2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

SEEC: Stable End-Effector Control with Model-Enhanced Residual Learning for Humanoid Loco-Manipulation

Jaehwi Jang*[1], Zhuoheng Wang*[1,2], Ziyi Zhou[1], Feiyang Wu[1], and Ye Zhao[1]
*Equal Contribution [1]Georgia Institute of Technology [2]Tsinghua University

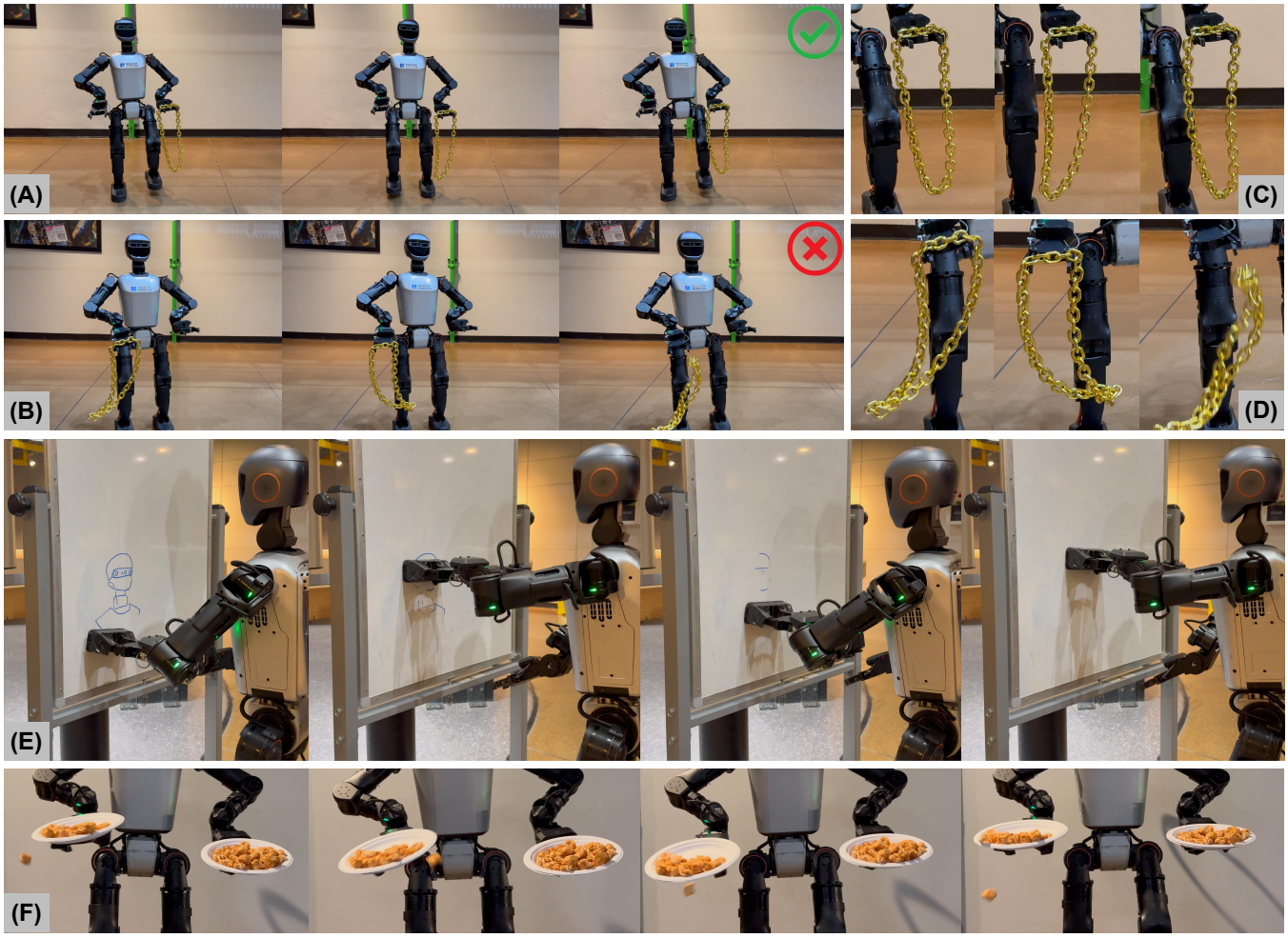


Fig. 1: Our SEEC framework enables a Booster T1 humanoid robot to perform stable loco-manipulation tasks while dynamic locomotion. Demonstrated skills include (A-D) holding a flexible chain while walking, (E) wiping a whiteboard surface with teleoperation, and (F) carrying a plate of snacks while walking. In (A-D), our SEEC framework (A, C) enables the robot to firmly hold the chain and suppress oscillatory dynamics, whereas the IK baseline (B, D) induces large oscillations that eventually cause the chain to drop. In (F), our SEEC framework (left arm) successfully kept the snacks in the plate, whereas the IK baseline (right arm) failed and dropped the snacks.

***Abstract*— Arm end-effector stabilization is essential for humanoid loco-manipulation tasks, yet it remains challenging due to the high degrees of freedom and inherent dynamic instability of bipedal robot structures. Previous model-based controllers achieve precise end-effector control but rely on precise dynamics modeling and estimation, which often struggle to capture real-world factors (e.g., friction and backlash) and thus degrade in practice. On the other hand, learning-based methods can better mitigate these factors via exploration and domain randomization, and have shown potential in real-world use. However, they often overfit to training conditions, requiring retraining with the entire body, and still struggle to adapt to unseen scenarios. To address these challenges, we propose a novel stable end-effector control (SEEC) framework with model-enhanced residual learning that learns to achieve precise and robust end-effector compensation for lower-body induced disturbances through model-guided reinforcement learning (RL) with a perturbation generator. This design allows the upper-body policy to achieve accurate end-effector stabilization as well as adapt to unseen locomotion controllers with no additional training. We validate our framework in different simulators and transfer trained policies to the Booster T1 humanoid robot. Experiments demonstrate that our method consistently outperforms baselines and robustly handles diverse and demanding loco-manipulation tasks. More details and videos are available at: https://seec-humanoid.github.io.**

979-8-3315-8160-2/26/$31.00 ©2026 IEEE

I. INTRODUCTION

Humanoid robots promise seamless integration into human environments, where they must walk and manipulate simultaneously. From carrying objects while moving to performing collaborative tasks [1]–[3], this capability is fundamental to practical humanoid deployment (see Fig. 1). Yet, achieving stable and precise arm end-effector control during dynamic locomotion remains an open challenge. Even modest base movements could induce large end-effector accelerations, causing tracking errors, destabilizing contact forces, and limiting humanoid utility in real-world settings.

Recently, learning-based approaches [4]–[7] have sought to achieve humanoid loco-manipulation by training end-to-end reinforcement learning (RL) policies. While effective at capturing nonlinearities and handling uncertainty, these policies often rely on imitating joint or task reference trajectories [8]–[10], and struggle to ensure accurate end-effector stabilization. For example, in HOVER [10], uncontrolled hand motions emerge as a byproduct of locomotion. Prior work [7] has attempted to stabilize end-effector control by directly penalizing its acceleration, but this approach heavily relies on policy optimization to "discover" an appropriate compensation strategy. Additionally, the learned behavior degenerates into static hand-holding motions, limiting general applicability. Moreover, when tasks require reactive whole-body coordination, sudden locomotion disturbances further exacerbate end-effector instability. Conventional RL training, as in [7], fails to provide robustness under such out-of-distribution (OOD) scenarios.

Inspired by model-based approaches [11]–[15], which achieve precise stabilization through dynamics modeling and online estimation, we introduce SEEC: **S**table **E**nd-**E**ffector **C**ontrol, a model-enhanced RL framework for humanoid loco-manipulation. SEEC leverages model-based expertise to provide analytic acceleration compensation signals during training. Instead of relying on naive penalization, the compensation torque from the model-based formulation is distilled into the RL policy, addressing instability in a more principled manner.

Furthermore, unlike prior works that jointly train manipulation and locomotion policies, we introduce a *perturbation generation strategy* that exposes the upper-body policy to a wide spectrum of locomotion-induced disturbances. By modeling these disturbances as base movement patterns, the upper-body controller learns to maintain stable arm end-effector control independent of any specific locomotion policy. This modular design not only improves robustness across diverse walking patterns, allowing seamless transfer across different walking patterns, including previously unseen locomotion controllers, but also facilitates integration into complex loco-manipulation tasks that demand coherent whole-body coordination.

Our core contributions can be summarized as follows.

- We propose a model-enhanced residual learning framework that integrates model-based expertise with learning-based adaptability, achieving precise acceleration compensation while effectively addressing model inaccuracies and parameter uncertainties.
- We introduce a *base-movement data generation and perturbation generation strategy* that exposes the policy to a broad spectrum of locomotion-relevant disturbances during training. This enables the controller to acquire robust compensation behaviors that *can transfer to unseen locomotion controllers and gaits* without requiring joint re-training.
- We demonstrate the first deployment of such a hybrid framework on a full humanoid Booster T1, validating it both in simulation and on the real hardware via zero-shot transfer. The system achieves more stable and precise end-effector control across a variety of loco-manipulation tasks, compared to the baselines.

II. RELATED WORK

Traditional works on whole-body controllers for legged and mobile manipulators rely on model-based methods, which often employ numerical optimization for precise control [16]–[19]. Although effective, these approaches depend on accurate dynamics model and contact scheduling, which are difficult to maintain in complex or unstructured environments. In contrast, learning-based methods have rapidly advanced humanoid whole-body control, driven by reinforcement learning (RL) and imitation learning (IL) [1], [2], [4], [6], [9], [20]–[22]. These frameworks produce expressive and robust behaviors, but accurate and stable end-effector control remains a fundamental challenge. The difficulty arises because locomotion-induced disturbances rapidly amplify tracking errors, especially during agile maneuvers or in dynamic environments. SoFTA [7] attempts to stabilize the end-effector by penalizing its acceleration in the reward function. However, this approach often degenerates into static hand-holding behaviors and generalizes poorly across loco-manipulation tasks. By contrast, we show that compensating for the torque induced by locomotion disturbances enables effective stabilization across diverse walking motions, including motions produced by controllers not seen during training.

Effective control of manipulators or arms on mobile/legged robots is crucial to achieve expressive motion and complex task execution. To manage complexity, many recent works adopt a decoupled architecture, splitting into upper-body and lower-body modules [4], [7], [9], [15], [23]. Ma et al. [15] model the influence of the manipulator on locomotion as a disturbance, training the locomotion controller to compensate. While this improves gait stability, it sidesteps the harder problem of stabilizing the arm under dynamic base motion. Moreover, existing frameworks [7] jointly train locomotion and manipulation policies, coupling them tightly. This prevents modular reuse and limits robustness: when deployed with different locomotion controllers, or in the face of the inevitable dynamically changing real-world environments, the upper-body policy cannot adapt to unseen disturbances and often fails. Our framework departs from this paradigm. We explicitly model lower-to-upper body coupling and introduce a *perturbation generation strategy*, which

independently trains the upper-body controller to compensate for a wide range of locomotion-induced disturbances. This design enables stable and precise end-effector control that handles unseen perturbations in the real world and even maintains performance across locomotion controllers.

Recent works have explored how to combine MPC and RL controllers [1]. One line of remarkable research uses a model-based controller or trajectory optimizer to supervise learning, where the RL policies imitate expert actions [8], [24]–[26]. Another line blends outputs from RL and MPC, using MPC for constraint satisfaction and RL for adaptability [27], [28]. These approaches improve sample efficiency and stability, but are generally evaluated on flat terrain or simplified tasks and rarely address the unique challenge of maintaining arm end-effector stability during dynamic loco-manipulation. In this work, we adopt a residual policy learning approach [27], [29], but tailored for humanoid loco-manipulation, achieving robust end-effector stabilization.

III. METHOD

We formulate our control problem as the coordination of two controllers: (i) a *lower-body controller* responsible for locomotion, and (ii) an *upper–body controller* responsible for manipulation tasks. Both policies are trained in IsaacLab [30] and modeled as Markov Decision Processes (MDPs). At time t, the agent (each policy) receives observation o_t, and then samples an action $a_t \sim \pi(\cdot|o_t)$ according to policy π and transitions to a new observation o_{t+1} while receiving a reward $r(o_t, a_t)$. The goal of the agent is to maximize the expected return $\mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_t]$, where $\gamma \in [0, 1)$ is the discount factor.

In our framework, the lower-body policy is trained to achieve stable and robust locomotion, following conventional sim-to-real training pipelines for robust lower-body control works [31]–[33]. This allows it to handle diverse locomotion tasks without additional adaptation. The main difficulty lies in the upper–body control, which compensates for the disturbances induced by the freely moving base. To make this problem tractable, we introduce two assumptions:

Assumption 1: Negligible arm-to-base back–coupling. The arms are dynamically light relative to the lower body. Control actions in the arms induce negligible reaction forces on the base, allowing us to model base motion as an exogenous input when computing compensation torques.

Assumption 2: Robust locomotion controller. The locomotion controller is robust enough to maintain balance and track desired trajectories despite disturbances generated by arm movements and control torques.

With the first assumption, we can simplify the disturbance model by treating the base motion as independent of arm movements. This allows us to focus solely on compensating for external base motion and ignore the lower-body's dynamic response to the upper-body. The second assumption allows us to design a controller without torque constraints.

A. Model-Enhanced Residual Learning

We frame upper-body stabilization as controlling an arm subject to disturbances from a moving base. Our method trains an RL residual policy that actively counteracts these locomotion-induced disturbances through a three-stage pipeline: (i) simulate base-induced inertial effects in a physically consistent manner, (ii) compute the analytic compensation torque that can cancel the resulting arm end-effector accelerations, and (iii) distill these signals into policy via reward reshaping, guiding it to output a joint command that stabilizes the arm end-effector. The overall framework is illustrated in Fig. 2.

*1) **Simulated Base Acceleration**:* Directly applying spatial accelerations to a floating base in simulation is numerically unstable due to the nonlinear inverse dynamics problem. Instead, we emulate base motion on a fixed-base model by injecting the *equivalent fictitious wrench* that would be induced by a generic base twist $V_b = [v_b^\top; \omega_b^\top]^\top \in se(3)$ and spatial acceleration $A_b = [\dot{v}_b^\top; \dot{\omega}_b^\top]^\top \in \mathbb{R}^6$. Under Assumption 1, for each link of mass m and inertia I located at position r relative to the base with local velocity v, the induced inertial force and torque are $F_b = m(-\dot{v}_b - \dot{\omega}_b \times r - \omega_b \times (\omega_b \times r) - 2\,\omega_b \times v)$ and $T_b = -I\,\dot{\omega}_b - \omega_b \times (I\omega_b)$. The terms correspond respectively to linear, Euler, centrifugal, and Coriolis forces, plus angular-acceleration and gyroscopic torques. Together, they reproduce the accelerations experienced under real base motion, but can be applied stably to a fixed-base model.

To approximate locomotion-induced perturbations, we construct base acceleration signals from two characteristic sources: (i) an impulse acceleration signal from the foot-ground reaction force, and (ii) a rhythmic sway from the body's center of mass (CoM) shifting with each step [34]. We represent the composed signal as

$$\mathbf{A}_b(t) = \sum_{k=1}^{N} \left[\underbrace{\boldsymbol{p}_k \; g(t; T_k)}_{\substack{\text{Foot contact} \\ \text{impulse}}} + \underbrace{\boldsymbol{s}_k \; \sin\left(2\pi t / T_k + \phi_k\right)}_{\text{Periodic CoM swing}} \right], \quad (1)$$

where $g(t; T_k)$ is a Gaussian impulse with standard deviation $0.01\,\mathrm{s}$ and unit peak amplitude, $\boldsymbol{p}_k \in \mathbb{R}^6$ is the impulse amplitude, $\boldsymbol{s}_k \in \mathbb{R}^6$ is the oscillation amplitude, and ϕ_k is a phase offset. We sample disturbance parameters to ensure their coverage of a diverse range of realistic signals. The base motion periods $\{T_k\}_{k=1}^K$ are drawn from a log-uniform distribution in the range $[0.64\,\mathrm{s}, 1.28\,\mathrm{s}]$, covering the range of natural human-like gait cycles while avoiding bias toward short or long strides. Impulse amplitudes $\boldsymbol{p}_k$ are sampled from $[-100\,\mathrm{m/s^2}, 100\,\mathrm{m/s^2}]^6$, spanning strong ground-contact transients. Oscillation amplitudes $\boldsymbol{s}_k$ are sampled from $[-10\,\mathrm{m/s^2}, 10\,\mathrm{m/s^2}]^6$, representing lateral and vertical CoM sways. Phase offsets ϕ_k are sampled uniformly from $[-\pi, \pi]$ to generate diverse periodic acceleration profiles.

This random sampling procedure produces a rich set of disturbance profiles that capture the variability of realistic base acceleration signals. By repeatedly exposing the policy to such disturbances during training, we encourage it to learn compensation strategies that are robust to different walking styles, contact timings, and gait controllers.

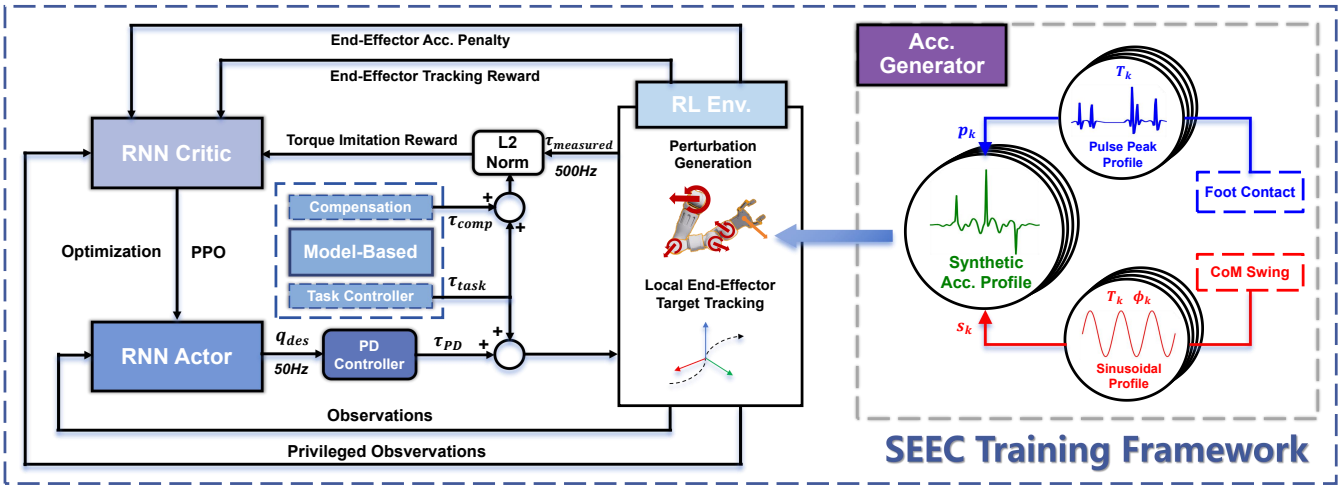


Fig. 2: **System framework overview of SEEC.** Our SEEC framework decouples the humanoid loco-manipulation controller into upper-body and lower-body controllers. The figure describes our core upper-body reinforcement learning (RL) module, which trains a residual policy that compensates lower-body induced disturbances. We leverage model-based acceleration compensation signals to guide RL training, ensuring more principled end-effector stability than naive penalization. Additionally, we generate base acceleration profiles to simulate external perturbations to promote robustness to unseen locomotion controllers. For the deployment, we transfer trained upper-body and lower-body policies to the robot without additional joint training.

*2) **Compensation Torque**:* In a fixed-base model, the local end–effector acceleration is $a_{ee}^{\text{loc}} = J(q)\ddot{q} + \dot{J}(q)\dot{q}$, where $J(q) \in \mathbb{R}^{6\times n}$ is the end-effector Jacobian at configuration $q \in \mathbb{R}^n$, and $\dot{q} \in \mathbb{R}^n, \ddot{q} \in \mathbb{R}^n$ are the joint velocity and acceleration with n-DoF of the arm. With base motion (V_b, A_b), the global end–effector acceleration is $a_{ee}^{\text{glob}} = a_{ee}^{\text{base}} + a_{ee}^{\text{loc}}$, where $a_{ee}^{\text{base}} = \dot{v}_b + 2\,\omega_b \times v_{ee}^{\text{loc}} + \omega_b \times (\omega_b \times r_{ee}^{\text{loc}}) + \dot{\omega}_b \times r_{ee}^{\text{loc}}$ for given end-effector velocity v_{ee}^{loc} and displacement r_{ee}^{loc} in the local frame.

In addition, fictitious wrenches on the arm links (Sec. III-A.1) induce a local responsive acceleration

$$a_{\text{resp}} = J(q)M^{-1}(q)\sum_i^B J_i(q)^\top [F_b^{i\,\top}; T_b^{i\,\top}]^\top \tag{2}$$

where B is the number of linkages, $M(q)$ is the joint–space inertia matrix, $J_i(q) \in \mathbb{R}^{6\times n}$ is a Jacobian for link i and $[F_b^{i\,\top}; T_b^{i\,\top}]^\top \in \mathbb{R}^6$ denotes the fictitious wrench acting on a link i from the base.

To cancel these effects under Assumption 2, we compute a compensating acceleration a_{comp} by exerting a joint torque $\tau_{\text{comp}} \in \mathbb{R}^n$, such that $a_{ee}^{\text{base}} + a_{\text{resp}} + a_{\text{comp}} \approx 0$. Using the operational–space formulation, and the minimum-norm torque [35][1] renders

$$\tau_{\text{comp}} = -J(q)^\top \Lambda(q)\left(a_{ee}^{\text{base}} + a_{\text{resp}}\right), \tag{3}$$

where $\Lambda(q) \in \mathbb{R}^{6\times 6}$ is the operational–space inertia matrix. This torque is combined with task-oriented controller signal τ_{task} (e.g., operational space tracking [36] of x_{des} and $\dot{x}_{\text{des}}$) to stabilize the end-effector. i.e. $\tau_{\text{task}} = J(q)^\top[\Lambda(q)\left(\bar{K}_p(x_{\text{des}} - x) + \bar{K}_d(\dot{x}_{\text{des}} - \dot{x})\right) + Q(q,\dot{q}) + G(q)]$, where $Q(q,\dot{q})$ is the operational-space centrifugal/Coriolis term, $G(q)$ is the operational-space gravitational term, and $\bar{K}_p, \bar{K}_d$ are operational-space gains.

[1]While we adopt the minimum–norm solution, alternative formulations are possible that can incorporate a constraint solver with additional cost functions and constraints.

*3) **Compensation Residual Policy Training**:* Directly deploying the analytically computed compensation torque τ_{comp} on hardware is infeasible, due to sensor noise, missing angular acceleration signals on the hardware IMU, and sim-to-real gap, such as motor delays and friction. Instead, we train an RL policy which outputs joint targets to a low–level PD controller that matches the desired compensation torque τ_{comp}. The observation space consists of the history of an end-effector command and proprioception data - IMU data ($\dot{v}_b$, ω_b), upper-body joint states, and previous actions. Let $\tau_{\text{PD}} = K_p(q_{\text{des}} - q) + K_d(\dot{q}_{\text{des}} - \dot{q})$ be the torque generated by the PD loop from the policy's targets q_{des}, where target joint velocities are fixed to be zero ($\dot{q}_{\text{des}} = 0$). In addition, we add an operational-space tracking control torque τ_{task} to achieve ideal target-tracking behavior. To improve robustness, we inject observation noise and friction during training, and regularize both control effort and end-effector accelerations to discourage excessive actuation and jittering. The reward encourages the measured torque τ_{measured} to match the ideal torque $\tau_{\text{comp}} + \tau_{\text{task}}$:

$$r_\tau = -\left\|\tau_{\text{measured}} - (\tau_{\text{comp}} + \tau_{\text{task}})\right\|_2, \tag{4}$$

along with auxiliary rewards that penalize the global accelerations similar to [7], action smoothness, and tracking tolerances. The policy is trained with PPO [37] using recurrent actor–critic networks with hidden sizes $[256, 128, 128]$ to capture short-term temporal disturbances under partial observability. We train the policy on the left arm only, and at deployment mirror its joint commands to the corresponding right-arm joints for symmetric bimanual control.

In this work, we set an end-effector target in the local frame. In this case, the policy minimizes the local tracking error while stabilizing the end-effector under base perturbations. Although target tracking and stabilization objectives may conflict, we address this issue by adding a tolerance

Components	Equations	Weights (stddev.)
Alive	1	10
Position	$\exp(-\|r^{\text{cmd}} - r\|^2/\sigma_r^2)$*	10 (0.1)
Orientation	$\exp(-\|Q^{\text{cmd}} \ominus Q\|^2/\sigma_r^2)$*	10 (0.1)
Torque guide	$\|\tau_{\text{measured}} - (\tau_{\text{comp}} + \tau_{\text{task}})\|$	-0.1
exp form	$\exp\left(-\|\tau_{\text{measured}} - (\tau_{\text{comp}} + \tau_{\text{task}})\|\right)$	5
EE Lin. Acc.	$\|a_e\|$	-0.1
exp form	$\exp(-\|a_e\|^2/\sigma_a^2)$	1.0 (3.0)
EE Ang. Acc.	$\|\alpha_e\|$	-0.01
exp form	$\exp(-\|\alpha_e\|^2/\sigma_\alpha^2)$	1.0 (10)
Action rates	$\|a_{\text{prev}} - a_{\text{current}}\|$	-0.1

TABLE I: Summary of upper-body training rewards. *Note that for position and orientation tracking rewards, we assign a small tolerance of $0.05\,\text{m}$ and $0.1\,\text{rad}$ each. All norms in the table are L_2 norm. ($\ominus$: quaternion subtraction)

margin in the tracking reward functions described in Table. I, allowing policy to balance precise tracking with robust stabilization.

Finally, to achieve greater stability, the target can be specified in the world frame and converted into the local frame at runtime. However, this requires an accurate, real-time estimation of the robot's world pose, which we leave for future work.

B. Locomotion Training

Following state-of-the-art locomotion works [32], [33], [38], the policy observation space consists of four components: (i) clock signals (sine/cosine of gait phase), (ii) proprioception (base angular velocity, joint states, previous actions), (iii) base velocity command, and (iv) 5-step observation history for short-term memory. The action space controls 13 lower-body joints via target positions tracked by PD controllers.

For the design of reward functions, we build upon the formulations provided in Booster Gym [31], which include balance stability, smoothness, and velocity tracking task progress with carefully assigned weights. Robustness is improved by randomizing upper-body joint targets, end-effector mass, and environment parameters across episodes. Locomotion policies are trained in IsaacLab using PPO. Both actor and critic are MLPs with hidden sizes $[256, 128, 128]$.

IV. EXPERIMENTS AND RESULTS

In the experiments, we use the Booster T1 humanoid, which stands $1.2\,\text{m}$ tall and possesses 29 degrees of freedom.

A. Simulation Results

To systematically demonstrate the advantages of our framework, we address the following key questions:

Q1: Does our model-enhanced residual learning controller achieve superior end-effector stability?

Q2: How much does our perturbation generation strategy improve end-effector stability?

Q3: Does our perturbation generation strategy enable robust generalization to previously unseen locomotion controllers?

To address the above questions, we evaluate our proposed framework using two key metrics: (1) **end-effector stability**, measuring the effectiveness of acceleration compensation; (2) **robustness**, reflecting the capability to adapt across diverse and unseen locomotion skills.

Task	Method	LinAcc (m/s^2) ↓ mean	LinAcc (m/s^2) ↓ max	AngAcc (rad/s^2) ↓ mean	AngAcc (rad/s^2) ↓ max
Stepping	IK	5.73±0.15	15.2±0.29	18.1±0.58	74.3±1.44
	RL w/o Sim. Acc.	4.01±0.07	16.2±0.70	18.5±0.37	78.9±3.27
	RL w Sim. Acc.	3.91±0.05	16.8±1.14	18.8±0.32	69.6±3.90
	Ours w/o τ_{task}	3.35±0.02	9.48±0.34	16.4±0.09	78.4±1.79
	Ours w/o r_τ	2.42±0.04	6.40±0.29	13.3±0.38	69.3±11.7
	Ours w/o Sim. Acc.	2.60±0.04	8.41±0.44	13.5±0.30	57.8±0.34
	Ours	**2.26**±0.02	**5.92**±0.20	**11.9**±0.18	**56.9**±0.46
Forward	IK	5.28±0.27	15.5±0.63	17.3±0.35	77.4±4.78
	RL w/o Sim. Acc.	3.54±0.20	15.7±1.57	16.7±1.10	77.4±7.42
	RL w Sim. Acc.	3.41±0.06	12.3±0.85	16.3±0.51	58.6±3.95
	Ours w/o τ_{task}	2.67±0.15	7.04±0.58	12.0±0.54	47.3±1.85
	Ours w/o r_τ	2.38±0.08	6.59±1.64	13.1±0.55	59.3±7.70
	Ours w/o Sim. Acc.	2.45±0.02	8.46±0.33	12.3±0.10	**42.8**±0.14
	Ours	**2.29**±0.01	**5.56**±0.19	**11.4**±0.11	43.5±1.51
Lateral	IK	5.95±0.12	18.6±6.99	19.5±6.62	94.0±37.9
	RL w/o Sim. Acc.	4.18±0.13	17.4±0.59	19.2±0.55	83.8±1.75
	RL w Sim. Acc.	5.74±0.15	15.2±0.29	18.1±0.58	74.3±1.44
	Ours w/o τ_{task}	3.43±0.09	11.9±0.24	17.4±0.38	95.9±1.76
	Ours w/o r_τ	2.67±0.03	6.73±0.30	13.4±0.07	55.7±1.66
	Ours w/o Sim. Acc.	3.21±0.02	12.0±0.21	14.4±0.08	62.8±1.96
	Ours	**2.40**±0.00	**6.03**±0.63	**12.2**±0.06	**54.8**±0.50
Rotation	IK	6.06±0.50	16.7±2.17	21.0±2.41	89.4±8.40
	RL w/o Sim. Acc.	4.87±0.13	20.2±2.76	22.7±0.53	100.±5.38
	RL w Sim. Acc.	4.31±0.09	17.8±0.61	19.8±0.28	72.5±1.91
	Ours w/o τ_{task}	3.76±0.56	8.90±0.86	16.9±1.15	67.4±13.7
	Ours w/o r_τ	2.93±0.01	7.48±0.11	15.6±0.08	72.8±0.78
	Ours w/o Sim. Acc.	2.89±0.03	9.67±0.75	14.5±0.14	67.4±1.29
	Ours	**2.75**±0.01	**7.13**±0.95	**14.1**±0.09	**66.6**±1.44

TABLE II: Benchmark results (MuJoCo) on end-effector stability.

For **end-effector stability**, we compare the acceleration compensation of several baselines: (1) an *IK-based* approach (2) *learning-based* approaches *with or without* our perturbation generation (denoted as RL w or w/o Sim. Acc.), trained in a fixed-base scenario, and (3) our proposed SEEC framework *without* operational space torque, and (4) our SEEC framework *without* torque guide reward.

For **robustness**, we assess performance under different locomotion policies by comparing: (1) the *Pre-Train* framework, trained with a specific pretrained locomotion policy provided beforehand, (2) the *Co-Train* framework, where we adopt the training setup from [7], locomotion and manipulation policies are trained simultaneously, while having the same control frequency for fair comparison. For the evaluation, we replace the locomotion part of each trained framework with a new locomotion policy trained with Sec. III-B, and compare the end-effector stability. We denote this experiment as testing "With Unseen Locomotion Policy".

To ensure a comprehensive evaluation, we design a diverse set of simulation tasks that expose the robot to distinct locomotion scenarios and dynamic variations, including: (a) stepping in place, (b) forward walking at $0.4\,\text{m/s}$, (c) lateral walking at $0.4\,\text{m/s}$, and (d) rotational walking at $0.4\,\text{rad/s}$. For each scenario, we perform three roll-outs and record the end-effector acceleration in the world frame. Additionally, for a fair comparison, all the methods share the same K_p and K_d gains for low-level PD control: 10.0 and 0.5, respectively.

Results and Analysis: As shown in Table II, our proposed SEEC framework outperforms the baselines in end-effector acceleration stability across most locomotion tasks. Note that removing either the operational space torque component or the torque-guided reward substantially degrades the perfor-

		With Trained Locomotion Policy				With Unseen Locomotion Policy			
		LinAcc (m/s^2)		AngAcc (rad/s^2)		LinAcc (m/s^2)		AngAcc (rad/s^2)	
Task	Method	mean	max	mean	max	mean	max	mean	max
Stepping	RL (Pre-Train)	6.57±0.27	18.2±1.39	28.0±0.34	84.4±7.94	–	–	–	–
	RL (Co-Train)	5.81±0.11	26.2±1.07	24.8±0.44	143.±9.38	10.6±0.04	25.7±0.46	48.5±0.44	173.±5.88
	Ours (w Pre-Train Loco. Policy)	3.27±0.28	9.89±0.70	17.5±0.69	92.7±17.1	5.32±0.02	18.4±0.69	26.1±0.25	94.0±1.13
	Ours (w Co-Train Loco. Policy)	3.07±1.75	11.5±0.42	20.4±0.52	158.±1.17				
Forward	RL (Pre-Train)	5.30±0.30	12.4±0.29	21.3±0.44	63.2 ±6.79	–	–	–	–
	RL (Co-Train)	5.36±0.06	22.5±0.39	23.9±0.24	142.±0.87	8.29±0.05	23.7±1.87	34.1±0.13	177.±11.56
	Ours (w Pre-Train Loco. Policy)	3.44±0.14	11.5±1.07	17.4±1.86	90.1±19.0	4.76±0.06	18.0±1.76	24.8±0.27	82.5±0.68
	Ours (w Co-Train Loco. Policy)	4.16±0.05	11.6±0.525	20.0±0.34	169.±13.83				
Lateral	RL (Pre-Train)	6.50±0.23	19.4±0.35	28.2±1.53	136.±1.44	–	–	–	–
	RL (Co-Train)	6.70±0.09	30.1±0.45	27.7±0.30	141.±4.60	9.28±0.02	25.2±0.29	40.2±0.18	168.±5.00
	Ours (w Pre-Train Loco. Policy)	3.74±0.32	14.4±2.73	17.0±1.19	70.7±6.83	4.97±0.01	18.1±1.61	25.2±0.14	88.0±4.18
	Ours (w Co-Train Loco. Policy)	4.05±0.05	12.4±0.49	21.9±0.39	156.±12.8				
Rotation	RL (Pre-Train)	6.47±0.47	23.2±9.03	29.4±0.16	154.±7.44	–	–	–	–
	RL (Co-Train)	6.38±0.14	24.0±1.73	27.0±0.26	135.±15.5	9.84±0.04	27.8±1.93	42.4±0.40	187.±8.38
	Ours (w Pre-Train Loco. Policy)	4.31±0.10	17.8±0.61	19.8±0.28	72.5±1.91	5.28±0.12	21.0±1.82	27.2±0.29	104.±0.97
	Ours (w Co-Train Loco. Policy)	4.15±0.02	12.32±0.31	22.8±0.10	143.±8.56				

–: The transfered upper-body policy has failed the locomotion policy due to excessive arm accleration.

TABLE III: Benchmarking results (MuJoCo) on robustness.

mance. This shows that simulated base accelerations with these components lead to effective compensation learning. Additionally, among the three ablation components, removing the operational space torque leads to the largest performance degradation, 36.11% for mean linear acceleration and 26.39% for mean angular acceleration, likely because this term provides a precise tracking signal that enables the RL policy to focus on learning only the compensation term, thereby improving overall performance.

Table III further demonstrates superior robustness over both the pre-train and co-train baselines when evaluated under a previously unseen locomotion policy. The pre-training method fails in all cases due to excessive arm movements, as the hierarchical training paradigm restricts the state-space exploration for the manipulation policy. In addition, the co-trained method exhibits an average degradation of 57.45% and 60.14% for mean linear and angular acceleration under an unseen locomotion policy, whereas ours shows an average degradation of 34.40% and 21.52%. This may be because the co-training setup relies heavily on coordinated interaction between the upper-body and lower-body for acceleration compensation.

B. Hardware results

1) End-effector acceleration comparison: In hardware demonstrations, we deploy our SEEC framework on the T1 robot. To evaluate the effectiveness of our approach, we compare our controller against the IK-based baseline on the real robot and compute the end-effector acceleration from pose data collected by a motion capture system operating at 120 Hz, as in Table IV. To obtain the end-effector acceleration, we apply numerical double differentiation to the recorded pose trajectories, removing abnormal or noisy measurements, as in Fig. 3. This provides a reliable measure of how the end-effector moves in a dynamically stable fashion.

We observe consistent results as in the simulation, where both linear and angular accelerations remain stabilized over time. Notably, while the absolute acceleration magnitudes are in a similar range, our method shows a smoother acceleration profile, underscoring that our framework is more stable.

	LinAcc (m/s^2) ↓		AngAcc (rad/s^2) ↓	
Method	mean	max	mean	max
IK-Based Method	3.57 ±0.46	11.6 ±4.63	41.1 ±4.31	151. ±14.6
SEEC (Ours)	**2.82** ±0.11	**6.36** ±0.36	**24.2** ±4.62	**78.6** ±9.81

TABLE IV: Real-world evaluation results on end-effector stability.

2) Solving loco-manipulation tasks: Furthermore, we test the proposed framework on tasks that require stable arm end-effector control under dynamic locomotion. The objective is to compensate for the hand acceleration when the robot is subject to whole-body motion and ground reaction forces. To this end, we design a set of representative tasks that combine locomotion with manipulation involving payloads.

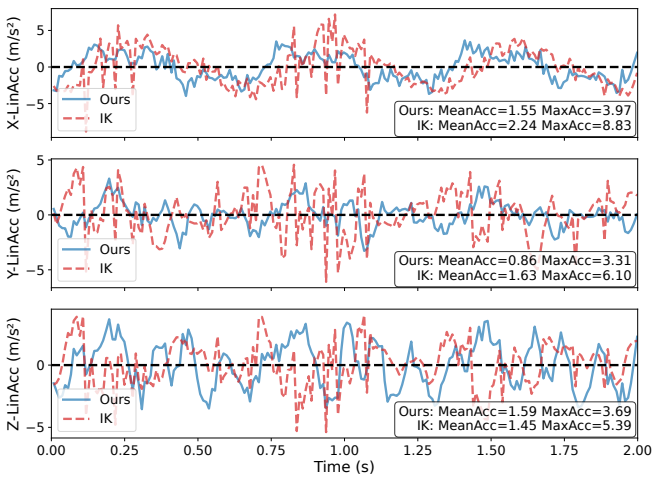
Chain Holding: The T1 robot needs to grasp a chain with its hands and attempts to minimize its oscillation through stable motion control during walking. This task requires minimizing damping oscillations introduced by locomotion and highlights the robot's ability to regulate dynamic external objects, which is shown in Fig. 1.

Mobile Whiteboard Wiping: The robot needs to hold an eraser with its gripper and wipe a vertical whiteboard while continuously stepping, with VR teleoperation [39]. The task requires the robot to maintain stable contact pressure and smooth wiping, ensuring effective cleaning while compensating for locomotion-induced disturbances.

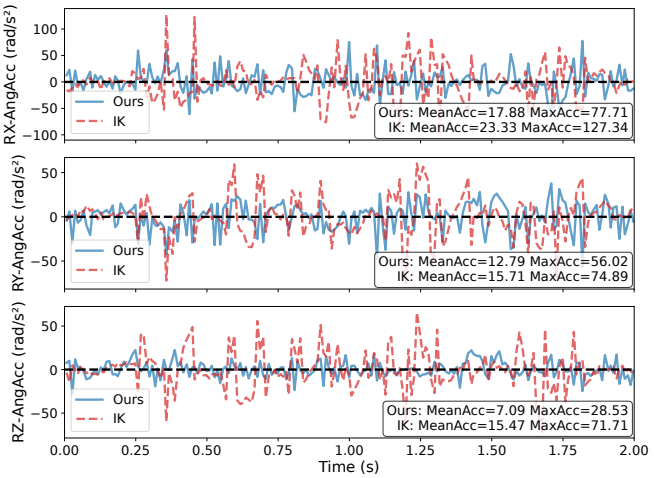
Plate Holding: The T1 carries a plate of snacks while walking. This task requires minimizing plate acceleration to prevent spilling and demands precise stabilization of the upper-body during movement.

Bottle Holding: The robot carries a bottle of liquid while walking. To prevent spilling, the robot must suppress oscillations and avoid sudden accelerations.

Results and Analysis. For the **Chain Holding** task, as in Fig. 1 (A-D), without acceleration compensation, the chain exhibits large-amplitude oscillations due to base motion, leading to a final dropping off from the robot hand. With our



(a) End-effector linear acceleration plots.



(b) End-effector angular acceleration plots.

Fig. 3: End-effector acceleration plots in real-world evaluation. The blue line indicates the acceleration profile of our method, and the dotted red line represents the baseline (IK) method.

SEEC framework, the robot effectively suppresses oscillatory dynamics, reducing oscillation amplitude and maintaining the chain nearly vertical during walking. This highlights the robustness of our framework in regulating external objects subject to dynamic excitations. For the **Mobile Whiteboard Wiping** task in Fig. 1(E), our framework consistently maintains smooth trajectories and steady end-effector contact forces, leading to clean wiping performance. For the **Plate Holding** task, as shown in Fig. 4, the baseline method spills the snacks as the robot walks. In contrast, our approach produces stable motions that allow the robot to carry the plate without spilling. For the **Bottle Holding** task, as shown in Fig. 5, the time-lapse shows the liquid shaking violently, leading to a sudden splash, while our method keeps the bottle steady with minimized sloshing.

These real-world tasks verify that our framework enables stable end-effector control under dynamic locomotion. The results highlight its robustness to disturbances and dynamic loco-manipulation scenarios.

Fig. 4: Plate holding task. With our method, the robot can stably hold the plate without spilling the snacks, whereas the IK-based method causes noticeable end-effector oscillations, leading to significant spillage.

V. Conclusions and Discussion

In this work, we introduce SEEC, a framework designed to achieve stable end-effector control for humanoid loco-manipulation. Our approach integrates model-based strategies into a learning-based end-effector controller, leveraging base acceleration data from simulation to enhance acceleration compensation. Experimental results from simulation demonstrate that our method consistently outperforms baseline approaches, yielding reduced end-effector acceleration and thereby improving stability.

Our method could benefit from integrating more advanced model-based controllers and RL training strategies. While we have verified the effectiveness of model-enhanced residual learning, our model can benefit from a model-based controller that can handle constraints and promote safe, stable operation, especially combined with constrained learning strategies.

Additionally, richer state inputs and more accurate state estimation would improve compensation and enable global target tracking. Our policy currently relies on upper-body proprioceptive input, which by design reacts to disturbances rather than proactively counteracting them. Whole-body state estimation could address this limitation and enable global target tracking to achieve more versatile loco-manipulation tasks, such as collaborative transports.

Acknowledgment

We sincerely thank Zhaoyuan Gu and Amelie Minji Kim for their thoughtful discussion and help with experiments. We would also like to express our gratitude for the hardware support provided by Booster Robotics.

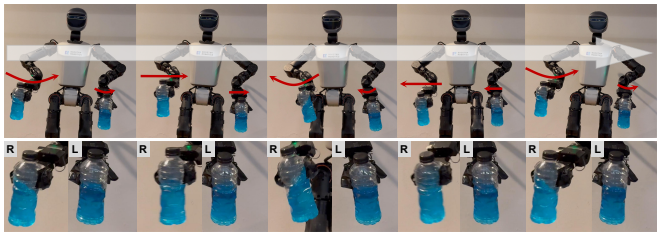


Fig. 5: Bottle holding task. The left arm is controlled by our approach, achieving stable holding with minimal liquid surface vibration, while the right arm is controlled by the IK baseline, resulting in pronounced liquid oscillations.

REFERENCES

[1] Z. Gu, J. Li, W. Shen, W. Yu, Z. Xie, S. McCrory, X. Cheng, A. Shamsah, R. Griffin, C. K. Liu *et al.*, "Humanoid locomotion and manipulation: Current progress and challenges in control, planning, and learning," *IEEEASME transactions on mechatronics*, 2025.

[2] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. M. Kitani, C. Liu, and G. Shi, "Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," in *Conference on Robot Learning*. PMLR, 2025, pp. 1516–1540.

[3] Z. Su, B. Zhang, N. Rahmanian, Y. Gao, Q. Liao, C. Regan, K. Sreenath, and S. S. Sastry, "Hitter: A humanoid table tennis robot via hierarchical planning and learning," *arXiv preprint arXiv:2508.21043*, 2025.

[4] Z. Fu, X. Cheng, and D. Pathak, "Deep whole-body control: Learning a unified policy for manipulation and locomotion," in *Conference on Robot Learning (CoRL)*, 2022.

[5] M. Liu, Z. Chen, X. Cheng, Y. Ji, R. Qiu, R. Yang, and X. Wang, "Visual whole-body control for legged loco-manipulation," *The 8th Conference on Robot Learning*, 2024.

[6] T. Portela, A. Cramariuc, M. Mittal, and M. Hutter, "Whole-body end-effector pose tracking," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 11 205–11 211.

[7] Y. Li, Y. Zhang, W. Xiao, C. Pan, H. Weng, G. He, T. He, and G. Shi, "Hold my beer: Learning gentle humanoid locomotion and end-effector stabilization control," *arXiv:2505.24198*, 2025.

[8] F. Liu, Z. Gu, Y. Cai, Z. Zhou, H. Jung, J. Jang, S. Zhao, S. Ha, Y. Chen, D. Xu *et al.*, "Opt2skill: Imitating dynamically-feasible whole-body trajectories for versatile humanoid loco-manipulation," *arXiv preprint arXiv:2409.20514*, 2024.

[9] 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.

[10] 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.

[11] 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," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 9772–9777, 2020.

[12] M. V. Minniti, F. Farshidian, R. Grandia, and M. Hutter, "Whole-body mpc for a dynamically stable mobile manipulator," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3687–3694, 2019.

[13] D. Wang, J. Yu, S. Wu, Z. Li, C. Li, R. Xiong, S. Qu, and Y. Wang, "A hierarchical mpc for end-effector tracking control of legged mobile manipulators," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 4855–4866, 2024.

[14] J. Woolfrey, W. Lu, and D. Liu, "Predictive end-effector control of manipulators on moving platforms under disturbance," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 2210–2217, 2021.

[15] Y. Ma, F. Farshidian, T. Miki, J. Lee, and M. Hutter, "Combining learning-based locomotion policy with model-based manipulation for legged mobile manipulators," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2377–2384, 2022.

[16] P. M. Wensing, M. Posa, Y. Hu, A. Escande, N. Mansard, and A. Del Prete, "Optimization-based control for dynamic legged robots," *IEEE Transactions on Robotics*, vol. 40, pp. 43–63, 2023.

[17] J.-P. Sleiman, F. Farshidian, and M. Hutter, "Versatile multicontact planning and control for legged loco-manipulation," *Science Robotics*, vol. 8, no. 81, p. eadg5014, 2023.

[18] S. Kuindersma, R. Deits, M. Fallon, A. Valenzuela, H. Dai, F. Permenter, T. Koolen, P. Marion, and R. Tedrake, "Optimization-based locomotion planning, estimation, and control design for the atlas humanoid robot," *Autonomous robots*, vol. 40, no. 3, pp. 429–455, 2016.

[19] C. D. Bellicoso, K. Krämer, M. Stäuble, D. Sako, F. Jenelten, M. Bjelonic, and M. Hutter, "Alma-articulated locomotion and manipulation for a torque-controllable robot," in *2019 International conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 8477–8483.

[20] 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.

[21] Z. Zhuang, S. Yao, and H. Zhao, "Humanoid parkour learning," in *8th Annual Conference on Robot Learning*, 2024.

[22] Z. Wang, J. Zhou, and Q. Wu, "Dribble master: Learning agile humanoid dribbling through legged locomotion," *arXiv preprint arXiv:2505.12679*, 2025.

[23] G. Pan, Q. Ben, Z. Yuan, G. Jiang, Y. Ji, S. Li, J. Pang, H. Liu, and H. Xu, "Roboduet: Learning a cooperative policy for whole-body legged loco-manipulation," *IEEE Robotics and Automation Letters*, 2025.

[24] D. Kang, J. Cheng, M. Zamora, F. Zargarbashi, and S. Coros, "Rl+ model-based control: Using on-demand optimal control to learn versatile legged locomotion," *IEEE Robotics and Automation Letters*, vol. 8, no. 10, pp. 6619–6626, 2023.

[25] D. Youm, H. Jung, H. Kim, J. Hwangbo, H.-W. Park, and S. Ha, "Imitating and finetuning model predictive control for robust and symmetric quadrupedal locomotion," *IEEE Robotics and Automation Letters*, vol. 8, no. 11, pp. 7799–7806, 2023.

[26] H. Jung, Z. Gu, Y. Zhao, H.-W. Park, and S. Ha, "Ppf: Pre-training and preservative fine-tuning of humanoid locomotion via model-assumption-based regularization," *IEEE Robotics and Automation Letters*, pp. 1–8, 2025.

[27] J. Cheng, D. Kang, G. Fadini, G. Shi, and S. Coros, "Rambo: Rl-augmented model-based whole-body control for loco-manipulation," *IEEE Robotics and Automation Letters*, 2025.

[28] S. H. Bang, C. A. Jové, and L. Sentis, "Rl-augmented mpc framework for agile and robust bipedal footstep locomotion planning and control," in *2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids)*. IEEE, 2024, pp. 607–614.

[29] T. Silver, K. Allen, J. Tenenbaum, and L. Kaelbling, "Residual policy learning," *arXiv preprint arXiv:1812.06298*, 2018.

[30] M. Mittal, C. Yu, Q. Yu, J. Liu, N. Rudin, D. Hoeller, J. L. Yuan, R. Singh, Y. Guo, H. Mazhar, A. Mandlekar, B. Babich, G. State, M. Hutter, and A. Garg, "Orbit: A unified simulation framework for interactive robot learning environments," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3740–3747, 2023.

[31] Y. Wang, P. Chen, X. Han, F. Wu, and M. Zhao, "Booster gym: An end-to-end reinforcement learning framework for humanoid robot locomotion," *arXiv preprint arXiv:2506.15132*, 2025.

[32] X. Gu, Y.-J. Wang, and J. Chen, "Humanoid-gym: Reinforcement learning for humanoid robot with zero-shot sim2real transfer," *arXiv preprint arXiv:2404.05695*, 2024.

[33] F. Wu, X. Nal, J. Jang, W. Zhu, Z. Gu, A. Wu, and Y. Zhao, "Learn to teach: Sample-efficient privileged learning for humanoid locomotion over real-world uneven terrain," *IEEE Robotics and Automation Letters*, 2025.

[34] E. R. Westervelt, J. W. Grizzle, and D. E. Koditschek, "Hybrid zero dynamics of planar biped walkers," *IEEE transactions on automatic control*, vol. 48, no. 1, pp. 42–56, 2003.

[35] B. Siciliano, L. Sciavicco, L. Villani, and G. Oriolo, *Robotics: modelling, planning and control*. Springer, 2009.

[36] O. Khatib, "A unified approach for motion and force control of robot manipulators: The operational space formulation," *IEEE Journal on Robotics and Automation*, vol. 3, no. 1, pp. 43–53, 2003.

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

[38] 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.

[39] Z. Zhao, L. Yu, K. Jing, and N. Yang, "Xrobotoolkit: A cross-platform framework for robot teleoperation," *arXiv preprint arXiv:2508.00097*, 2025.