Hold My Beer: Learning Gentle Humanoid Locomotion and End-Effector Stabilization Control

Yitang Li[1,2], **Yuanhang Zhang**[2], **Wenli Xiao**[2]
Chaoyi Pan[2], **Haoyang Weng**[1,2], **Guanqi He**[2], **Tairan He**[2], **Guanya Shi**[2]
[1]IIIS, Tsinghua University, [2]Carnegie Mellon University
Website: https://lecar-lab.github.io/SoFTA/
Code: https://github.com/LeCAR-Lab/SoFTA

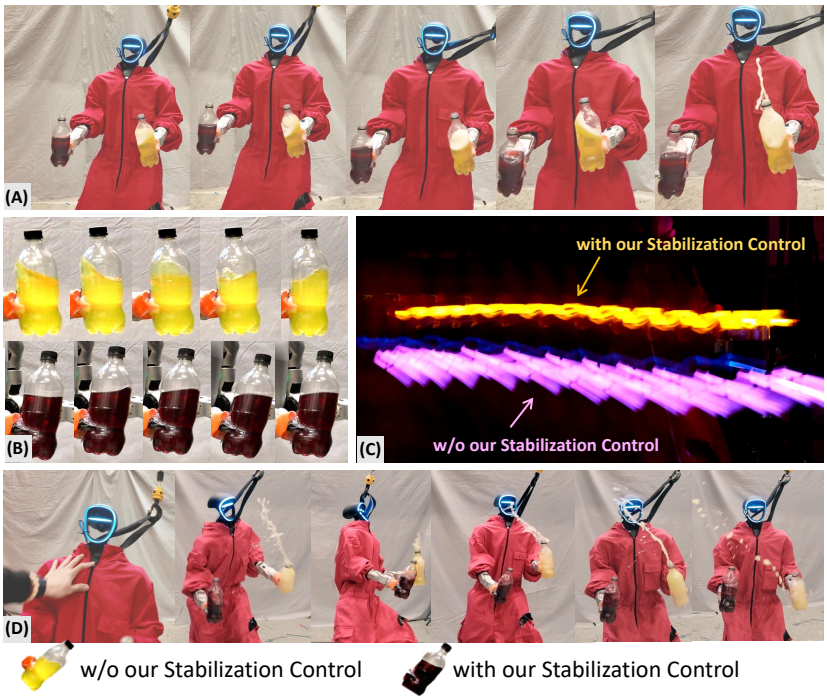


Figure 1: Learning Gentle Humanoid Locomotion and End-Effector Stabilization Control with SoFTA: **(A)** Carrying bottles of drink during a 1m/s large-step walk. **(B)** Liquid surface when the robot is tapping in place. **(C)** Long-exposure photo of the robot holding a glow stick walks forward. **(D)** SoFTA keeps the drink from spilling, even when subjected to a fierce push. See the website for more video: https://lecar-lab.github.io/SoFTA/

Abstract: Can your humanoid walk up and hand you a full cup of beer—without spilling a drop? While humanoids are increasingly featured in flashy demos—dancing, delivering packages, traversing rough terrain—fine-grained control during locomotion remains a significant challenge. In particular, stabilizing a filled end-effector (EE) while walking is far from solved, due to a fundamental mismatch in task dynamics: locomotion demands slow-timescale, robust control, whereas EE stabilization requires rapid, high-precision corrections. To address this, we propose SoFTA, a Slow-Fast Two-Agent framework that decouples upper-body and lower-body control into separate agents operating at different frequencies and with distinct rewards. This temporal and objective separation mitigates policy interference and enables coordinated whole-body behavior. SoFTA executes upper-body actions at 100 Hz for precise EE control and lower-body actions at 50 Hz for robust gait. It reduces EE acceleration by 2-5$\times$ to baselines and performs much closer to human-level stability, enabling delicate tasks such as carrying nearly full cups, capturing steady video during locomotion, and disturbance

RSS 2025 Workshop on Whole-Body Control and Bimanual Manipulation (RSS2025WCBM).

rejection with EE stability. We validate SoFTA on both Unitree G1 and Booster T1, showing strong cross-platform generalizatin.

Keywords: Humanoid Robots, Reinforcement Learning, Stable Locomotion

1 Introduction

Humanoid robots are designed to operate in human-centric environments, with general-purpose structures that make them well-suited for diverse tasks. Recent advances in locomotion [1–11] and manipulation [12–16] have pushed humanoid performance toward human-level ability [17]. However, one critical capability remains underexplored: fine-grained end-effector (EE) stabilization during locomotion. This capability is essential for safe and precise physical interaction with objects—such as handing over a cup of water or recording stable video—yet current humanoids fall short. For instance, the default Unitree G1 controller yields average EE accelerations around 5m/s^2 —over 10× higher than human levels—leading to excessive shaking and making delicate tasks infeasible.

We identified a fundamental performance gap stemming from the disparities in task characteristics between EE stabilization and locomotion, both in terms of objectives and dynamics. At the *objective* level, locomotion requires traversability, which naturally introduces non-quasi-static dynamics. In contrast, EE stabilization requires a base with minimal motion to maintain precision. At the *dynamics* level, lower body locomotion operates with "slow" dynamics meaning it can only be controlled through discrete long time-scale contacts. The nature of ground contacts makes it more susceptible to the sim-to-real gap, demanding greater robustness against noise and disturbances. On the other hand, EE control involves "fast" dynamics, with fully actuated and more controllable arms to produce continuous torques, allowing for fast and precise corrections.

To bridge the gap, we propose **SoFTA**—a **S**l**o**w-**F**ast **T**wo-**A**gent reinforcement learning (RL) framework that decouples the action and value spaces of the upper and lower body. This design enables different execution frequencies and reward structures: the upper-body agent acts at high frequency for precise EE control with compensate behavior, while the lower-body agent prioritizes robust locomotion at a slower frequency. SoFTA facilitates stable training and whole-body coordination by this decoupling, resulting in fast and accurate EE control alongside robust locomotion. Like shown in Figure 1, our system achieves a 50–80% reduction in EE acceleration over baselines. SoFTA can achieve EE acceleration less than 2m/s^2 in diverse locomotion, which is much closer to human-level stability, enabling tasks like serving coffee or stable video recording. Our key contributions are:

- We introduce SoFTA, a novel slow-fast two-agent RL framework that decouples control for locomotion and EE stabilization in both time and objective space, enabling robust locomotion and precise, stable EE control through frequency separation and task-specific reward design.
- We demonstrate real-world deployment of SoFTA on a Unitree G1 humanoid, enabling tasks such as walking while carrying liquids or recording stable first-person videos.
- Extensive experiments are conduct in both simulation and real-world with in-depth analysis across control frequencies, showing that SoFTA can effectively stabilizes the end-effector during locomotion through its frequency design.

2 Related Work

Learning-based Humanoid Control Recent advances in learning-based whole-body control have enabled humanoid robots to acquire a wide range of skills in simulation. Efforts such as domain randomization and system identification to better align simulation with real-world dynamics [18–21] have proven to be effective for sim-to-real transfer of humanoid policies. These capabilities span robust locomotion [22–24, 1–11, 25], advanced manipulation [12–16] and integrated loco-manipulation behaviors [26–32, 17]. Despite these promising developments, relatively little attention has been paid to achieving precise and stable EE control, which is essential for fine-grained

humanoid loco-manipulation. In this work, we focus on enabling humanoid robots maintain end-effector stability during locomotion.

End-Effector Control for Mobile Manipulators Stabilizing EE during motion is crucial for mobile manipulation. Prior work predominantly focuses on wheeled robots, where model-based approaches unify base and arm control through optimization [33–40], but they rely on accurate dynamics models and predefined contact schedules, limiting their scalability to complex humanoid systems. Hybrid approaches [41–43] combine learned locomotion with planned arms but often freeze the base, reducing coordination. Joint learning [44, 45] improves tracking, mostly on non-humanoids. In contrast, we present the first method to achieve fine-grained end-effector stabilization during dynamic humanoid locomotion.

Humanoid Policy Architecture To enable effective humanoid policy learning, researchers have explored various architectural designs. Single-agent whole-body policies [21, 46] offer flexibility for complex tasks but are challenged by high-dimensional state-action spaces. Inspired by MARL [47], recent work decoupling policies to simplify training. Multi-critic methods [48, 49] handle value conflicts, while decentralized control [50] assigns body parts to separate controllers. Others [32, 51] split locomotion and manipulation, enabling diverse behaviors. Still, few leverage structural decoupling fully. Our method extends this idea by separating reward design and update timing, achieving stable EE control and robust locomotion.

3 SoFTA for Learning Stable End-effector Control and Robust Locomotion

3.1 Problem Statement

Observations and Actions. We aim to control a humanoid robot to stabilize its end-effector at target positions while also following locomotion commands. We formulate the problem as a goal-conditioned RL task, where the policy $\pi(s_t^{\text{prop}}, s_t^{\text{goal}})$ is trained to output an action $a_t \in \mathbb{R}^{27}$, representing target joint positions. The proprioceptive input s_t^{prop} includes a 5-step history of joint positions $q_t \in \mathbb{R}^{27}$, joint velocities $\dot{q}_t \in \mathbb{R}^{27}$, root angular velocities $\omega_t^{\text{root}} \in \mathbb{R}^3$, projected gravity vectors $g_t \in \mathbb{R}^3$, and past actions $a_t \in \mathbb{R}^{27}$. The goal state s_t^{goal} contains target root linear velocity $v_t^{\text{goal}} \in \mathbb{R}^2$, target yaw angular velocity $\omega_t^{\text{goal}} \in \mathbb{R}$, desired base heading $h_t^{\text{goal}} \in \mathbb{R}$, $c_t^{\text{goal}} \in \mathbb{R}^2$ (with a binary stand/walk command and a gait frequency), and $c_t^{\text{EE}} \in \mathbb{R}^{5\times n}$ encodes the EE command. Here, n denotes the number of potential end-effectors, each with a 5-dimensional command specifying whether it is activated for stabilization, the x- and y-coordinates in the local frame, the z-coordinate in the global frame, and the tracking tolerance σ.

Objective Formulation for Stable EE control We use PPO [52] to maximize the cumulative discounted reward $\mathbb{E}\left[\sum_{t=1}^{T} \gamma^{t-1} r_t\right]$. Several rewards r_t are defined to achieve stable end-effector control: 1) penalizing high linear/angular acceleration, $r_{\text{acc}} = -\|\ddot{p}_{\text{EE}}\|_2^2$, $r_{\text{ang-acc}} = -\|\dot{\omega}_{\text{EE}}\|_2^2$; 2) encouraging near-zero linear/angular acceleration, $r_{\text{zero-acc}} = \exp\left(-\lambda_{\text{acc}}\|\ddot{p}_{\text{EE}}\|_2^2\right)$, $r_{\text{zero-ang-acc}} = \exp\left(-\lambda_{\text{ang-acc}}\|\dot{\omega}_{\text{EE}}\|_2^2\right)$; 3) penalizing gravity tilt in the end-effector frame, $r_{\text{grav-xy}} = -\left\|\mathbf{P}_{xy}(R_{\text{EE}}^T \mathbf{g})\right\|_2^2$. $\ddot{p}_{\text{EE}} \in \mathbb{R}^3$ is the linear acceleration, $\dot{\omega}_{\text{EE}} \in \mathbb{R}^3$ is the angular acceleration, $\lambda_{\text{acc}}, \lambda_{\text{ang–acc}} > 0$ are exponential reward scale factors, $R_{\text{EE}} \in \text{SO}(3)$ is the rotation matrix, $\mathbf{g}$ is the gravity vector, and $\mathbf{P}_{xy}(\cdot) : \mathbb{R}^3 \to \mathbb{R}^2$ projects onto the xy-plane.

Task Characteristics. Stable End-Effector Control and Robust Locomotion are fundamentally different tasks in both their objectives and dynamics.

At the *objective level*, end-effector control demands extreme stability, requiring the base to remain as static as possible, while locomotion must accommodate varying gaits and momentum changes. Precise end-effector control benefit from sharp, fine-grained, continuous rewards, whereas locomotion favors long-horizon, robustness-focused rewards. Given these differences, using a single critic to aggregate all reward signals may not be the most effective way.

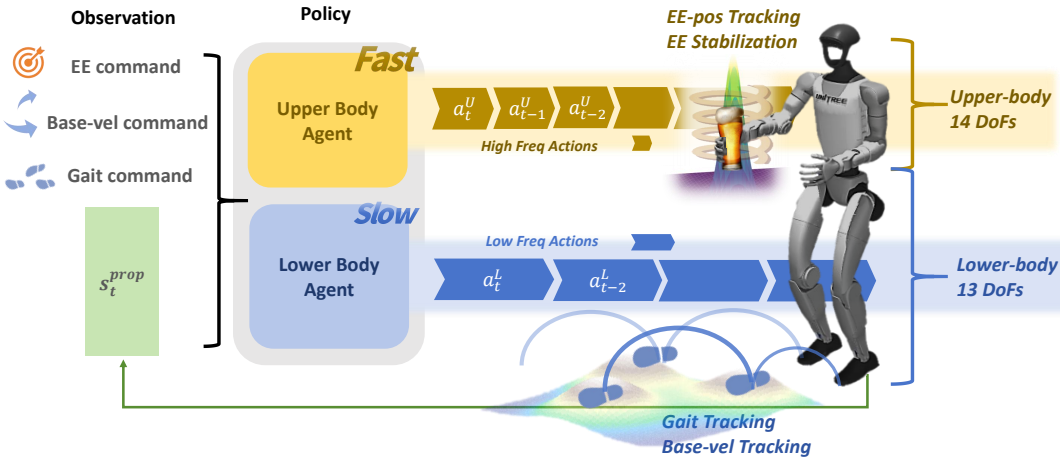


Figure 2: Overview of the SoFTA framework: The framework employs two distinct agents that share the same observation but act within separate action spaces at different rates, targeting two fundamentally different task: stable end-effector control and robust locomotion. Stable end-effector control requires a sharp reward landscape and rapid upper-body actions for precise manipulation, whereas robust locomotion focuses on maintaining robustness under gait rewards.

At the *dynamics level*, locomotion is governed by discrete ground contact forces and exhibits "slower" dynamics due to its long time-scales. In contrast, the upper body has a "faster" dynamics, and is often more controllable by fully actuated arms, affording more aggressive and faster control strategies. Given that higher control frequencies tend to increase sensitivity and exacerbate the sim-to-real gap [53–55, 19], while lower frequencies are less precise but more deployable and robust, it is advantageous to modulate control rates accordingly.

3.2 SoFTA: Slow-Fast Two-Agent Framework.

Slow-Fast Two-Agent Framework Design. Given these distinct task characteristics, we propose SoFTA, a two-agent framework in which each agent independently controls a disjoint subset of the robot's degrees of freedom at different control frequencies (Figure 2). Both agents in SoFTA share the full-body observation to facilitate coordinated behavior while allowing each agent to specialize. Specifically, the upper-body agent operates at a high frequency to control 14 arm joints, enabling precise and rapid adjustments for end-effector stability while the lower-body agent runs at a lower frequency, managing the legs and waist to ensure stable locomotion and balance. This asymmetric control frequency matches the longer characteristic timescale of gait cycles compared to the fast, precision motion required for stabilization tasks.

Training SoFTA with Separate Reward Groups. Due to the differing control dynamics and timescales of upper-body and lower-body tasks, their reward signals are inherently heterogeneous, which can lead to interference and suboptimal learning. To improve credit assignments [56–59], we decompose the overall reward into two semantically aligned components, each tailored to the respective PPO agent. This decomposition provides more targeted feedback, preventing overloading of any agent and promoting fair cooperation. To further encourage collaborative behavior and sustained task execution, we include the termination reward in both reward streams. While both agents share the same observation space, they operate with separate actor and critic networks and do not share parameters. More details are summarized in Appendix A.1.

4 Experimental Results

In this section, we evaluate the performance of SoFTA in both simulation and real-world environments. Our experiments aim to answer the following key questions:

Simulation Results (Isaac Gym)		Acc (m/s^2) ↓		AngAcc (rad/s^2) ↓		Acc-Z (m/s^2) ↓		Grav-XY (m/s^2) ↓	
Task	Method	mean	max	mean	max	mean	max	mean	max
Tapping	Lower-body RL + IK	1.83±0.05	4.92±0.18	10.3±0.4	37.8±0.8	0.77±0.02	3.46±0.08	0.15±0.01	0.53±0.02
	Whole-body RL	1.29±0.03	4.24±0.14	9.76±0.25	**31.7**±0.7	0.58±0.01	1.71±0.05	**0.09**±0.01	0.46±0.02
	SoFTA	**1.08**±0.03	**3.45**±0.11	**8.10**±0.20	32.1±0.7	**0.44**±0.01	**0.87**±0.04	0.11±0.01	**0.44**±0.01
RandCommand	Lower-body RL + IK	3.17±0.07	6.51±0.21	15.5±0.6	53.4±1.2	1.53±0.04	3.33±0.10	0.19±0.01	0.52±0.02
	Whole-body RL	2.47±0.05	5.19±0.19	**11.0**±0.3	44.2±1.0	1.66±0.03	2.98±0.09	**0.10**±0.01	**0.36**±0.01
	SoFTA	**1.48**±0.06	**4.78**±0.15	11.4±0.4	**42.3**±1.0	**0.33**±0.02	**1.97**±0.06	0.14±0.01	0.39±0.01
Push	Lower-body RL + IK	3.88±0.16	25.0±1.2	26.9±1.4	**65.1**±4.2	2.12±0.10	4.18±0.16	0.45±0.06	0.82±0.07
	Whole-body RL	4.62±0.17	29.6±1.80	20.1±1.4	70.3±5.1	2.38±0.11	6.20±0.20	0.91±0.09	1.83±0.31
	SoFTA	**2.98**±0.10	**18.8**±0.65	**14.6**±0.9	66.8±3.6	**0.54**±0.05	**2.35**±0.12	**0.31**±0.05	**0.67**±0.06

Table 1: Simulation Results: EE stability is evaluated in Isaac Gym across various tasks. SoFTA consistently outperforms the baselines in most metrics, demonstrating superior EE stability.

- **Q1** (Section 4.1): Can the *Two-Agent design* of SoFTA perform better in *simulation*?
- **Q2** (Section 4.2): What capabilities does SoFTA enable in real world?
- **Q3** (Section 4.3): How important is the *Slow-Fast frequency design* for SoFTA performance?

Baselines. We compare SoFTA with the following baselines. 1) *Robot Default Controller*[1] [60]: Utilizes the default Unitree locomotion, providing stable and low-impact locomotion. It serves as a naive baseline for EE stabilization. 2) *Lower-body RL + IK* [30]: Employs a learned lower body policy for locomotion followed by inverse kinematics to stabilize the EE. 3) *Whole-body RL*: A single RL agent is trained to jointly control the whole body for both locomotion and EE control.

Ablations of SoFTA. We evaluate variants of SoFTA using different upper-body and lower-body frequency pairings of 33.3 Hz, 50 Hz, and 100 Hz.

Experiment Setup. We train our policy in Isaac Gym at 200 Hz simulation frequency. During training, reward functions, termination conditions, and curriculum design are consistent and frequency-agnostic across all comparisons. For real-world evaluation, we deploy SoFTA on the Unitree G1 robot, following the sim-to-real pipeline of HumanoidVerse [61].

Metrics. We evaluated EE stability using the following metrics: linear acceleration norm (*Acc*), angular acceleration norm (*AngAcc*), and projected gravity in the XY plane of EE frame (*Grav-XY*). Specifically, during locomotion, the z-direction may experience sudden velocity changes due to contact, so we additionally report the z-acceleration (*Acc-Z*) for a more comprehensive evaluation. These metrics are reported as both the mean and maximum absolute values. Each metric is evaluated over 3 runs with mean and standard reported. For real-world acceleration, we collect pose data at 200 Hz using a mocap system for evaluation. The data is first interpolated, with abnormal points removed, and then double differentiation and filtering are applied to compute the acceleration.

4.1 Simulation Results

To answer **Q1** (*Can the* Two-Agent design *of SoFTA perform better in* simulation*?*), we assess EE stability across three locomotion scenarios: (1) *Tapping*: the robot steps in place to test stability under consistent, predictable contact events; (2) *RandCommand*: where random commands are issued every 10 seconds to evaluate robustness across diverse motions; and (3) *Push*: where the base is perturbed with a 0.5m/s velocity in a random direction every second to simulate unpredictable external disturbances. The results are summarized in Table 1.

We observe that *Lower-body RL + IK* performs the worst due to lack of dynamics awareness, while *Whole-body RL* improves but struggles to stabilize the EE in demanding scenarios like *Push*, where external disturbances amplify the instability. In contrast, SoFTA achieves the best overall performance, significantly reducing EE accelerations, especially in the vertical direction, highlighting the advantage of our decoupled design with frequency scheduling.

[1]This baseline is applicable only in the real world due to the accessibility of the built-in controller.

Task	Method	Acc (m/s^2) ↓ mean	Acc (m/s^2) ↓ max	AngAcc (rad/s^2) ↓ mean	AngAcc (rad/s^2) ↓ max	Acc-Z (m/s^2) ↓ mean	Acc-Z (m/s^2) ↓ max	Grav-XY (m/s^2) ↓ mean	Grav-XY (m/s^2) ↓ max
Tapping	Robot Default Controller	4.67±0.41	9.71±0.88	17.3±2.0	60.4±6.8	1.25±0.08	4.01±0.42	0.41±0.05	1.07±0.10
	Whole-body RL	1.86±0.21	6.11±0.43	16.1±1.5	47.9±5.3	1.34±0.10	5.10±0.44	**0.17**±0.02	0.97±0.11
	SoFTA	**1.35**±0.12	**4.96**±0.38	**11.2**±1.1	**41.2**±5.5	**0.52**±0.06	**2.33**±0.36	0.43±0.02	**0.75**±0.08
TrajTrack	Robot Default Controller	4.88±0.33	11.6±0.9	18.2±4.1	47.8±5.0	1.41±0.09	5.53±0.35	0.86±0.06	1.72±0.15
	Whole-body RL	2.95±0.48	12.2±1.2	13.4±1.4	60.5±8.7	2.02±0.21	9.51±0.82	0.54±0.04	1.72±0.11
	SoFTA	**1.51**±0.08	**6.25**±0.47	**10.7**±0.7	**42.4**±3.3	**0.62**±0.03	**3.17**±0.21	**0.48**±0.03	**1.18**±0.09
Turning	Robot Default Controller	5.55±0.28	14.0±0.4	23.2±3.7	62.1±8.6	1.80±0.09	7.33±0.37	0.90±0.05	1.83±0.09
	Whole-body RL	4.21±0.21	8.93±0.45	16.2±1.1	**57.9**±7.4	1.84±0.11	5.97±0.30	**0.31**±0.02	0.87±0.04
	SoFTA	**1.61**±0.08	**4.01**±0.20	**9.41**±0.81	62.8±8.8	**0.72**±0.04	**3.94**±0.20	0.36±0.02	**0.71**±0.04

Table 2: Real-World Results: EE stability evaluated in Real World across diverse task settings. SoFTA consistently outperforms baselines, especially in Acc-Z metric.

Benefit from Two-Agent Reward Group Separation. Figure 3 shows the reward conflict between `EE-AngAcc-Penalty` for EE stabilization and `Angular-Vel-Tracking-Reward` for locomotion. For Whole-Body RL, optimizing both is difficult: prioritizing locomotion increases EE penalties, while a dominant EE penalty will make the RL to not keep standing all the time, sacrificing locomotion quality (see the last half of the blue line). In contrast, SoFTA resolves this by decoupling the objectives into two separate agents. Even with a significant EE penalty, the lower body keeps improving locomotion, then coordinates, enabling more stable learning and better performance.

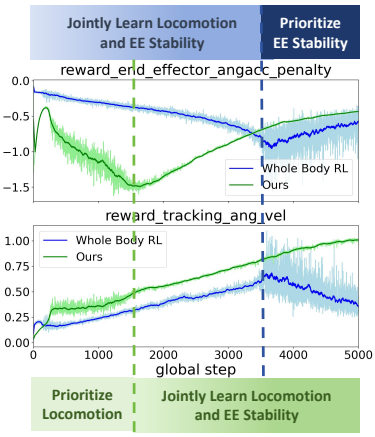
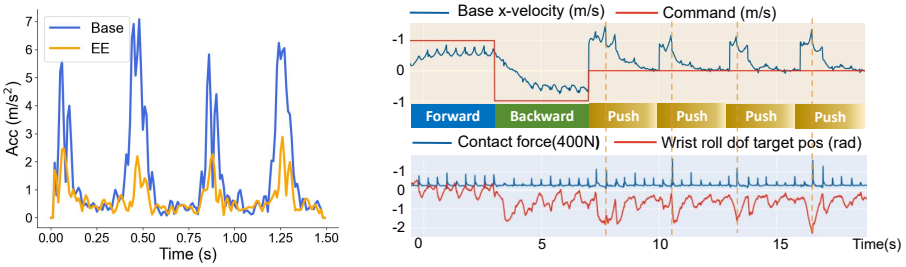


Figure 3: Reward Curves of EE-term and locomotion-term during Training.

Emergent Compensation Behavior. Figure 4(a) shows the acceleration curves of the base and EE. Our policy reduces sharp base accelerations caused by ground contacts, indicating stability through effective compensation, not just reduced base motion. To illustrate this, we visualize arm DoF target positions and contact force patterns. As seen in Figure 4(b), DoF activations align with locomotion rhythm and contact events, with compensation peaking during external pushes and ground impacts, highlighting the upper body's role in stabilizing the EE.



(a) Acc curves of Base and EE during locomotion.

(b) Compensate behaviors of wrist roll dof.

Figure 4: Emergent Compensation Behavior.

4.2 Real-World Results

To answer **Q2** (*What capabilities does SoFTA enable in real world?*), we assess EE stability in three real-world locomotion scenarios: (1) *Tapping*; (2)*TrajTrack* to move periodically along a straight line trajectory, and (3) *Turning* to do in-place rotation. Note that the IK-based method relies heavily on motion capture system. Even with perfect state information in simulation, it fails to produce strong results, so we do not include it in the real-world experiments.

The results in Table 2 show that the *Robot Default Controller* exhibits the highest acceleration across nearly all metrics, highlighting that even carefully designed locomotion controllers with gentle step-

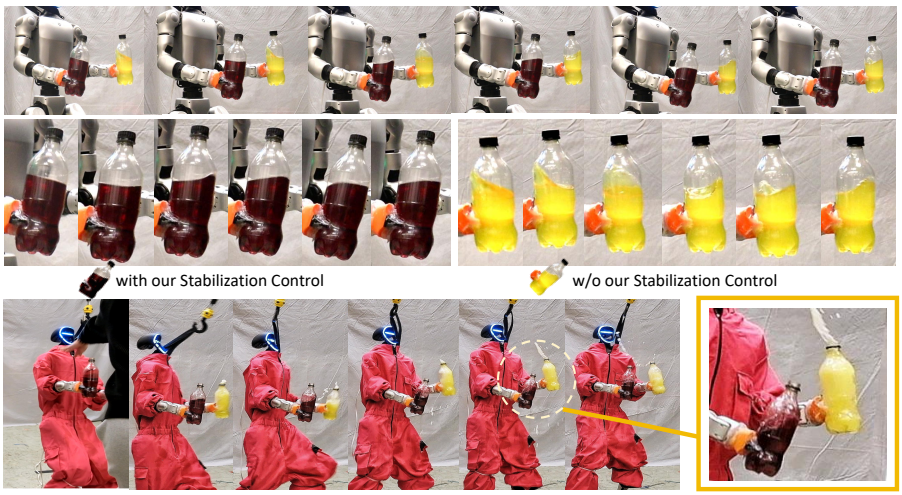


Figure 5: *Top:* Humanoid carring bottle of water without spillage during tepping. *Bottom:* Humanoid disturbance rejection with EE stability.

ping are insufficient for tasks requiring precise EE stability. While *Whole-body RL* offers moderate improvements, it struggles under motions with large movement like *TrajTrack*. In contrast, SoFTA maintains consistent and robust performance even during diverse locomotion. Compared to simulation results, real-world tests reveal that despite using the same domain randomization, observation noise, and reward functions, SoFTA demonstrates stronger sim-to-real transferability. Whole-body RL, by comparison, shows noticeably sluggish and hesitant steps, with shifts during foot tapping, likely due to excessive upper-body influence.

With the EE stability plus robust locmotion, SoFTA enables the robot to perform the following precise and stable upper-body tasks during locomotion.

Case 1: Humanoid Carrying Bottle without Spillage. Figure 5 shows the humanoid carrying a water bottle during locomotion. Even in tapping, without stabilization (YELLOW), contact impacts cause the liquid to slosh noticeably. In contrast, SoFTA (**RED**) greatly suppresses liquid motion, allowing the robot to carry an almost full cup of water smoothly while walking. Beyond periodic locomotion, our policy also demonstrates strong disturbance rejection capabilities. As shown in the Figure 5, when subjected to sudden and forceful pushes, the robust locomotion of the robot quickly adapts to avoid falling, while the upper body actively compensates to keep the end effector as steady as possible, effectively preventing the liquid from spilling.

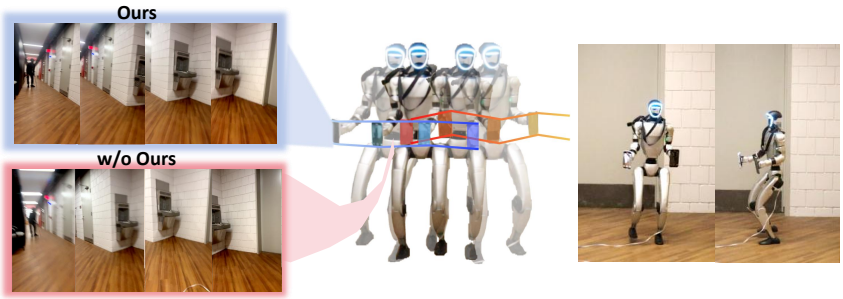


Figure 6: Humanoid as Camera Stabilizer to record videos.

Case 2: Humanoid as Camera Stabilizer. Figure 6 shows video footage recorded by the robot during continuous turning, comparing with and without stabilization. SoFTA ensures smooth and consistent camera motion, avoiding visible jitter with off-gantry-level robust locomotion. This allows the robot to record long, uninterrupted videos.

	IsaacGym		Sim2Sim		Sim2Real	
Acc (m/s^2) ↓	mean	max	mean	max	mean	max
w/o Tracking	**0.82**	**2.06**	2.29	7.83	2.91	8.15
with Tracking	1.08	3.45	**1.14**	**4.05**	**1.35**	**4.96**

Table 3: EE Tracking Improve Transferability

Benefits of Random EE Position Command for Sim2Real. While both the whole-body RL and SoFTA incorporate EE tracking, we found that tracking, specifically by forcing the EE to remain stable at a given position, plays a crucial role in generalization. Fixing the EE position during training often led to overfitted behaviors tied to specific poses and simulator dynamics, resulting in poor transferability. In contrast, training with random EE position commands promotes more reactive and adaptable motions, fostering compensation patterns that transfer more effectively (Table 3).

4.3 In-Depth Analysis on Frequency Design

To answer **Q3** (*How important is the* Slow-Fast frequency design *for SoFTA performance?*), we compare peak EE acceleration under various frequency settings in both simulation and real-world environments, using two scenarios: *Tapping* (predictable contacts) and *Stop* (sudden switch from walking to standing). As shown in Figure 7, our slow-fast design, 50 Hz for the lower body and 100 Hz for the upper, consistently achieves lower acceleration across tasks and domains.

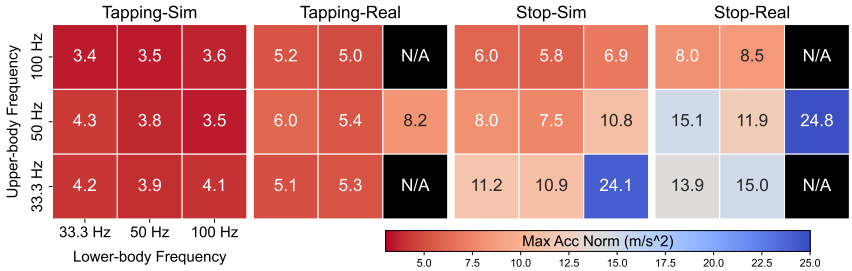


Figure 7: Max Acc under Different Control Frequencies in Simulation and Real World: Higher values reflect reduced stability. N/A indicates unstable or failed trials in the real-world testing.

We observe that in simulation, a 50 Hz lower-body policy is sufficient for maintaining stable locomotion, even under unpredictable conditions. In contrast, a higher-frequency upper-body policy proves beneficial for rapid recovery during sudden stops. From a sim-to-real perspective, the results indicate that deploying a high-frequency lower-body policy may introduce stricter deployment constraints. In our real-*Stop* trials, a 100 Hz lower-body policy caused more oscillations and degraded EE performance, with some instances resulting in failure. This degradation may be due to increased sensitivity to observation noise and control delays. On the other hand, running the upper-body agent at 100 Hz did not exhibit such issues and consistently enhanced overall performance. Considering that 50 Hz locomotion is a widely adopted standard and that inference-time constraints (0.01 s) are in place, our Slow-Fast Frequency configuration appears to be near-optimal. The further analysis of high-frequency upper body behaviors can be shown in Appendix A.2.

4.4 Cross Embodiment Validation

To evaluate the generalizability of our training method and control design, we apply it to a different robot, Booster T1 [62], which has distinct joint configurations and body proportions. Using the same framework, the T1 policy outperforms its default controller(Figure 8) in stability and precision, especially during sudden stops and fine end-effector control. This demonstrates that our approach captures transferable structural priors, enabling robust behavior across diverse humanoid embodiments without additional tuning.

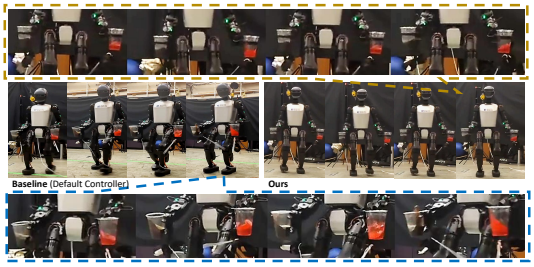


Figure 8: Real-world Results on Booster T1.

5 Conclusion

In this paper, we present SoFTA, a Slow-False Two-Agent reinforcement learning framework that enables robust locomotion and precise, stable EE control through frequency separation and task-specific reward design. Extensive experiments show up SoFTA can have a 50-80% reduction in EE acceleration, achieving 2-3$\times$ near-human-level stability. This allows successful deployment of tasks like walking while carrying liquids or recording stable video on the Unitree G1 humanoid and enabling humanoid robots to perform complex tasks with precision and reliability.

6 Limitation

Despite its strong performance, SoFTA still faces several limitations. First, while it significantly reduces EE acceleration, the achieved stability still falls short of human-level performance. Carrying a cup of water while walking is a task that humans can perform effortlessly with minimal spill. SoFTA yet match the subtlety and adaptability of human control. Second, the decoupling of locomotion and end-effector control creates a fixed task boundary. While this separation is effective for many loco-manipulation tasks, it becomes suboptimal when the two modules must closely coordinate, such as during dynamic reaching or complex interactions. Third, while SoFTA offers a flexible framework for many scenarios and introduces valuable insights on frequency assignment, its performance may vary depending on the specific task or robot configuration. Aspects like task complexity, robot morphology, or the need for more nuanced coordination may require further adjustments to the design.

Future work could focus on improving the adaptability of SoFTA to more diverse tasks and robot configurations, with particular attention to dynamic coordination and complex interactions. Additionally, addressing the human-level stability gap will be crucial, particularly in tasks requiring high precision and fine motor control. Exploring more advanced learning strategies and architecture, such as attention mechanism, could help achieve better generalization across various platforms and tasks.

Acknowledgments

We sincerely thank Zhengyi Luo and Zenghao Tang for their insightful discussions, which significantly contributed to shaping the direction of this work. We also thank Haotian Lin and Nikhil Sobanbabu for their help with real-world experiments. We gratefully acknowledge the hardware support provided by Unitree Robotics and Booster Robotics.

References

[1] Q. Liao, B. Zhang, X. Huang, X. Huang, Z. Li, and K. Sreenath. Berkeley humanoid: A research platform for learning-based control. *arXiv preprint arXiv:2407.21781*, 2024.

[2] Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath. Reinforcement learning for versatile, dynamic, and robust bipedal locomotion control. *The International Journal of Robotics Research*, page 02783649241285161, 2024.

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

[4] I. Radosavovic, B. Zhang, B. Shi, J. Rajasegaran, S. Kamat, T. Darrell, K. Sreenath, and J. Malik. Humanoid locomotion as next token prediction. arxiv. 2024. *arXiv preprint arXiv:2402.19469*, 2024.

[5] 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.

[6] 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)*, pages 11225–11231. IEEE, 2024.

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

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

[9] H. Wang, Z. Wang, J. Ren, Q. Ben, T. Huang, W. Zhang, and J. Pang. Beamdojo: Learning agile humanoid locomotion on sparse footholds, 2025. URL https://arxiv.org/abs/2502.10363.

[10] J. Ren, T. Huang, H. Wang, Z. Wang, Q. Ben, J. Pang, and P. Luo. Vb-com: Learning vision-blind composite humanoid locomotion against deficient perception, 2025. URL https://arxiv.org/abs/2502.14814.

[11] W. Xie, C. Bai, J. Shi, J. Yang, Y. Ge, W. Zhang, and X. Li. Humanoid whole-body locomotion on narrow terrain via dynamic balance and reinforcement learning, 2025. URL https://arxiv.org/abs/2502.17219.

[12] R.-Z. Qiu, S. Yang, X. Cheng, C. Chawla, J. Li, T. He, G. Yan, L. Paulsen, G. Yang, S. Yi, et al. Humanoid policy˜ human policy. *arXiv preprint arXiv:2503.13441*, 2025.

[13] T. Lin, K. Sachdev, L. Fan, J. Malik, and Y. Zhu. Sim-to-real reinforcement learning for vision-based dexterous manipulation on humanoids, 2025. URL https://arxiv.org/abs/2502.20396.

[14] J. Li, Y. Zhu, Y. Xie, Z. Jiang, M. Seo, G. Pavlakos, and Y. Zhu. Okami: Teaching humanoid robots manipulation skills through single video imitation, 2024. URL https://arxiv.org/abs/2410.11792.

[15] S. Atar, X. Liang, C. Joyce, F. Richter, W. Ricardo, C. Goldberg, P. Suresh, and M. Yip. Humanoids in hospitals: A technical study of humanoid surrogates for dexterous medical interventions, 2025. URL https://arxiv.org/abs/2503.12725.

[16] X. Shu, F. Ni, X. Fan, S. Yang, C. Liu, B. Tu, Y. Liu, and H. Liu. A versatile humanoid robot platform for dexterous manipulation and human–robot collaboration. *CAAI Transactions on Intelligence Technology*, 9(2):526–540, 2024. doi:https://doi.org/10.1049/cit2.12214. URL https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/cit2.12214.

[17] Z. Gu, J. Li, W. Shen, W. Yu, Z. Xie, S. McCrory, X. Cheng, A. Shamsah, R. Griffin, C. K. Liu, A. Kheddar, X. B. Peng, Y. Zhu, G. Shi, Q. Nguyen, G. Cheng, H. Gao, and Y. Zhao. Humanoid locomotion and manipulation: Current progress and challenges in control, planning, and learning, 2025. URL https://arxiv.org/abs/2501.02116.

[18] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3803–3810. IEEE, 2018.

[19] 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.

[20] X. Chen, J. Hu, C. Jin, L. Li, and L. Wang. Understanding domain randomization for sim-to-real transfer, 2022. URL https://arxiv.org/abs/2110.03239.

[21] T. He, J. Gao, W. Xiao, Y. Zhang, Z. Wang, J. Wang, Z. Luo, G. He, N. Sobanbab, C. Pan, Z. Yi, G. Qu, K. Kitani, J. Hodgins, L. J. Fan, Y. Zhu, C. Liu, and G. Shi. Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills, 2025. URL https://arxiv.org/abs/2502.01143.

[22] T. Li, H. Geyer, C. G. Atkeson, and A. Rai. Using deep reinforcement learning to learn high-level policies on the atrias biped. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 263–269. IEEE, 2019.

[23] Z. Xie, P. Clary, J. Dao, P. Morais, J. Hurst, and M. Panne. Learning locomotion skills for cassie: Iterative design and sim-to-real. In *Conference on Robot Learning*, pages 317–329. PMLR, 2020.

[24] Z. Li, X. Cheng, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath. Reinforcement learning for robust parameterized locomotion control of bipedal robots. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2811–2817. IEEE, 2021.

[25] 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. *arXiv preprint arXiv:2410.21229*, 2024.

[26] T. He, Z. Luo, W. Xiao, C. Zhang, K. Kitani, C. Liu, and G. Shi. Learning human-to-humanoid real-time whole-body teleoperation. *arXiv preprint arXiv:2403.04436*, 2024.

[27] C. Lu, X. Cheng, J. Li, S. Yang, M. Ji, C. Yuan, G. Yang, S. Yi, and X. Wang. Mobile-television: Predictive motion priors for humanoid whole-body control. *arXiv preprint arXiv:2412.07773*, 2024.

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

[29] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. Kitani, C. Liu, and G. Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *arXiv preprint arXiv:2406.08858*, 2024.

[30] Q. Ben, F. Jia, J. Zeng, J. Dong, D. Lin, and J. Pang. Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit, 2025. URL https://arxiv.org/abs/2502.13013.

[31] H. Shi, W. Wang, S. Song, and C. K. Liu. Toddlerbot: Open-source ml-compatible humanoid platform for loco-manipulation, 2025. URL https://arxiv.org/abs/2502.00893.

[32] J. Shi, X. Liu, D. Wang, O. Lu, S. Schwertfeger, F. Sun, C. Bai, and X. Li. Adversarial locomotion and motion imitation for humanoid policy learning, 2025. URL https://arxiv.org/abs/2504.14305.

[33] B. U. Rehman, M. Focchi, J. Lee, H. Dallali, D. G. Caldwell, and C. Semini. Towards a multi-legged mobile manipulator. pages 3618–3624, 2016.

[34] L. Sentis and O. Khatib. Synthesis of whole-body behaviors through hierarchical control of behavioral primitives. *International Journal of Humanoid Robotics*, 2(04):505–518, 2005.

[35] H. Ferrolho, V. Ivan, W. Merkt, I. Havoutis, and S. Vijayakumar. RoLoMa: Robust loco-manipulation for quadruped robots with arms. *Autonomous Robots*, 47(8):1463–1481, 2023.

[36] H. Ferrolho, W. Merkt, V. Ivan, W. Wolfslag, and S. Vijayakumar. Optimizing Dynamic Trajectories for Robustness to Disturbances Using Polytopic Projections. pages 7477–7484, 2020.

[37] L. Shi, X. Yu, C. Zhou, W. Jin, W. Chi, S. Zhang, D. Zhang, X. Li, and Z. Zhang. Whole-body impedance coordinative control of wheel-legged robot on uncertain terrain, 2024. URL https://arxiv.org/abs/2411.09935.

[38] J. Pankert and M. Hutter. Perceptive model predictive control for continuous mobile manipulation. *IEEE Robotics and Automation Letters*, 5(4):6177–6184, 2020. doi:10.1109/LRA.2020.3010721.

[39] M. Osman, M. W. Mehrez, S. Yang, S. Jeon, and W. Melek. End-effector stabilization of a 10-dof mobile manipulator using nonlinear model predictive control, 2021. URL https://arxiv.org/abs/2103.13153.

[40] M. V. Minniti, F. Farshidian, R. Grandia, and M. Hutter. Whole-body mpc for a dynamically stable mobile manipulator. *IEEE Robotics and Automation Letters*, 4(4):3687–3694, Oct. 2019. ISSN 2377-3774. doi:10.1109/lra.2019.2927955. URL http://dx.doi.org/10.1109/LRA.2019.2927955.

[41] Y. Ma, F. Farshidian, T. Miki, J. Lee, and M. Hutter. Combining Learning-Based Locomotion Policy With Model-Based Manipulation for Legged Mobile Manipulators. 7(2):2377–2384, 2022.

[42] M. Liu, Z. Chen, X. Cheng, Y. Ji, R. Yang, and X. Wang. Visual Whole-Body Control for Legged Loco-Manipulation. 2024.

[43] G. Pan, Q. Ben, Z. Yuan, G. Jiang, Y. Ji, S. Li, J. Pang, H. Liu, and H. Xu. RoboDuet: Whole-body Legged Loco-Manipulation with Cross-Embodiment Deployment, 2024.

[44] Z. Fu, X. Cheng, and D. Pathak. Deep Whole-Body Control: Learning a Unified Policy for Manipulation and Locomotion. pages 138–149, 2023.

[45] T. Portela, A. Cramariuc, M. Mittal, and M. Hutter. Whole-body end-effector pose tracking, 2025. URL https://arxiv.org/abs/2409.16048.

[46] C. Zhang, W. Xiao, T. He, and G. Shi. Wococo: Learning whole-body humanoid control with sequential contacts, 2024. URL https://arxiv.org/abs/2406.06005.

[47] K. Zhang, Z. Yang, and T. Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms, 2021. URL https://arxiv.org/abs/1911.10635.

[48] Z. Zhuang and H. Zhao. Embrace collisions: Humanoid shadowing for deployable contact-agnostics motions. *ArXiv*, abs/2502.01465, 2025. URL https://api.semanticscholar.org/CorpusID:276107247.

[49] 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, 2025. URL https://arxiv.org/abs/2502.08378.

[50] Y. Guo, Z. Jiang, Y.-J. Wang, J. Gao, and J. Chen. Decentralized motor skill learning for complex robotic systems, 2023. URL https://arxiv.org/abs/2306.17411.

[51] C. Lu, X. Cheng, J. Li, S. Yang, M. Ji, C. Yuan, G. Yang, S. Yi, and X. Wang. Mobile-television: Predictive motion priors for humanoid whole-body control, 2025. URL https://arxiv.org/abs/2412.07773.

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

[53] A. Rajeswaran, S. Ghotra, B. Ravindran, and S. Levine. Epopt: Learning robust neural network policies using model ensembles, 2017. URL https://arxiv.org/abs/1610.01283.

[54] 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*, 4(26), Jan. 2019. ISSN 2470-9476. doi:10.1126/scirobotics.aau5872. URL http://dx.doi.org/10.1126/scirobotics.aau5872.

[55] Y. Yang, K. Caluwaerts, A. Iscen, T. Zhang, J. Tan, and V. Sindhwani. Data efficient reinforcement learning for legged robots, 2019. URL https://arxiv.org/abs/1907.03613.

[56] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, and T. Graepel. Value-decomposition networks for cooperative multi-agent learning, 2017. URL https://arxiv.org/abs/1706.05296.

[57] S. Iqbal and F. Sha. Actor-attention-critic for multi-agent reinforcement learning. *CoRR*, abs/1810.02912, 2018. URL http://arxiv.org/abs/1810.02912.

[58] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *CoRR*, abs/1706.02275, 2017. URL http://arxiv.org/abs/1706.02275.

[59] H. Yarahmadi, M. E. Shiri, H. Navidi, A. Sharifi, and M. Challenger. Bankruptcy-evolutionary games based solution for the multi-agent credit assignment problem. *Swarm Evol. Comput.*, 77:101229, 2023. URL https://api.semanticscholar.org/CorpusID:255700663.

[60] U. Robotics. Unitree g1 humanoid agent ai avatar, 2024. URL https://www.unitree.com/g1.

[61] C. L. Lab. Humanoidverse: A multi-simulator framework for humanoid robot sim-to-real learning. https://github.com/LeCAR-Lab/HumanoidVerse, 2025.

[62] B. Robotics. Booster t1 humanoid robot, 2025. URL https://www.boosterobotics.com/.

A Appendix

A.1 Training Details

Observation We adopt an asymmetric observation structure to enable efficient policy learning in simulation while ensuring robust real-world deployment under partial observability. The actor relies solely on onboard-accessible inputs—proprioception, command signals, and recent actions—excluding global position data, thus removing dependence on odometry or external tracking. Observations are stacked over five timesteps to provide short-term temporal context.

Type	Observation	Actor	Critic	Scale	Noise Scale
Privileged	base_lin_vel	✗	✓	2.0	0.0
	end_effector_relative_pos	✗	✓	1.0	0.0
	end_effector_gravity	✗	✓	1.0	0.0
Proprioception	base_ang_vel	✓	✓	0.25	0.1
	projected_gravity	✓	✓	1.0	0.0
	dof_pos	✓	✓	1.0	0.01
	dof_vel	✓	✓	0.05	0.1
	actions	✓	✓	1.0	0.0
	sin_phase / cos_phase	✓	✓	1.0	0.0
Command	command_lin_vel	✓	✓	1.0	0.0
	command_ang_vel	✓	✓	1.0	0.0
	command_EE	✓	✓	1.0	0.0
	command_gait	✓	✓	1.0	0.0

Table 4: Comparison of actor and critic observations with scaling factors. Privileged observations used only by the critic are shaded and marked in red.

During training, the critic is granted privileged access to additional information, including `base_lin_vel`, `end_effector_relative_pos`, and `end_effector_gravity`, which help robot to understand its current state and task success more accurately. To improve robustness, noise is injected into selected observations. Observation scales and noise scales are summarized in Table 4.

This setup improves value estimation and training stability while ensuring deployable policies grounded in realistic sensor inputs, supporting robust sim-to-real transfer for locomotion and end-effector tasks.

Task Definition We define our task as a combination of robust locomotion and end-effector (EE) stabilization under general body configurations. The EE stabilization command, denoted as $c^{\text{EE}} \in \mathbb{R}^5$, encodes task-specific requirements. The first dimension is a binary flag indicating whether EE stabilization is enabled. If this value is zero, all stabilization-related rewards are disabled for that EE. The next two values specify the desired EE position (x, y) in the local frame of the body. The fourth value defines the target EE height along the global z-axis, given as an offset relative to the desired base z-position. The final element of c^{EE} is a tolerance parameter σ_{EE} that controls the precision of EE tracking. A higher tolerance leads to smoother motion with lower accelerations, which is beneficial for tasks such as bottle carrying where precise EE positioning is not critical. Conversely, a lower tolerance prioritizes accurate tracking, which is essential for tasks like camera stabilization where EE pose must be tightly maintained.

For locomotion, the control command includes both target base velocity and gait information. The velocity command comprises desired linear velocities (v_x, v_y) and angular velocity ω, all defined in the base frame. The system is expected to track these velocities within specified tolerances $\sigma_x, \sigma_y, \sigma_\omega$. Gait control is represented by a two-dimensional vector. The first value is a binary indicator of whether the desired gait is a double-stance (both feet in contact). If not (i.e., in dynamic gait mode), the second value specifies the desired gait period. From this gait period, we compute

the gait phase using sinusoidal signals $(\sin(\phi), \cos(\phi))$, where ϕ denotes the phase. This allows the derivation of target contact timings for each foot $\hat{C}$. A phase-based reward is then introduced to guide the agent to follow the desired contact sequence $\hat{C}$.

We list all command ranges in Table 5, with $\sigma_x = 0.5$m/s, $\sigma_y = 0.5$m/s, $\sigma_\omega = 0.5$rad/s respectively.

Component	Range / Value
`command_lin_vel`	x: $\mathcal{U}$ (-1, 1) m/s y: $\mathcal{U}$ (-1, 1) m/s
`command_ang_vel`	$\mathcal{U}$ (-1, 1) rad/s
`command_gait`	mode: 0/1 period: $\mathcal{U}$ (0.5, 1.3) m/s
`command_EE`	activation: 0/1 x: $\mathcal{U}$ (-0.1, 0.1) m y: $\mathcal{U}$ (-0.1, 0.1) m z: $\mathcal{U}$ (-0.1, 0.1) m tolerance: $\mathcal{U}$ (0.1, 0.2) m

Table 5: Command ranges used during training.

Domain Randomization To enhance the robustness and generalization of SoFTA, we apply domain randomization techniques, as detailed in Table 6. We first train SoFTA with all domain randomization strategies listed, excluding push perturbations. After obtaining a stable policy, we introduce push disturbances to further improve robustness under external disturbance.

Component	Range / Value
P Gain	$\mathcal{U}$ (0.95, 1.05) × `default`
D Gain	$\mathcal{U}$ (0.95, 1.05) × `default`
Friction Coefficient	$\mathcal{U}$ (-0.5, 1.25)
Base Mass	$\mathcal{U}$ (-1.0, 3.0) `kg`
Control Delay	$\mathcal{U}$ (20, 40) `ms`
Push Perturbations	`Interval:` $\mathcal{U}$ (5, 16) `s` `Max velocity:` 0.5 `m/s`

External Force on EE	`Position std:` 0.03 `m`
	`X force:` $\mathcal{U}$ (-0.5, 0.5) `N`
	`Y force:` $\mathcal{U}$ (-0.5, 0.5) `N`
	`Z force:` $\mathcal{U}$ (-7, 2) `N`

Table 6: Domain randomization parameters used during training.

Rewards Design We show the grouped SoFTA task reward components in Table 7. Notice that the termination is a shared reward component Also, we introduce several penalties and energy regularization in order to achieve robust sim-to-real performance like *dof limit, stand symmetry, contact force, feet height on the air, action rate* and so on. Follow [21], We adjust the scaling factor $s_{t,i}$ in the cumulative discounted reward formula to handle small rewards differently based on their sign: $\mathbb{E}\left[\sum_{t=1}^{T} \gamma^{t-1} \sum_i s_{t,i} r_{t,i}\right]$, where $s_{t,i} = s_{\text{current}}$ if $r_{t,i} < 0$, and 1 if $r_{t,i} \geq 0$. The factor s_{current} starts at 0.5 and is adjusted dynamically—multiplied by 0.9999 when episode length is under 0.4s, and by 1.0001 when it exceeds 2.1s, with an upper bound of 1. This allows our policy to first focus on task terms and them regular the behavior to be smooth and reasonable for sim-to-real.

Group	Term	Weight	Expression
Lower Body	`tracking_lin_vel_x`	1.5	$\exp(-\frac{1}{\sigma_x^2}\|v_x - \hat{v}_x\|^2)$
	`tracking_lin_vel_y`	1.0	$\exp(-\frac{1}{\sigma_y^2}\|v_y - \hat{v}_y\|^2)$
	`tracking_ang_vel`	2.0	$\exp(-\frac{1}{\sigma_\omega^2}\|\omega_z - \hat{\omega}_z\|^2)$
	`tracking_base_height`	0.5	$\exp(-\frac{1}{\sigma_h}\|h - \hat{h}\|)$
	`tracking_gait_contact`	0.5	$\sum(\mathbb{1}(C = \hat{C}) - \mathbb{1}(C \neq \hat{C}))$
	`termination`	-500.0	$\mathbb{1}_{\text{terminate}}$
Upper Body	`tracking_end_effector_pos`	1.0	$\exp\left(-\frac{1}{\sigma_{EE}^2}\|p_{\text{EE}} - \hat{p}_{\text{EE}}\|^2\right)$
	`tracking_zero_end_effector_acc`	10	$\exp\left(-\lambda_{\text{acc}}\|\ddot{p}_{\text{EE}}\|^2\right)$
	`tracking_zero_end_effector_ang_acc`	1.5	$\exp\left(-\lambda_{\text{ang-acc}}\|\dot{\omega}_{\text{EE}}\|^2\right)$
	`penalty_end_effector_acc`	-0.1	$-\|\ddot{p}_{\text{EE}}\|^2$
	`penalty_end_effector_ang_acc`	-0.01	$-\|\dot{\omega}_{\text{EE}}\|^2$
	`penalty_end_effector_tilt`	-5.0	$-\left\|\mathbf{P}_{xy}(R_{\text{EE}}^T\mathbf{g})\right\|^2$
	`termination`	-100.0	$\mathbb{1}_{\text{terminate}}$

Table 7: Reward terms categorized by body group, including task rewards and penalties with corresponding expressions and weights. C means the contact sequence. Hat over variables represents the desired value. In implementation, we set $\lambda_{acc} = 0.25$, $\lambda_{acc} = 0.0044$.

Training Hyperparameter We summarize the main hyperparameters used in our PPO multi-actor-critic training setup in Table 8. These include general PPO settings, action std for different body modules, and the network architecture shared across policy and value networks.

Parameter	**Value**
General PPO Settings	
Clip Parameter	0.2
Gamma (γ)	0.99
GAE Lambda (λ)	0.95
Value Loss Coef	1.0
Entropy Coef	0.01
Actor Learning Rate	1×10^{-3}
Critic Learning Rate	1×10^{-3}
Max Grad Norm	1.0
Use Clipped Value Loss	True
Desired KL	0.01
Num Steps per Env	48
Noise Settings	
Init Noise Std	lower_body: 0.8, upper_body: 0.6
Std Threshold	lower_body: 0.15, upper_body: 0.10
Network Architecture	
Hidden Layers	[512, 256, 128]
Activation Function	ELU

Table 8: PPO Multi-Actor-Critic Training Configuration

A.2 More Analysis on Frequency Ablation

Methods	Response Time (s) ↓	Max Acc (m/s^2) ↓	Max Vel (m/s) ↓
Ours (L50-U33)	0.598	43.5	1.90
Ours (L50-U50)	0.338	40.5	1.48
Ours (L50-U100)	**0.167**	**37.8**	**1.17**

Table 9: Response time and maximum error magnitudes under different upper-body frequencies.

Across both simulation and real-world environments, our experiments show that a 50 Hz lower-body control frequency consistently achieves stable locomotion, regardless of the upper-body control frequency, whereas other lower-body frequencies may lead to degraded performance under certain upper-body control frequencies. We further investigated how higher upper-body frequencies enhance EE stability in challenging scenarios, such as sudden external pushes. As shown in Figure 9 (top), higher-frequency policies (100 Hz) react faster to base motion changes and recover balance quicker. In Figure 9 (bottom), we observe that higher frequencies lead to faster EE velocity recovery. Table 9 presents the quantitative results. We observe that increasing the upper-body control frequency reduces recovery time (defined as the time when the error first falls below $\frac{1}{e}$ of its maximum), as well as peak acceleration and velocity errors. This indicates enhanced disturbance compensation and faster recovery dynamics.

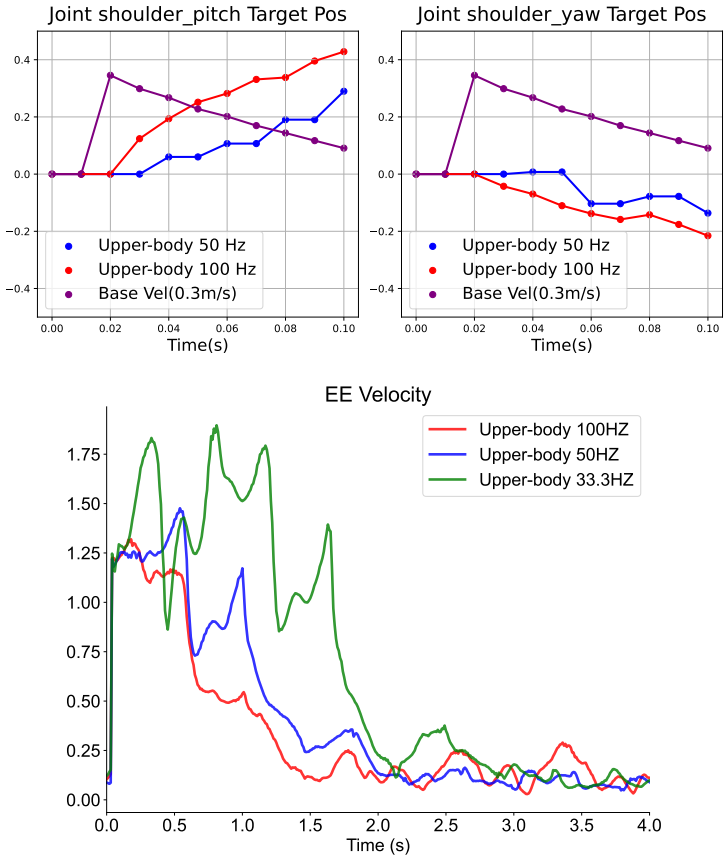


Figure 9: Effect of upper-body control frequency on EE stabilization. top: EE velocity (m/s) recovery with different upper-body frequencies. bottom: Response comparison at 100 Hz vs. 50 Hz.