \cdot \mathbf{1}_{\{x_t \notin x_g\}}$	4
	Forward Incr.	$\mathbf{1}_{\{p_{CoM,t}^{(x)} > p_{CoM,t-1}^{(x)}\}} \cdot \mathbf{1}_{\{x_t \notin x_g\}}$	4

achieves two full cycles of traversal consecutively, demonstrating the reliability of this context-aware gait switching.

Route (b) consists of a complete side-to-side traversal of the high platform with the following sequence: walk, climb-up, crawl, climb-down, and walk. Besides the successful execution of full-body skills, it also highlights the robustness of our perception pipeline in accurately identifying environmental contexts during dynamic maneuvers.



(a) Robot executing two consecutive full-skill cycles.



(b) Robot completing platform traversal.

Fig. 10: Experimental validation of the unified policy. The distilled system demonstrates real-world robustness by successfully completing diverse routes requiring complex full-body coordination.

C. Extended Experiments on Robustness

The adaptability and robustness of our policy is showcased in three extreme cases: (i) large external perturbation; (ii) significant perception artifacts; (iii) soft high platform;

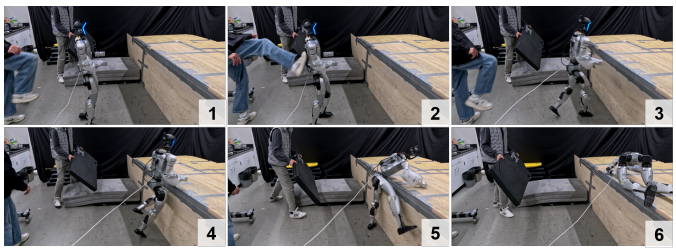


Fig. 11: The robot regains balance and climbs up the high platform after being heavily kicked.

1) Robustness to Perturbation: Our system exhibits significant robustness and context-aware adaptability when subjected to perturbations. As shown in Fig. 11, the robot is heavily kicked from behind when approaching the platform. Despite the stumble and unintended contact with the front surface of the platform, the robot adjusts rapidly. Notably, it adjusts the gait, changing the pivoting leg to maintain balance and initiate the climbing behavior. This behavior indicates that the distilled policy has obtained the ability to complete skill transitions even when initialized from near-failure states. Furthermore, the robot demonstrates the capacity to leverage these transitions as stabilizing maneuvers, utilizing the environment to regain balance once the transition conditions are met. This suggests that the distillation process successfully transfers the teachers' specialized robustness into a unified policy that can effectively modulate its behavior based on the physical context.

2) Robustness to Perception Artifacts: Figure 12 illustrates a typical elevation map observed by the robot and the robot's corresponding climbing maneuver. The map contains a significant batch of "ghost points", which form a fake obstacle behind the robot of a scale comparable to the target platform. Despite these substantial perceptual artifacts, the robot successfully climbs up the platform with nominal movement, demonstrating the policy's perceptual robustness. This resilience is obtained from the noises injected during the training process, especially the outlier clusters, which showcases the necessity of rigorous perception and noise modeling for reliable real-world deployment.

3) Robustness to Varying Contact Property: We further evaluate the policy's climbing capabilities by placing a soft mat, made of vinyl and foam, on top of the target platform (Fig. 13). This setup challenges the robot to climb up an unseen material with significantly different compliance and friction properties compared to the rigid training environments. The robot successfully climbs the soft mat in the first trial, maintaining the same levels of stability and efficiency observed when climbing rigid platforms.

The stability of the climbing skill originates from the policy's quasi-static and contact-rich moving patterns. Unlike dynamic jumping or lunging behaviors, our policy does not rely on impulsive supporting forces provided by rigid surfaces, nor does it depend on a limited number of contact points to maintain equilibrium. Instead, the policy learns to distribute

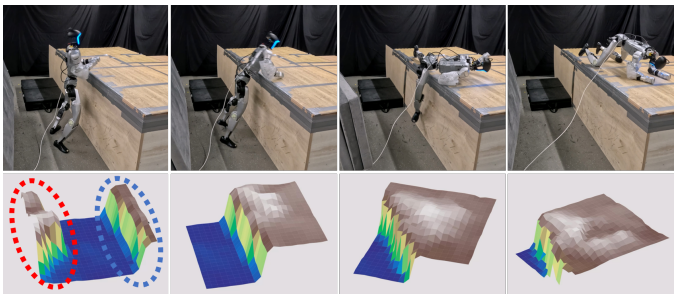


Fig. 12: A typical artifacted elevation map with significant outlier cluster. The red circle denotes the outliers while the blue circle denotes the target platform. The map is aligned with the platform edge for a clearer view.

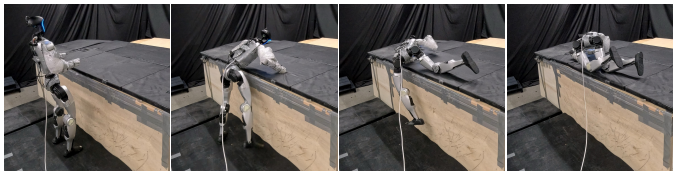


Fig. 13: The robot successfully and stably climbs up the platform with a black soft mat on top.

loads across multiple contact points, ensuring balance through consistent interaction with the platform. These results further demonstrate the adaptability of our policy to varying contact properties and highlight the advantages of our proposed reward formulation in learning safe, robust humanoid behaviors.

D. Effectiveness of Multi-Teacher Distillation

To evaluate our distillation pipeline, we compare the teacher and the distilled student across the four contact-rich full-body maneuvers in simulation. We report success rate and maximum contact force (Tab. IV), evaluated over 1,000 parallel environments with the same randomization ranges used during training. As expected, the specialized teacher policies achieve the strongest performance on their respective tasks. Nevertheless, the distilled student attains comparable success rates across all maneuvers, while maintaining statistically similar maximum contact forces within a safe range. Overall, these results indicate that the unified student policy faithfully captures the behaviors of diverse experts and retains near-teacher-level performance in simulation.

TABLE IV: Comparison of Teacher and Student Performance. **SR:** Success Rate; **M.C.F.:** Max Contact Force.

	Teacher Policy		Student Policy	
Skill	**SR (%)**	**M.C.F. (N)**	**SR (%)**	**M.C.F. (N)**
Climb-up	98.8	638 ± 479	98.6	657 ± 324
Climb-down	99.9	754 ± 241	99.0	762 ± 539
Stand-up	99.5	632 ± 222	99.1	680 ± 237
Lie-down	100.0	576 ± 125	100.0	637 ± 124

E. Details for Teacher Policy Training

1) Environment Configuration: We train the teacher policy in IsaacLab using 4096 parallel environments for each single skill. While the six skills share basic observations, climbing skills are additionally based on height scan dots, and walking skill additionally utilize a phase signal to lead the gait pattern. The full list of observations is in Table V. We also introduce perturbations and domain randomization, including previously discussed perception artifacts, to improve robustness (Table VI).

TABLE V: Observations and Noise for Teacher Training.

Skill	**Observation**	**Noise Range**
All Skills	Root angular velocity (rad/s)	$[-0.2, 0.2]$
	Projected gravity	$[-0.05, 0.05]$
	Joint position (rad)	$[-0.1, 0.1]$
	Joint velocity (rad/s)	$[-1.5, 1.5]$
	Last action	—
Climb-up/down	Elevation Map (m)	refer to Tab. VI
Walk / Crawl	Phase signal	—
	Velocity Commands (m/s)	—

TABLE VI: Perturbations and Domain Randomization Ranges.

Perturbed Terms	**Perturbed Range**
Torso CoM Position (m)	$x, y : [-0.05, 0.05]$
	$z : [-0.02, 0.02]$
Torso Mass (kg)	$m : [-1.0, 1.0]$
Static Friction	$\mu_s : [0.3, 1.6]$
Dynamic Friction	$\mu_d : [0.3, 1.2]$
Restitution	$e : [0.0, 0.5]$
Joint Default Position (rad)	$q : [-0.01, 0.01]$
Joint Initial Position (rad)	$\dot{q} : [-0.15, 0.15]$
External Push (m/s) (rad/s)	$v_x, v_y : [-0.5, 0.5]$
	$v_z : [-0.2, 0.2]$
	$\omega_r, \omega_p : [-0.5, 0.5]$
	$\omega_y : [-0.78, 0.78]$
	Interval (s): $[1, 3]$
Elevation Map Noise (m)	Gaussian: $[-0.15, 0.15]$
	Drift $d_x, d_y : [-0.05, 0.05]$
	Drift $d_z : [-0.1, 0.05]$
	Outliers: 20%

The six skills are categorized into two groups: (i) Non-periodic full-body maneuvers: climb-up, climb-down, stand-up, lie-down; (ii) Periodic locomotion skills: walk, crawl; Each group widely shares common rewards with several task-related reward terms. The full list of rewards is defined in Table X.

2) Algorithm Design and Network Architecture: We use Proximal Policy Optimization (PPO) to optimize the actor and the critic during the teacher policy training stage. The network architecture and hyperparameter are listed in Table VII.

F. Details for Multi-Teacher Distillation

1) Discussion on Data Distribution Construction: The construction of an appropriate data distribution is essential for the performance of the unified student policy. Our system encounters two primary challenges during distillation: (i) the vast

TABLE VII: Hyperparameter of Teacher Policy

Environment and Architecture	
Num. of Environments	4096
Episode Length	350 / 1000
Network Type	MLP
Activation	ELU
Actor Network	[512, 256, 128]
Critic Network	[512, 256, 128]
PPO Optimization Parameters	
Num. Epochs	5
Num. Mini Batches	4
Num. Steps per Batch	24
Num. Steps per Env	24
Normalization	Observation
Learning Rate lr	1.0e-3
Clip Parameter	0.2
Entropy Coefficient	0.01
Gamma γ	0.99
Lambda λ	0.95
Desired KL value	0.01
Max Gradient Norm	1.0

and highly diverse state distribution introduced by multiple teachers; and (ii) the stringent requirements for adaptability within each primitive skill and across their transitions.

In long-sequence traversal tasks, the amount of data sample for downstream skills is often bottlenecked by the success rate of all upstream skills. This dependency can lead to underfitting of specific maneuvers and strictly limited state coverage, thereby compromising robustness. For example, if the student's climb-up behavior is supervised only by trajectories where walk skill was the predecessor, the policy tends to overfit to that specific transition. Consequently, real-world deployment becomes fragile if the student is initialized from a state that slightly differs from the terminal walking state observed during training. To mitigate this, our environment design follows two key principles: (i) ensuring the student is trained on a state distribution equivalently wide to that of the individual teacher policies; and (ii) maintaining an evenly sampled and purposely controlled data distribution to ensure sufficient sample for every single skill.

To realize these principles, we employ a distributed environment configuration. Unlike typical multi-expert distillation frameworks that train students on sequential terrains, we utilize a "divide-and-conquer" approach where all skills and transitions are represented in parallel sub-environments. We initialize the robot across a broad range of states to maximize state-space coverage. We also observe that mastering particular skills is more challenging than learning other skills or the transition between skills. This is especially evident for the walking and climb-up skills, both of which necessitate a high degree of robustness for varied behavior or terrains. Consequently, we allocate higher proportions of training environments to these tasks to account for their extensive data distribution. The specific skills associated with each type of environment and their respective proportions are detailed in Table VIII.

The distillation process incorporates the full suite of configurations utilized during teacher training, including domain randomization, physical perturbations, perception artifacts, and varied terrain properties. To further enhance robustness, we apply action noise and symmetry augmentation during distillation. By injecting action noise, the student is forced into slightly perturbed, off-equilibrium states, allowing it to acquire expert labels for these critical near-failure cases. Additionally, symmetry augmentation mirrors the student's observations and the teacher's corresponding actions to synthesize additional data pairs, significantly improving data efficiency. These techniques prove essential for learning a student policy that maintains resilience under diverse real-world conditions.

TABLE VIII: Environment Distribution for Distillation.

Skills	**Env. Prop.**	**Terrains**	**Vel. Cmd.** (m/s)
Walk	0.17	Rough + Plane	Omni.
Crawl	0.08	Plane	Omni.
Stand-up + Walk	0.07	Plane	Zero + Omni.
Walk + Climb-up	0.16	Platform	Forward
Climb-up + Crawl	0.12	Platform	Forward + Lateral
Crawl + Climb-down	0.20	Platform	Lateral
Climb-down + Walk	0.15	Platform	Lateral + Backward
Lie-down + Crawl	0.05	Plane	Zero + Omni.

2) Algorithm Design and Network Architecture: We employ Behavioral cloning with DAgger in the distillation stage, minimizing the MSE loss between student actions and teacher actions as in Eq. 6. The network architecture and hyperparameters are in the Table IX.

$$\mathcal{L}(\theta) = \mathbb{E}_{o\sim\mathcal{D}}\left[\left\|\pi_\theta(o) - \mathbf{a}_{\text{teacher}}\right\|_2^2\right] \tag{6}$$

TABLE IX: Hyperparameter for Distillation

Environment and Architecture	
Num. of Environments	1000
Episode Length	400
Activation	ELU
Network Type	MLP
Student Network	[2048, 1024, 512, 256]
Optimization Parameters	
BC Iterations	4
DAgger Iterations	16
Num. Epochs	1500
Num. Steps per Batch	20000
Num. Steps per Env	400
Normalization	Observation
Learning Rate lr	3.0e-4
Action Noise Std	0.1
Gradient Length	1.0
Max Gradient Norm	1.0

TABLE X: Reward Formulations for Teacher Policy Training.

Non-periodic full-body maneuvers			
Skill	**Reward**	**Formulation**	**Weight**
All Skills	Survival	$\mathbf{1}_{\{\neg term\}}$	15
	Termination	$\mathbf{1}_{\{term,\neg timeout\}}$	-800
	Force Penalty	$\exp(0.01 \cdot \max(0, \|F\| - 500)) - 1$	-1
	Head Safety	$\exp(0.1 \cdot \|F_{head}\|) - 1$	-1
	Joint Limits	$\|\max(0, \|q_t\| - q_{soft})\|_1$	-10
	Hip Deviation	$\mathbf{1}_{\{\|q_{hip,yaw}\|>1.5 \vee \|q_{hip,roll}\|>1.4\}}$	-1
	Waist Deviation	$\mathbf{1}_{\{\|q_{waist\_yaw}\|>1.4\}}$	-6
	Joint Velocity	$\|\dot{q}_t\|^2$	-0.001
	Joint Accel.	$\|\ddot{q}_t\|^2$	-2e-8
	Action Rate	$\|a_t - a_{t-1}\|^2$	-0.2
	Torque	$\|\tau_t\|^2$	1.5e-5
	Power	$\sum \|\tau_t \cdot \dot{q}_t\|$	-1e-5
	Body Slip	$\sum_{i \in \mathcal{C}} \|v_{i,t}^{(xy)}\|$	-0.1
	Base Ang. Vel.	$\|\omega_{base,t}^{(xy)}\|^2$	-0.005
	Base Accel.	$\|\ddot{p}_{base,t}\|^2 + 0.02\|\dot{\omega}_{base,t}\|^2$	-0.0001
	Body Accel.	$\sum_i \|\ddot{p}_{i,t}\|$	-0.0002
Climb-up	Upward Progress	$\mathbf{1}_{\{p_{LB,max}^{(z)} \geq p_{LB,t}^{(z)}\}} \cdot \mathbf{1}_{\{x_t \notin x_g\}}$	-4
	Edge Approach	$\mathbf{1}_{\{p_{CoM,max}^{(x)} \geq p_{CoM,t}^{(x)}\}} \cdot \mathbf{1}_{\{x_t \notin x_g\}}$	-4
	Terminal Posture	$\mathbf{1}_{\{t>H-1s\}} \cdot \mathbf{1}_{\{x_t \in x_g\}} \cdot \exp(-0.1 \cdot \|q_t - q_{prone}\|)$	7
Climb-down	Descent Progress	$\mathbf{1}_{\{p_{LB,min}^{(z)} \leq p_{LB,t}^{(z)}\}} \cdot \mathbf{1}_{\{x_t \notin x_g\}}$	-4
	Edge Clearance	$\mathbf{1}_{\{p_{CoM,min}^{(x)} \leq p_{CoM,t}^{(x)}\}} \cdot \mathbf{1}_{\{x_t \notin x_g\}}$	-4
	Terminal Posture	$\mathbf{1}_{\{t>H-1s\}} \cdot \mathbf{1}_{\{x_t \in x_g\}} \cdot \exp(-0.1 \cdot \|q_t - q_{standing}\|)$	7
Stand-up	Height Progress	$\mathbf{1}_{\{p_{head,max}^{(z)} \geq p_{head,t}^{(z)}\}} \cdot \mathbf{1}_{\{x_t \notin x_g\}}$	-4
	Balance Progress	$\mathbf{1}_{\{d_{bal,min} \leq d_{bal,t}\}} \cdot \mathbf{1}_{\{x_t \notin x_g\}}$	-4
	Terminal Posture	$\mathbf{1}_{\{t>H-1s\}} \cdot \mathbf{1}_{\{x_t \in x_g\}} \cdot \exp(-0.1 \cdot \|q_t - q_{standing}\|)$	7
Lie-down	Descent Progress	$\mathbf{1}_{\{p_{CoM,min}^{(z)} \leq p_{CoM,t}^{(z)}\}} \cdot \mathbf{1}_{\{x_t \notin x_g\}}$	-4
	Head Placement	$\mathbf{1}_{\{p_{head,min}^{(z)} \leq p_{head,t}^{(z)}\}} \cdot \mathbf{1}_{\{x_t \notin x_g\}}$	-4
	Terminal Posture	$\mathbf{1}_{\{t>H-1s\}} \cdot \mathbf{1}_{\{x_t \in x_g\}} \cdot \exp(-0.1 \cdot \|q_t - q_{prone}\|)$	7
Periodic locomotion skills			
Skill	**Reward**	**Formulation**	**Weight**
All Skills	Track lin. Velocity	$\exp(-\|v^{(x,y)} - v_{cmd}\|^2 / 0.5^2)$	1.3
	Track Ang. Velocity	$\exp(-\|\omega^{(z)} - \omega_{cmd}\|^2 / 1.0^2)$	1.3
	Vertical Lin. Velocity	$\|v^{(z)}\|^2$	-2
	Horizontal Ang. Velocity	$\|\omega^{(x,y)}\|^2$	-0.15 / -0.05
	Height Penalty	$(p_{root}^{(z)} - h_{des})^2$	-10
	Joint Acc. Penalty	$\sum_{j \in \mathcal{A}} \ddot{q}_{j,t}^2$	$-2.5e-7$
	Joint Vel. Penalty	$\sum_{j \in \mathcal{A}} \dot{q}_{j,t}^2$	$-1.5e-3$
	Action Rate	$\|a_t - a_{t-1}\|^2$	-0.1
	Joint Limits	$\sum(\max(0, q_{min} - q) + \max(0, q - q_{max}))$	-5
	Survival	$\mathbf{1}_{\{\neg terminated\}}$	0.2 / 10
	Torque Penalty	$\|\tau_t\|^2$	$-1.0e-5$
	Undesired Contact	$\sum_{b \in \mathcal{B}} \mathbf{1}_{\{\|F_{contact,b}\|>0.1\}}$	-1
Walk	Base Orientation	$\|g_b^{(x,y)}\|^2$	-1.0
	Hip Deviation	$\|q_t - q_{default}\|^2$	-1.0
	Contact Slip	$\sum_{b \in \text{ankle\_roll}} \mathbf{1}_{\{\|F_{c,b}\|>1\}} \cdot \|v_b\|^2$	-0.2
	Feet Swing Height	$\sum_{f \in \text{ankle\_roll}} \mathbf{1}_{swing} \cdot (0.08 - h_f)^2$	-20
	Gait Phase	$\sum_i \mathbf{1}_{\{C_i = C_{target}(\phi_t)\}}$	0.18
	Feet Air Time	$\sum_{f \in \text{ankle\_roll}} (\tau_{air,f} - 0.5) \cdot \mathbf{1}_{contact} \cdot \mathbf{1}_{move}$	0.1
Crawl	Termination	$\mathbf{1}_{terminated}$	-100
	Lying Deviation	$\|q_t - q_{lying}\|^2$	-1.0
	Contact Force Penalty	$\sum_b \max(0, \|F_b\| - 500)$	-0.01