

HULK: Learning Whole-Body Forceful Loco-Manipulation for Humanoids

An Dang^{*1,2}, Arturo Flores Alvarez^{*1,3}, Yu-Ming Chen¹, Conor Mc Gartoll¹, Helen Sun¹, Aaron Ames^{1,4}, Nima Fazeli^{1,2}, Manikantan Nambi¹

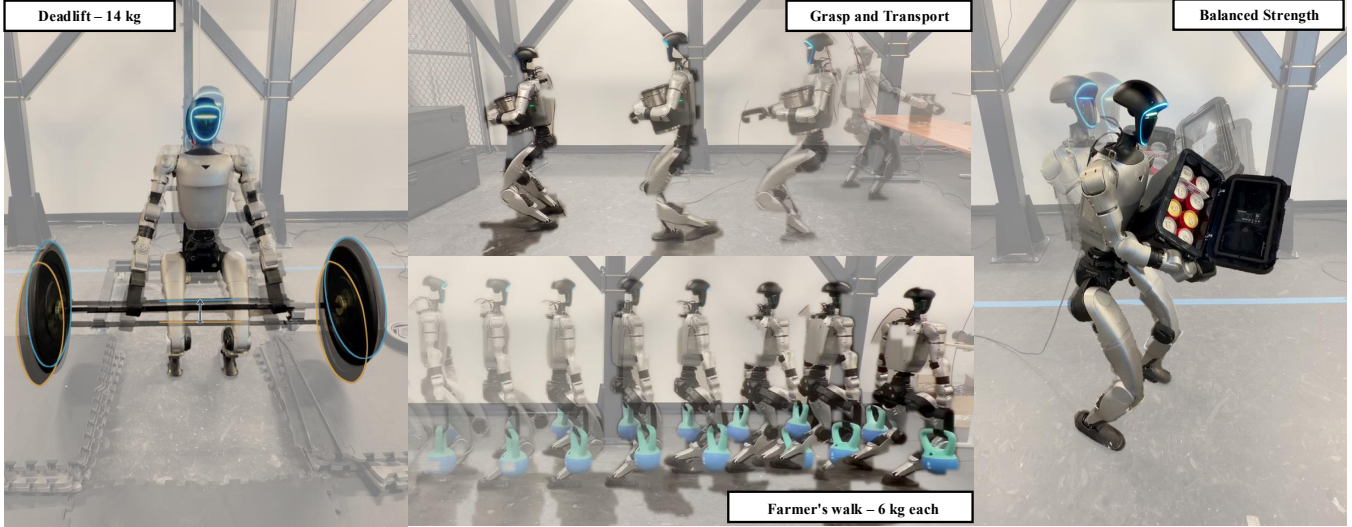


Fig. 1: **HULK enables loco-manipulation of heavy loads, anywhere on the upper body.** HULK first trains two specialists: one that tracks arm motions under forces applied at the wrists, and one that locomotes while holding large objects against the body. Next, it distills them into a single policy that drives one whole-body command interface. A predictive model of the loaded robot guides both specialists during training, so the policy anticipates how a payload alters the robot’s centroidal dynamics rather than discovering this relationship from reward alone. This pipeline enables the robot to perform on unseen tasks such as farmer’s carry and deadlift. Project website: www.hulk-forceful-wbc.com.

Abstract—Humanoid loco-manipulation of large, heavy objects demands forceful interaction across the entire body. However, such payloads shift a humanoid’s center of mass and impose sustained loads across the upper body, challenging balance and tracking. We present HULK, a whole-body control framework for forceful loco-manipulation. Using model predictive control (MPC) to guide reinforcement learning with predictions of the loaded dynamics, we train two teachers: one tracks arm motions under wrist forces, and the other locomotes while holding large objects against the body. A capture-point control barrier function augments the wrist-force teacher during training to improve balance under load. We distill both teachers into a single policy. We evaluate our method on a Unitree G1 in both simulation and the real-world. In simulation, the teacher with the barrier function achieves the lowest forward and lateral velocity tracking errors at 10 kg per arm among evaluated controllers and reduces aggregate balance metric by 35.7% relative to MPC-guided reinforcement learning alone. Our wrist-force teacher withstands torso push disturbances of up to 130 N.

I. INTRODUCTION

Humanoid robots are a compelling form factor for a world built around the human body. The promise is a single platform that can perform the same physical labor as people, in the same unmodified spaces. Realizing this promise, however, requires forceful whole-body work: lifting a loaded tote, leaning into a cart, or carrying a heavy object against the chest while walking. These tasks recur throughout logistics, manufacturing, and assistive applications. What these tasks share is a load the robot must sustain rather than reject: it persists for the task’s duration, shifts where the robot’s weight sits, and competes for actuation that keeps it balanced.

Current approaches to humanoid loco-manipulation have converged on a layered control strategy. At the bottom sits a learned whole-body controller (WBC): a policy that turns a compact set of task-space commands, such as base velocity, torso pose, and arm targets, into joint actions that keep the robot balanced. Above it, planners, teleoperation interfaces, and agentic pipelines issue those commands without reasoning about balance themselves, treating the WBC as a low-level “System 0” [1], [2]. Massively parallel reinforcement learning (RL) in simulation has made this bottom layer

^{*}Equal contribution. This work was done during an internship at Amazon.

¹Amazon, <andangt, amfloalv, yumingch, cmcgartoll, sunhele, aames, nfz, mnambi>@amazon.com

²University of Michigan, <andang, nfz>@umich.edu

³University of California, Los Angeles, aflores@ucla.edu

⁴California Institute of Technology, ames@caltech.edu

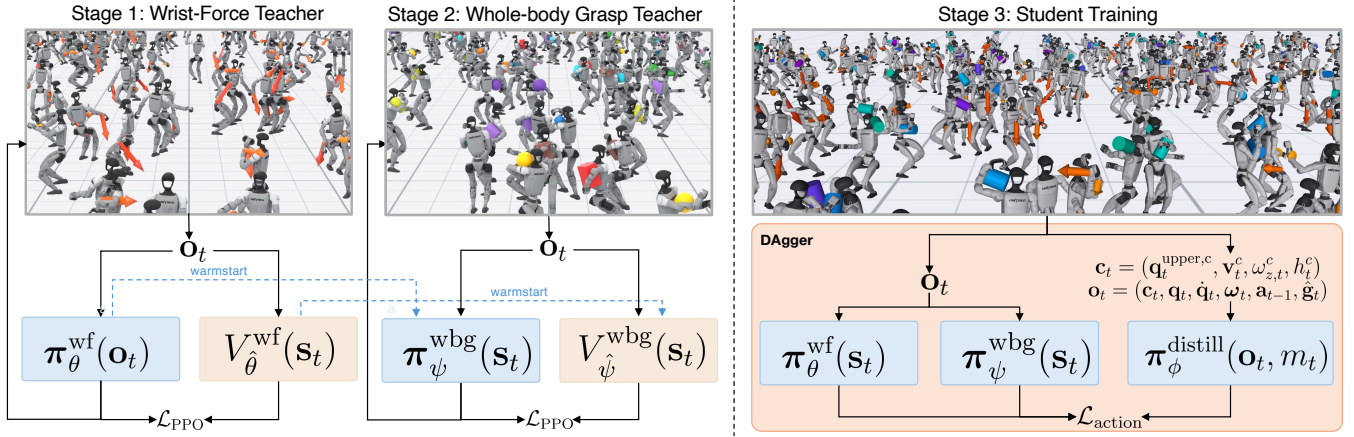


Fig. 2: **Training overview.** We decompose training into 3 stages. In stage 1, we train a teacher π_{θ}^{wf} that follows high-level commands c_t while undergoing force perturbations at the wrists. High-level commands include upper body joint targets $\mathbf{q}_t^{\text{upper},c}$, commanded linear velocity $\mathbf{v}_t^c$, commanded angular velocity in z direction ω_t^c , and height command h_t^c . In stage 2, the teacher π_{ψ}^{wbg} follows these commands while maintaining whole-body grasps across multiple object geometries. In stage 3, the student policy $\pi_{\phi}^{\text{distill}}$ is trained using DAGger with teacher mode conditioning $m_t \in \{0, 1\}$. The student takes in sensor observations $\mathbf{o}_t$ only present on a humanoid such as high-level commands, joint position $\mathbf{q}_t$, joint velocity $\dot{\mathbf{q}}_t$, angular velocity ω_t , previous action $\mathbf{a}_{t-1}$, and projected gravity $\hat{\mathbf{g}}_t$.

reliable enough to serve as that substrate [3], [4], [5], [6], [7], [8], and progress above it has been rapid along two fronts: *agility*, pushing the dynamism of whole-body skills [9], [10], [11], and *perception*, closing the loop on terrain and obstacles [12], [13].

Forceful interaction, however, remains comparatively underexplored. Unlike agility and perception, it requires the WBC to adapt its balance strategy to large and sustained external wrenches. A heavy payload changes the forces and moments that must be supported by the robot, with their effect varying continuously with the arm configuration and the object’s pose relative to the body. Existing approaches address parts of this problem. Compliance-based methods lower the effective stiffness of the robot so that it yields under unexpected contact [14]; while effective for incidental interaction, this behavior is poorly suited to supporting a heavy payload against gravity. The strongest forceful controllers instead learn to resist applied forces, but only *at the wrists* [15], or only for horizontal push–pull moments [16], [17]. Carrying a load further requires this adaptation to persist: the disturbance is not a transient to be rejected, but a condition that remains throughout the task. The robot must therefore maintain balance under a continuously applied wrench while simultaneously producing the forces required for manipulation.

We address the displaced center of mass and sustained balance under load in a single learned policy. Model predictive control (MPC) supplies reward guidance during RL [18], giving the policy a model-based signal for how a load moves the centroidal state rather than requiring it to discover that relationship from reward alone. We then augment the wrist-force teacher with a capture-point control barrier function (CBF) [19], [20]. Following CBF-RL [21], the barrier acts during training to improve learned balance recovery under

load while retaining network-only inference. Our contributions are:

- 1) We extend learned whole-body control to forceful interactions between external objects and contacts across the humanoid upper body.
- 2) We use MPC reward guidance to train robust locomotion and loco-manipulation under sustained payloads.
- 3) We augment the wrist-force teacher with training-time capture-point constraints, improving loaded tracking and balance recovery while retaining network-only inference.

II. RELATED WORKS

Whole-body Control. Whole-body controllers have lately been trained using reinforcement learning to follow high-level commands while actively stabilizing the humanoid. Several works [3], [4], [5], [6], [8] demonstrate the capabilities of sim-to-real reinforcement learning for motion tracking or teleoperation of the humanoid for human-like behavior on hardware. Large-scale models now expose whole-body control as a general-purpose interface: SONIC [7] supersedes motion tracking toward natural whole-body control using a large human motion tracking dataset. These methods address kinematic tracking rather than forces and persistent loading.

Forceful Whole-body Control. Two efforts define the forceful direction. FALCON [15] rejects force perturbations at the wrists with a wrist-force curriculum and a dual-network architecture, but struggles to hold commands while doing so. Thor [16] leans the torso to recruit the robot’s own weight, using ZMP analysis of sagittal-plane wrist forces, and so addresses horizontal push–pull rather than downward payload forces. Neither models whole-body object interaction during training, so the payload-augmented center of mass degrades locomotion under heavy load. HAFO [17] and

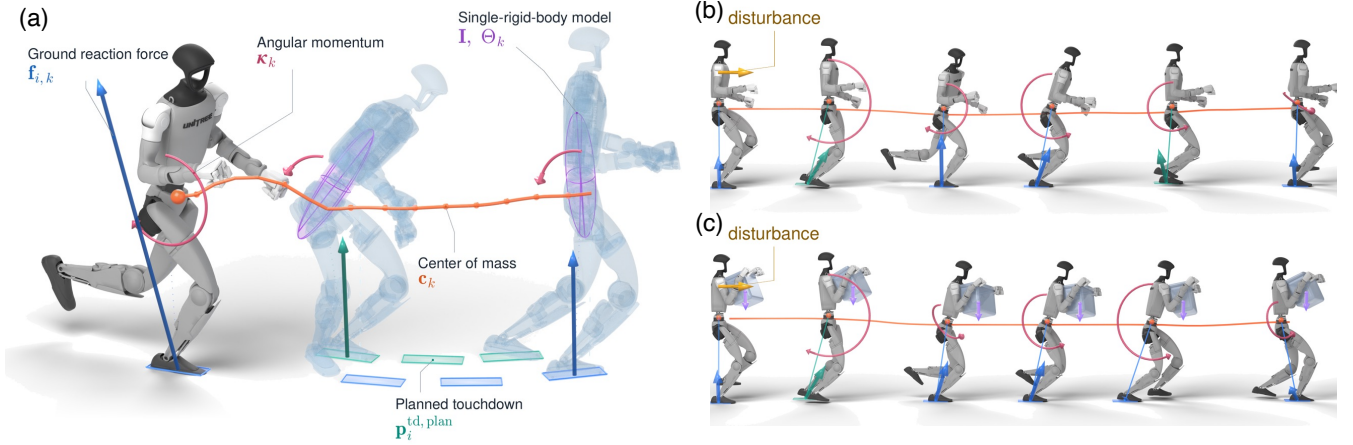


Fig. 3: **Centroidal MPC Force Reasoning.** In (a), we show the G1 about to fall over with forward angular momentum. The MPC generates a plan of the CoM trajectory and ground reaction forces to re-stabilize the angular momentum of the robot and keep an upright posture. The green annotations correspond to left foot while the blue ones correspond to right foot. The light blue ghosts illustrate the G1 following the MPC plan. (b) and (c) showcase the G1 stabilizing from an external pushes in a hands-free and minifridge grasp scenario.

load-aware adaptation [22] target force-adaptive control, but heavy payloads remain open, while Hold My Beer [23] and FACET [24] stabilize a payload under gentle motion rather than sustaining large loads.

Outside the learned-WBC setting, payload carrying has been addressed by hierarchical optimization [25] and by distilling trajectory optimization into skills [26], both indicating that a model of the loaded dynamics is valuable; Dao et al. [27] show sim-to-real box loco-manipulation without a whole-body command interface. AMO [28] provides such an interface but assumes both feet remain in contact.

Closest to our approach, Li et al. [18] use model predictive control for reward guidance and demonstrates cart-pushing and payload carrying. However, this method treats interactions as forces applied at the wrists rather than incorporating a payload wrench into reference generation, and it does not address contact against the torso. We extend it to a whole-body controller accepting height, torso-tilt and arm commands, with payloads entering the model through force perturbations and whole-body grasp force profiles.

Balance Constraints in Learned Control. Control barrier functions are ordinarily enforced by a runtime quadratic program that minimally corrects a nominal controller [19], [20], [29], which for a learned policy means shipping a reduced-order model, gradient and solver alongside the network. CBF-RL [21] instead applies the correction during training and penalizes the residual in the reward, so the requirement is internalized and deployment carries no extra machinery. We adopt that formulation. Because a sustained external wrench displaces the center of mass for a whole episode, the quantity to constrain is a balance state under load, for which the capture point model [30], is the natural choice.

III. METHODOLOGY

Our goal is to learn a single whole-body controller for payload-driven loco-manipulation: a policy that follows pla-

nar velocity, base height, torso orientation and arm joint commands while a load acts anywhere on the upper body, and that transfers to hardware from proprioception alone. We train two specialist teachers separately, one that tracks arm poses under forces applied at the wrists and one that locomotes while maintaining whole-body grasps across multiple objects. We distill the two into the single student we deploy (Sec. III-G, Fig. 2). The teachers both share a reward whose force-aware terms are supplied by a predictive model of the loaded robot rather than hand-designed (Sec. III-B), using a centroidal model as the predictor (Sec. III-C) together with the references that inform it of the load (Sec. III-D), and balance imposed by a barrier function during training instead of a runtime filter (Sec. III-F). The two differ only in the interaction each is exposed to and how it enters the model, which we make concrete once per teacher in Sec. III-E.

A. Problem Formulation

We formulate forceful whole-body control as an infinite-horizon discrete-time partially observable Markov decision process (POMDP) defined by the tuple $\mathcal{M} := (\mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \rho_0, \gamma)$ with state space $\mathcal{S}$, observation space $\mathcal{O}$, action space $\mathcal{A}$, reward $\mathcal{R}(s, a)$, transition probability $\mathcal{P}(s' | s, a)$, initial state distribution ρ_0 , and discount factor $\gamma \in [0, 1)$. The robot seeks a policy $\pi_\theta(a | o)$, parameterized by θ , that maximizes the expected discounted return $\mathcal{J} = \mathbb{E}_{\pi_\theta} [\sum_{t=0}^{\infty} \gamma^t r_t]$ over control steps t , where $s \in \mathcal{S}$, $o \in \mathcal{O}$, $a \in \mathcal{A}$, and $r_t = \mathcal{R}(s_t, a_t)$.

B. MPC Reward Guidance

We use MPC to generate outputs that will be tracked in the reward formulation of both teachers. These outputs include contact/interaction forces between the robot and its environment, the robot center of mass (CoM), and the rotation of the robot pelvis. We use MPC to solve the

following optimization problem over a horizon of N steps of duration Δt , given a dynamics model and constraints:

$$\begin{aligned} \min_{\mathcal{Z}} \quad & \sum_{k=0}^{N-1} L_k(\mathbf{x}_k, \mathbf{u}_k) + L_{\Delta,k}(\mathbf{u}_k, \mathbf{u}_{k-1}) + L_N(\mathbf{x}_N) \quad (1a) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = f_{\Delta t}(\mathbf{x}_k, \mathbf{u}_k), \quad k = 0, \dots, N-1, \quad (1b) \\ & \mathbf{u}_k \in \mathcal{U}_k, \quad k = 0, \dots, N-1, \quad (1c) \\ & \mathbf{x}_0 = \Phi(\mathbf{q}_t, \dot{\mathbf{q}}_t), \quad (1d) \end{aligned}$$

where $\mathcal{Z} = \{\mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{u}_0, \dots, \mathbf{u}_{N-1}\}$, $\mathbf{x}_k$ and $\mathbf{u}_k$ are the state and control input at horizon step k (both made concrete in Sec. III-C), $f_{\Delta t}$ is the dynamics model discretized at Δt , Φ maps the measured full-order configuration and velocity $(\mathbf{q}_t, \dot{\mathbf{q}}_t)$ to the reduced-order state, and $\mathcal{U}_k$ represents the control input set that each $\mathbf{u}_k$ is an element of. We define our stage and terminal costs as quadratic costs defined as:

$$\begin{aligned} L_k &= \|\mathbf{x}_k - \mathbf{x}_k^{\text{ref}}\|_{\mathbf{Q}}^2 + \|\mathbf{u}_k - \mathbf{u}_k^{\text{ref}}\|_{\mathbf{W}}^2, \quad (2) \\ L_{\Delta,k} &= \|\mathbf{u}_k - \mathbf{u}_{k-1}\|_{\mathbf{W}_{\Delta}}^2 \quad (3) \\ L_N &= \|\mathbf{x}_N - \mathbf{x}_N^{\text{ref}}\|_{\mathbf{Q}_N}^2, \quad (4) \end{aligned}$$

where $\|\mathbf{v}\|_{\mathbf{A}}^2 = \mathbf{v}^T \mathbf{A} \mathbf{v}$ for any vector $\mathbf{v}$ and weight matrix $\mathbf{A}$, $\mathbf{x}_k^{\text{ref}}$ is the reference trajectory generated from the high-level commands, $\mathbf{u}_k^{\text{ref}}$ is an input reference that is zero unless a known external load is being planned for (Sec. III-E), so that the second term is plain input regularization by default, $\mathbf{u}_{-1}$ is the input applied at the previous solve, and $\mathbf{Q}$, $\mathbf{W}$, and $\mathbf{Q}_N$ represent tunable diagonal positive semi-definite matrices for the MPC optimization problem.

Solving (1) yields an optimal state trajectory $\{\tilde{\mathbf{x}}_k\}_{k=0}^N$ and input trajectory $\{\tilde{\mathbf{u}}_k\}_{k=0}^{N-1}$. Rather than executing the plan, we use it as reward guidance for RL [18]: the policy is rewarded for keeping the measured robot state close to the plan,

$$r_t^{\text{mpc}} = \sum_{i \in \mathcal{T}} \exp(-\alpha_i \|\Lambda_i(\tilde{\mathbf{x}}_k(t), \tilde{\mathbf{u}}_k(t)) - \Lambda_i(\mathbf{x}_t, \mathbf{u}_t)\|^2), \quad (5)$$

where $\mathcal{T}$ indexes the tracked quantities, each Λ_i extracts one of them from the plan at the horizon step $k(t)$ corresponding to time t and from the measured reduced-order state $\mathbf{x}_t$ and measured interaction forces $\mathbf{u}_t$, and $\alpha_i > 0$ sets the sensitivity of the i -th term; the tracked quantities themselves follow from the choice of model (Sec. III-C). This gives the policy a signal for how the payload displaces the reduced-order state, rather than requiring it to discover the loaded dynamics from task rewards alone.

C. Centroidal Dynamics

For reward guidance the plan need only be dynamically consistent rather than full-order, so we adopt the reduced-order model of [18] and extend its interaction set to the upper body. The state of (1) is the centroidal state $\mathbf{x}_k = (\mathbf{c}_k, \boldsymbol{\Theta}_k, \mathbf{l}_k, \boldsymbol{\kappa}_k) \in \mathbb{R}^{12}$, collecting the CoM position, the base roll, pitch and yaw, and the centroidal linear and angular momenta, respectively. The input $\mathbf{u}_k = (\mathbf{f}_{i,k}, \boldsymbol{\tau}_{i,k})_{i \in \mathcal{C}}$ stacks the wrenches at the interaction points $\mathcal{C}$, the two feet and, when they carry load, the two wrists; a flat foot contributes a six-dimensional wrench and a wrist a three-dimensional

force, so $\mathbf{u}_k \in \mathbb{R}^{6|\mathcal{C}|}$. With m the total mass, $\mathbf{g}$ gravity, $\mathbf{I}$ the composite rigid-body inertia held constant over the horizon, $\sigma_{i,k} \in \{0, 1\}$ gating interaction i into the active set $\mathcal{C}_k$, and $\bar{\mathbf{c}}_k$ a reference CoM, an Euler discretization of the centroidal dynamics [31] gives

$$\begin{aligned} \mathbf{c}_{k+1} &= \mathbf{c}_k + \frac{\Delta t}{m} \mathbf{l}_k + \frac{\Delta t^2}{2} \left(\mathbf{g} + \frac{1}{m} \sum_{i \in \mathcal{C}} \sigma_{i,k} \mathbf{f}_{i,k} \right), \\ \boldsymbol{\Theta}_{k+1} &= \boldsymbol{\Theta}_k + \Delta t \mathbf{I}^{-1} \boldsymbol{\kappa}_k, \\ \mathbf{l}_{k+1} &= \mathbf{l}_k + \Delta t \left(m \mathbf{g} + \sum_{i \in \mathcal{C}} \sigma_{i,k} \mathbf{f}_{i,k} \right), \\ \boldsymbol{\kappa}_{k+1} &= \boldsymbol{\kappa}_k + \Delta t \sum_{i \in \mathcal{C}} \sigma_{i,k} [(\mathbf{p}_{i,k} - \bar{\mathbf{c}}_k) \times \mathbf{f}_{i,k} + \boldsymbol{\tau}_{i,k}], \end{aligned} \quad (6)$$

which is affine in $(\mathbf{x}_k, \mathbf{u}_k)$, hence a linear time-varying model $\mathbf{x}_{k+1} = \mathbf{A}_k \mathbf{x}_k + \mathbf{B}_k \mathbf{u}_k + \mathbf{e}_k$. A carried payload is not a decision variable: the force and moment it exerts about the CoM enter $\mathbf{e}_k$ in the same rows as an interaction wrench, so the plan leans into the load rather than rejecting it as a disturbance.

The input set $\mathcal{U}_k$ of (1c) collects the wrenches an active interaction can source. A contact force obeys Coulomb friction, $\|\mathbf{f}_i^t\|_2 \leq \mu_i f_i^n$ with $f_i^n \in [0, f_i^{n,\max}]$ for tangential and normal components $\mathbf{f}_i^t, f_i^n$ and friction coefficient μ_i ; bounding the two tangential components individually replaces the second-order cone by halfspaces, keeping the problem a quadratic program. Every row scales with f_i^n , so a gated swing foot can source no force.

D. Force-Aware References

The centroidal MPC prescribes footstep locations and gait timings to preserve convexity. Prior work, however, does not account for sustained external forces when generating these footstep references. The footstep planner used in Li et al. [18] uses a feedforward Raibert heuristic [32] that places the swing foot at the neutral point of the commanded stride,

$$\mathbf{p}_i^{\text{td}} = \mathbf{c}_{xy}^{\text{td}} + \frac{T_{\text{st}}}{2} \mathbf{v}_{xy}^{\text{cmd}} + \mathbf{R}_z(\psi) \mathbf{d}_i, \quad (7)$$

where $\mathbf{p}_i^{\text{td}} \in \mathbb{R}^2$ is the touchdown location of foot i , the subscript xy denotes planar components, T_{st} is the stance duration, $\mathbf{v}_{xy}^{\text{cmd}}$ the commanded planar velocity, $\mathbf{R}_z(\psi)$ the rotation by the base yaw ψ , $\mathbf{d}_i$ the lateral offset of foot i from the CoM in the base frame, and $\mathbf{c}_{xy}^{\text{td}}$ the CoM position predicted at touchdown by integrating the commanded twist forward from the measured $\mathbf{c}_{xy}$. Equation (7) is purely feedforward. Under a sustained external force the realized velocity deviates from $\mathbf{v}_{xy}^{\text{cmd}}$.

We make the Raibert heuristic force-aware by adding an offset for external force perturbations. Let $\mathcal{H}$ be the set of contacts on the robot's body, where contact j applies a force $\mathbf{f}_{j,xy}^e$ at height z_j above the ground. Balancing moments about the CoM, a foot displacement d contributes mgd while a force $\mathbf{f}_{j,xy}^e$ applied at height z_j contributes $z_j \mathbf{f}_{j,xy}^e$. Equating the two, the foot must land further out by $\sum_{j \in \mathcal{H}} z_j \mathbf{f}_{j,xy}^e / (mg)$. Adding this offset to Eq. (7) and scaling it by a gain k_f gives

$$\mathbf{p}_i^{\text{tdf}} = \mathbf{p}_i^{\text{td}} + \frac{k_f}{mg} \sum_{j \in \mathcal{H}} z_j \mathbf{f}_{j,xy}^e, \quad (8)$$

We also add the Raibert feedback term [32] where k_c is the gain controlling feedback law which results in $\mathbf{p}_i^{\text{tdf-fb}} = \mathbf{p}_i^{\text{tdf}} + \frac{k_c}{\omega_0} (\dot{\mathbf{c}}_{xy} - \mathbf{v}_{xy}^{\text{cmd}})$.

E. Teacher Training

Wrist-force teacher. We train a controller that braces and resists against forces applied at the wrists while tracking an upper-body pose command, using a force curriculum akin to [15]. The arm command is drawn from the BONES-RP1 dataset [33] retargeted to the 29-DoF robot: we keep the 14 arm joint angles of every retargeted frame, roughly 4.8×10^4 frames over 252 motions, and sample individual frames as static pose targets rather than replaying a trajectory, so the teacher has to hold arbitrary arm configurations under load instead of one scripted motion. Independent 3D forces are applied at the two wrists, up to ± 40 N horizontally and -50 to $+5$ N vertically, each axis zeroed with probability 0.25, held for 3 to 5 s, and ramped by a per-environment magnitude curriculum. The wrists enter the model (6) as two point interactions alongside the feet, placed by forward kinematics along the arm chain evaluated at the *commanded* arm target and gated on only while a force is live, so $\mathbf{u}_k \in \mathbb{R}^{18}$ collects two foot wrenches and two wrist forces.

Whole-body-grasp teacher. The teacher learns to locomote and stay stable while maintaining whole-body grasps across multiple objects. The interactions are the ground reaction forces and the contact forces of each environment’s assigned grasp. We generate 130,000 grasps offline across multiple object geometries using energy-based optimization from [34] (examples in Fig. 6). Each environment is initialized with an object and a compatible grasp to maintain, so the contact locations of (6) are read from that grasp and gated on for as long as the assigned object is held. The grasp contact forces are decision variables subject to the same friction constraints as the ground-reaction forces. The forces exerted by the object in the sampled whole-body grasp provide $\mathbf{u}_k^{\text{ref}}$, while the MPC optimizes the corresponding interaction forces. These forces also contribute to the centroidal angular-momentum dynamics in Eq. (6). All else in Secs. III-B to III-D is shared between the two teachers. During training, we warm-start from the wrist-force teacher so that locomotion and upper-body command tracking are not relearned from scratch.

F. Capture-Point Balance Guidance via CBF-RL

Forceful loco-manipulation demands sustained balance while the robot keeps tracking motion and force commands. We introduce a capture-point control barrier function (CP-CBF) and incorporate it into the wrist-force teacher through CBF-RL [21]. This corrects the wrist-force teacher’s commands with a control barrier function during training and penalizes the residual in the reward, so the exported policy runs unfiltered (Fig. 4). We defer to [21] for the training mechanics.

The barrier uses the instantaneous capture point [30], equivalently the planar divergent component of motion (DCM) under the constant-height LIPM, defined as $\xi =$

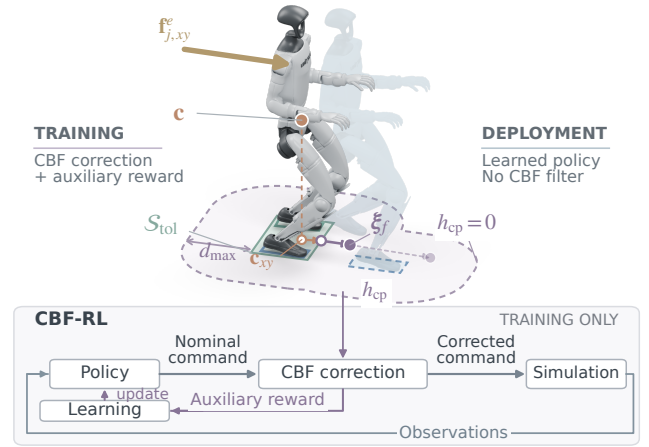


Fig. 4: **CP-CBF guidance.** The ghost illustrates a recovery step into a separate admissible footprint; the dashed boundary marks $h_{cp} = 0$. The inset shows command correction and reward feedback during training only.

$\mathbf{c}_{xy} + \dot{\mathbf{c}}_{xy}/\omega_0$, the support location that would arrest horizontal motion under a linear inverted pendulum. Here $\mathbf{c}_{xy}$ and $\dot{\mathbf{c}}_{xy}$ are the horizontal CoM position and velocity, and $\omega_0 = \sqrt{g/z_{\text{cmd}}}$ is the natural frequency with z_{cmd} the CoM height from the commanded base height. We make the barrier force-aware as $\xi_f = \xi + \Delta \mathbf{p}_{\text{CoP}}$, where $\Delta \mathbf{p}_{\text{CoP}}$ is the center-of-pressure shift that balances the external wrench. We then measure ξ_f against a support region $\mathcal{S}_{\text{tol}}$, a yaw-aligned union of tolerance-expanded boxes over the contacting soles and predicted touchdown footprints that admits recovery steps without covering the intervening region, and define

$$h_{cp} = 1 - \text{dist}(\xi_f, \mathcal{S}_{\text{tol}})/d_{\text{max}}, \quad (9)$$

where dist is the minimum distance to the region, zero inside it, and $d_{\text{max}} > 0$ is the admissible excursion, so that $h_{cp} > 0$ is remaining margin, $h_{cp} = 0$ its limit, and negative values an exceeded allowance. The admissible excursion we denote as the divergent component of motion (DCM) excursion. This will also be the balance metric that we use in our experiments. The CBF-QP and reward guidance are detailed in Sec. B.

G. Student Distillation

The teachers specialize in different interaction regimes. We freeze both and train a mode-conditioned student with DAgger [35]. At each student-visited state, the teacher selected by $m_t \in \{0, 1\}$ supplies the target action. The student regresses these actions using the same observation space $\mathcal{O}$ as the teacher actors, keeping supervision on its own rollout distribution.

IV. EXPERIMENTS

We evaluate whether HULK can maintain commanded motion under sustained load, recover from disturbances, and transfer these capabilities to forceful loco-manipulation tasks.

A. Experimental Setup

Controllers and comparisons. We compare against FALCON, Thor, and AMO [15], [16], [28]. The simulation rows *Ours* and *Ours + CP-CBF* evaluate the wrist-force teachers; hardware trials evaluate the distilled student deployed zero-shot on the Unitree G1.

In simulation, all controllers receive identical commands and perturbation schedules on the same platform. Commands use their shared interface: $v_x, v_y \in [-0.5, 0.5]$ m/s, $\omega_z \in [-0.5, 0.5]$ rad/s, base height 0.5–0.8 m, torso roll ± 0.2 rad, torso pitch ± 0.25 rad, and eight common arm joints. We report mean $\pm$ sample standard deviation across five evaluation-seed aggregates for each fixed policy. These seeds vary evaluation inputs, not policy training. Bold marks the best displayed value in each simulation-table column.

Carrying and push recovery. We apply 10 kg per arm at mid-forearm, ramp the load over 2 s, and allow another 2 s to settle before issuing commands. Each trial commands one velocity axis, measuring tracking under sustained load after the loading transient. Separately, standing push tests apply a one-second torso force at six magnitudes from 50 to 130 N, from behind, in front, and laterally. Each HULK teacher is evaluated on 90 push episodes. We measure the signed margin of the DCM defined in Section III-F relative to the benchmark’s two-foot support envelope. Negative margins indicate that the DCM lies outside this envelope; we call these departures DCM excursions. We report the average of each episode’s minimum post-push margin, with less negative values indicating smaller worst departures outside the envelope. Peak torso inclination is averaged over the three directions at 50 N, the largest push all controllers survive. Maximum push denotes the largest tested magnitude survived in every direction.

Wrist-force rejection. While the robot follows arm and velocity commands, we apply randomized 3D wrist forces with scales $\alpha_g \in \{0, 0.5, 1.0\}$, corresponding to 0, 20, and 40 N per wrist. The planar component opposes travel in the base frame. Upper-body error E_{upper} is the RMS tracking error over the common arm joints. Composite root error E_{root} is the time average of the numerical sum of horizontal velocity-error magnitude (m/s) and absolute yaw-rate error (rad/s); it is not a single-unit physical quantity.

Hardware tasks. Cart push/pull and payload carry load the wrists, while fridge and canister carry additionally introduce contact against the torso (Fig. 5). All tasks involve walking. For each task, controllers receive identical commands and object placement at a common load. A trial succeeds if the robot completes the commanded task without falling, triggering a joint-torque protective stop, or dropping the object. For payload carry, where all controllers can complete the carrying task, we additionally require the robot to remain within a prescribed path-deviation threshold, measured using an OptiTrack motion capture system. Trial counts vary by task and appear as the denominators in Tab. III.

For each controller and task, we additionally sweep the

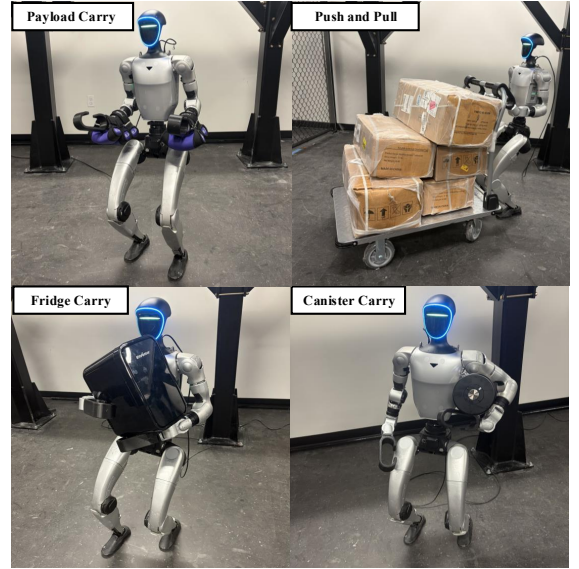


Fig. 5: **Real-world tasks.** Cart push/pull, payload carry, fridge carry, and canister carry on the Unitree G1. Together they span the upper body: force acts at the wrists in the first two and against the torso in the last two, and every task is performed while walking.

payload magnitude to determine the maximum load completed successfully. Carrying tasks use forward locomotion, whereas cart push/pull uses backward locomotion. Success rates are evaluated separately at a common fixed load for each task: 150 kg for cart push/pull, 11 kg for payload carry, 6 kg for fridge carry, and 5 kg for canister carry.

B. Simulation Results

Sustained load and recovery quality. The CP-CBF teacher combines lower loaded translational errors with smaller DCM excursions (Tab. I). Relative to *Ours*, the largest tracking improvement is lateral, suggesting that the benefit depends on the commanded direction. Because the load acts at the forearms rather than the training wrists, these results also show transfer across force application points. AMO retains the lowest mean turn-rate error, so the advantage does not extend uniformly across commands. The two HULK teachers achieve similar push-survival performance, but differ substantially in DCM excursion. The smaller DCM excursions of the CP-CBF teacher suggest improved recovery quality that push survival alone may hide. The smaller DCM excursions from our CP-CBF teacher suggest a difference in recovery quality that push survival may hide. This is consistent with the balance state targeted during training, although negative margins show that recovery involves leaving the support envelope.

Arm accuracy and locomotion under wrist forces. Both HULK teachers perform competitively against FALCON in arm and composite root tracking at the strongest tested wrist force (Tab. II) with only a mean difference of 0.02 rad. Their arm errors remain nearly constant as force increases, although FALCON retains the lowest means. CP-CBF has

Method	Tracking at 10 kg/arm ↓			Disturbance rejection		
	E_{v_x} (m/s)	E_{v_y} (m/s)	E_{ω_z} (rad/s)	Worst DCM margin (m) ↑	Peak tilt (deg) ↓	Max push (N) ↑
Thor [16]	0.263 ± 0.002	0.265 ± 0.003	0.450 ± 0.006	-0.275 ± 0.016	23.17 ± 0.13	80
AMO [28]	0.276 ± 0.002	0.350 ± 0.004	0.196 ± 0.003	-0.705 ± 0.030	9.94 ± 0.74	50
FALCON [15]	0.367 ± 0.012	0.369 ± 0.013	0.392 ± 0.005	-0.329 ± 0.034	7.91 ± 0.23	80
Ours	0.242 ± 0.009	0.257 ± 0.004	0.234 ± 0.008	-0.251 ± 0.011	5.90 ± 0.18	110
Ours + CP-CBF	0.210 ± 0.007	0.142 ± 0.003	0.197 ± 0.009	-0.162 ± 0.012	5.59 ± 0.24	130

TABLE I: **Carrying and disturbance rejection.** Tracking under 10 kg/arm at mid-forearm and, separately, standing torso pushes. Both Ours rows evaluate wrist-force teachers. DCM margin entries average episode-wise minimum signed margins; negative values indicate excursions outside the benchmark support envelope. Tilt is measured at 50 N. Protocols, aggregation, and boldface are defined in Sec. IV-A.

Method	E_{upper} (rad) ↓			E_{root} (composite) ↓		
	$\alpha_g = 0$	0.5	1.0	$\alpha_g = 0$	0.5	1.0
Thor [16]	0.337 ± 0.282	0.329 ± 0.282	0.337 ± 0.275	0.762 ± 0.264	0.849 ± 0.250	0.913 ± 0.185
AMO [28]	0.402 ± 0.207	0.416 ± 0.212	0.445 ± 0.222	0.680 ± 0.016	1.228 ± 0.064	2.111 ± 0.104
FALCON [15]	0.144 ± 0.072	0.148 ± 0.074	0.159 ± 0.079	0.444 ± 0.041	0.525 ± 0.050	0.868 ± 0.134
Ours	0.161 ± 0.052	0.160 ± 0.056	0.162 ± 0.059	0.620 ± 0.062	0.543 ± 0.044	0.839 ± 0.030
Ours + CP-CBF	0.172 ± 0.055	0.172 ± 0.059	0.176 ± 0.061	0.620 ± 0.012	0.752 ± 0.019	0.948 ± 0.033

TABLE II: **Force rejection while walking.** Arm and composite root tracking errors; both Ours rows evaluate wrist-force teachers. Force scales $\alpha_g \in \{0, 0.5, 1.0\}$ correspond to 0, 20, and 40 N per wrist. Metrics and reporting follow Sec. IV-A.

higher root error than our baselines at both nonzero force levels. Our teachers train on sampled poses held under load, with MPC and CP-CBF guidance emphasizing balance, while FALCON uses a torque-aware force curriculum [15].

C. Real-world Results

HULK achieves the highest observed completion rate and maximum completed load on three of the four hardware tasks, and ties AMO on payload carry (Tab. III). In cart push/pull, HULK completes 5/6 trials at the common load versus 3/6 for the best baseline, FALCON, and pulls up to 300 kg compared with 200 kg in the separate load sweep. In payload carry, all controllers can physically carry the load, but only HULK and AMO consistently remain within the prescribed path-deviation threshold, both achieving 4/4 successful trials at the common load. HULK and AMO also both reach 12.25 kg in the separate load sweep.

The largest completion-rate gains arise when objects are supported against the torso. HULK completes 4/4 trials with both the fridge in front of the torso and the canister at its side. AMO completes 2/4 fridge trials, whereas Thor and FALCON complete none; all baselines complete 0/4 canister trials. HULK’s maximum completed loads are 7.0 kg for the fridge and 6.1 kg for the canister, compared with baseline maxima of 6.0 and 2.6 kg, respectively. Together, these results show that the distilled policy combines cart and payload capabilities with successful whole-body carrying across the two tested load geometries.

V. DISCUSSION AND LIMITATIONS

Load geometry helps explain why forceful loco-manipulation requires whole-body coordination. A force’s

point of application determines its moment about the CoM, coupling arm configuration to the support needed to sustain the load (Eq. (6)). Force-aware MPC exposes this coupling in the training objective. Transfer from wrist forces to forearm loading suggests that learned compensation can extend beyond the trained contact location, although it does not establish generalization to arbitrary contact arrangements.

MPC and CP-CBF provide complementary guidance: MPC supplies references consistent with loaded dynamics, while the barrier shapes how the policy approaches a balance boundary. The observed tradeoff between root tracking and recovery (Tabs. I and II) suggests that carrying may benefit from allowing motion errors to preserve balance. The exported policy is unfiltered, so the barrier provides no runtime safety guarantee.

The model and training regime limit the behaviors this guidance can support. The centroidal predictor approximates articulated dynamics without explicitly resolving joint-level actuation feasibility, so a balanced centroidal plan need not be executable by every joint. The wrist-force teacher samples static arm poses, and the grasp teacher starts from established grasps generated across multiple objects. Coverage of stable configurations does not establish competence during transitions between them. Grasp acquisition, regrasping, and rapid contact changes therefore remain outside the demonstrated scope. Richer dynamics models and training on contact transitions are directions for extending to dynamic manipulation.

ACKNOWLEDGMENT

We acknowledge the use of AI-generated content (including but not limited to text, figures, images, and code) in this paper.

Method	Cart push/pull		Payload carry		Fridge carry		Canister carry	
	SR. $\uparrow$	Max $\uparrow$	SR. $\uparrow$	Max $\uparrow$	SR. $\uparrow$	Max $\uparrow$	SR. $\uparrow$	Max $\uparrow$
Thor [16]	2/6	130	2/4	7.25	0/4	2.0	0/4	0.6
AMO [28]	0/6	75	4/4	12.25	2/4	6.0	0/4	2.6
FALCON [15]	3/6	200	0/4	3.62	0/4	2.5	0/4	0.0
HULK	5/6	300	4/4	12.25	4/4	7.0	4/4	6.1

TABLE III: **Real-world evaluation on the Unitree G1.** SR. reports successful trials/total trials at a common load per task. Max reports the largest load completed in a separate sweep in kg. The HULK row evaluates the distilled student; see Sec. IV-A for the protocol.

REFERENCES

- [1] L. Yang, J. Li, N. Poddar, Y. Hou, G. Huh, R. Griffin, G. Gkioxari, and A. Ames, "HANDOFF: Humanoid agentic task-space whole-body control via distilled complementary teachers," *arXiv preprint arXiv:2606.06493*, 2026.
- [2] S. Wei, H. Jing, B. Li, Z. Zhao, J. Mao, Z. Ni, S. He, J. Liu, X. Liu, K. Kang, S. Zang, W. Yuan, M. Pavone, D. Huang, and Y. Wang, " Ψ_0 : An open foundation model towards universal humanoid locomanipulation," in *Robotics: Science and Systems (RSS)*, 2026.
- [3] 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," in *Conference on Robot Learning (CoRL)*, 2024.
- [4] T. He, W. Xiao, T. Lin, Z. Luo, Z. Xu, Z. Jiang, J. Kautz, C. Liu, G. Shi, X. Wang, L. Fan, and Y. Zhu, "HOVER: Versatile neural whole-body controller for humanoid robots," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.
- [5] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, "Expressive whole-body control for humanoid robots," in *Robotics: Science and Systems (RSS)*, 2024.
- [6] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "HumanPlus: Humanoid shadowing and imitation from humans," in *Conference on Robot Learning (CoRL)*, 2024.
- [7] Z. Luo, Y. Yuan, T. Wang, C. Li, F. Castañeda, S. Chen, Z.-A. Cao, J. Li, D. Minor, Q. Ben, J. Park, D. Sami, Z. Wang, X. Da, R. Ding, C. Hogg, L. Song, E. Lim, E. Jeong, T. He, H. Xue, W. Xiao, S. Yuen, J. Kautz, Y. Chang, U. Iqbal, L. Fan, and Y. Zhu, "SONIC: Supersizing motion tracking for natural humanoid whole-body control," *arXiv preprint arXiv:2511.07820*, 2025.
- [8] Q. Liao, T. E. Truong, X. Huang, Y. Gao, G. Tevet, K. Sreenath, and C. K. Liu, "BeyondMimic: From motion tracking to versatile humanoid control via guided diffusion," *arXiv preprint arXiv:2508.08241*, 2025.
- [9] 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. Fan, Y. Zhu, C. Liu, and G. Shi, "ASAP: Aligning simulation and real-world physics for learning agile humanoid whole-body skills," in *Robotics: Science and Systems (RSS)*, 2025.
- [10] D. Wang, X. Wang, C. Zhang, J. Shi, Y. Zhao, C. Bai, and X. Li, "X-LoCo: Towards generalist humanoid locomotion control via synergetic policy distillation," in *Robotics: Science and Systems (RSS)*, 2026.
- [11] Y. Wang, S. Zhu, P. Zhi, Y. Li, J. Li, Y.-L. Li, Y. Xiao, X. Wang, B. Jia, and S. Huang, "OmniXtreme: Breaking the generality barrier in high-dynamic humanoid control," in *Robotics: Science and Systems (RSS)*, 2026.
- [12] Y. Zhang, Y. Seo, J. Chen, Y. Yuan, K. Sreenath, P. Abbeel, C. Sferazza, K. Liu, R. Duan, and G. Shi, "RPL: Learning robust humanoid perceptive locomotion on challenging terrains," *arXiv preprint arXiv:2602.03002*, 2026.
- [13] Z. Wu, X. Huang, L. Yang, Y. Zhang, X. Chen, P. Abbeel, R. Duan, A. Kanazawa, C. Sferazza, G. Shi, and C. K. Liu, "Perceptive humanoid parkour: Chaining dynamic human skills via motion matching," *arXiv preprint arXiv:2602.15827*, 2026.
- [14] G. B. Margolis, M. Wang, N. Fey, and P. Agrawal, "SoftMimic: Learning compliant whole-body control from examples," *arXiv preprint arXiv:2510.17792*, 2025.
- [15] Y. Zhang, Y. Yuan, P. Gurunath, I. Gupta, S. Omidshafiei, A. Akbar Agha-mohammadi, M. Vazquez-Chanlatte, L. Pedersen, T. He, and G. Shi, "FALCON: Learning force-adaptive humanoid locomanipulation," in *Proceedings of the 8th Annual Learning for Dynamics and Control Conference (LADC)*, 2026.
- [16] G. Li, Q. Shi, Y. Hu, Z. Wang, X. Wang, and S. Luo, "Thor: Towards human-level whole-body reactions for intense contact-rich environments," *arXiv preprint arXiv:2510.26280*, 2025.
- [17] C. Dong, H. Xu, W. Feng, Z. Wang, Y. Zhou, Y. Zhao, and B. He, "HAFO: A force-adaptive control framework for humanoid robots in intense interaction environments," *arXiv preprint arXiv:2511.20275*, 2025.
- [18] J. Li, L. Wu, S. A. Esteban, L. Yang, J. Drgoña, and A. D. Ames, "Accelerating and scaling MPC-guided reinforcement learning for humanoid locomotion and manipulation," *arXiv preprint arXiv:2606.05687*, 2026.
- [19] A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada, "Control barrier function based quadratic programs for safety critical systems," *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 3861–3876, 2017.
- [20] A. D. Ames, S. Coogan, M. Egerstedt, G. Notomista, K. Sreenath, and P. Tabuada, "Control barrier functions: Theory and applications," in *18th European Control Conference (ECC)*, 2019, pp. 3420–3431.
- [21] L. Yang, B. Werner, M. de Sa, and A. D. Ames, "CBF-RL: Safety filtering reinforcement learning in training with control barrier functions," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2026, arXiv preprint arXiv:2510.14959.
- [22] J. Kang, H. Cho, S. Kang, and D. Koo, "Splitadapter: Load-aware humanoid loco-manipulation via factorized adaptation," 2026. [Online]. Available: <https://arxiv.org/abs/2606.03297>
- [23] 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," in *Conference on Robot Learning (CoRL)*, 2025. [Online]. Available: <https://arxiv.org/abs/2505.24198>
- [24] B. Xu, H. Weng, Q. Lu, Y. Gao, and H. Xu, "FACET: Force-adaptive control via impedance reference tracking for legged robots," 2025. [Online]. Available: <https://arxiv.org/abs/2505.06883>
- [25] A. Rigo, M. Hu, S. K. Gupta, and Q. Nguyen, "Hierarchical optimization-based control for whole-body loco-manipulation of heavy objects," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 15 322–15 328.
- [26] F. Liu, Z. Gu, Y. Cai, Z. Zhou, H. Jung, J. Jang, S. Zhao, S. Ha, Y. Chen, D. Xu, and Y. Zhao, "Opt2Skill: Imitating dynamically-feasible whole-body trajectories for versatile humanoid loco-manipulation," *IEEE Robotics and Automation Letters*, 2025. [Online]. Available: <https://arxiv.org/abs/2409.20514>
- [27] J. Dao, H. Duan, and A. Fern, "Sim-to-real learning for humanoid box loco-manipulation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 16 930–16 936.
- [28] J. Li, X. Cheng, T. Huang, S. Yang, R.-Z. Qiu, and X. Wang, "AMO: Adaptive motion optimization for hyper-dexterous humanoid whole-body control," in *Proceedings of Robotics: Science and Systems (RSS)*, Los Angeles, CA, USA, June 2025.
- [29] A. Taylor, A. Singletary, Y. Yue, and A. Ames, "Learning for safety-critical control with control barrier functions," in *Proceedings of the 2nd Conference on Learning for Dynamics and Control (LADC)*, ser. Proceedings of Machine Learning Research, vol. 120. PMLR, 2020, pp. 708–717.
- [30] J. Pratt, J. Carff, S. Drakunov, and A. Goswami, "Capture point: A step toward humanoid push recovery," in *IEEE-RAS International Conference on Humanoid Robots*, 2006, pp. 200–207.
- [31] D. E. Orin, A. Goswami, and S.-H. Lee, "Centroidal dynamics of a

- humanoid robot,” *Autonomous Robots*, vol. 35, no. 2, pp. 161–176, 2013.
- [32] M. H. Raibert, *Legged Robots That Balance*. Cambridge, MA: MIT Press, 1986.
- [33] Bones Studio, “Bones rp01,” <https://bones.studio/datasets>, 2026.
- [34] J. Ye, K. Wang, C. Yuan, R. Yang, Y. Li, J. Zhu, Y. Qin, X. Zou, and X. Wang, “Dex1B: Learning with 1B demonstrations for dexterous manipulation,” in *Robotics: Science and Systems (RSS)*, 2025.
- [35] S. Ross, G. Gordon, and D. Bagnell, “A reduction of imitation learning and structured prediction to no-regret online learning,” in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, ser. Proceedings of Machine Learning Research, vol. 15. PMLR, 2011, pp. 627–635.

APPENDIX

A. Capture Point Derivation

Starting from the linear inverted pendulum model (LIPM), we assume $\mathbf{c}_{xy}$ is the xy position of the robot CoM and its z -coordinate is held constant. We denote $\mathbf{p} \in \mathbb{R}^2$ as the center of pressure (CoP) of the foot contacts with the ground plane. This gives the following dynamics:

$$\ddot{\mathbf{c}}_{xy} = \frac{g}{z}(\mathbf{c}_{xy} - \mathbf{p}),$$

where g is the gravitational acceleration and z is the constant height of the robot CoM. In our teacher training, we set z to the commanded CoM height. If we define $\omega_0 = \sqrt{g/z}$ as the natural frequency of the LIPM, then we get

$$\ddot{\mathbf{c}}_{xy} = \omega_0^2(\mathbf{c}_{xy} - \mathbf{p}).$$

If we assume a fixed $\mathbf{p} = \mathbf{p}_0$ and set $\mathbf{y} = \mathbf{c}_{xy} - \mathbf{p}_0$, then we get

$$\ddot{\mathbf{y}} = \omega_0^2 \mathbf{y},$$

which is a linear system with solution

$$\begin{aligned} \mathbf{y}(t) = & \frac{1}{2} \left(\mathbf{c}_0 - \mathbf{p}_0 + \frac{\dot{\mathbf{c}}_0}{\omega_0} \right) e^{\omega_0 t} \\ & + \frac{1}{2} \left(\mathbf{c}_0 - \mathbf{p}_0 - \frac{\dot{\mathbf{c}}_0}{\omega_0} \right) e^{-\omega_0 t}. \end{aligned}$$

Here, $\mathbf{c}_0 = \mathbf{c}_{xy}(0)$, $\mathbf{p}_0 = \mathbf{p}(0)$, and $\dot{\mathbf{c}}_0 = \dot{\mathbf{c}}_{xy}(0)$. As $t \rightarrow \infty$, the second term approaches zero while the first term diverges. Setting the coefficient of the divergent term to zero,

$$\mathbf{c}_0 - \mathbf{p}_0 + \frac{\dot{\mathbf{c}}_0}{\omega_0} = 0,$$

gives

$$\mathbf{p}_0 = \mathbf{c}_0 + \frac{\dot{\mathbf{c}}_0}{\omega_0},$$

which is the capture point solution at $t = 0$. We therefore define

$$\boldsymbol{\xi} = \mathbf{c}_{xy} + \frac{\dot{\mathbf{c}}_{xy}}{\omega_0}$$

as the instantaneous capture point, where $\boldsymbol{\xi}$ is the CoP that would capture the current state. Under the constant-height LIPM, this quantity is also the planar divergent component of motion (DCM), whose margins and excursions are reported in Sec. IV-B. Differentiating with respect to time gives

$$\dot{\boldsymbol{\xi}} = \dot{\mathbf{c}}_{xy} + \frac{\ddot{\mathbf{c}}_{xy}}{\omega_0}.$$

Substituting the LIPM dynamics gives

$$\begin{aligned} \dot{\boldsymbol{\xi}} &= \dot{\mathbf{c}}_{xy} + \omega_0(\mathbf{c}_{xy} - \mathbf{p}) \\ &= \omega_0(\boldsymbol{\xi} - \mathbf{p}), \end{aligned}$$

showing that the capture point follows the divergent dynamics of the LIPM relative to $\mathbf{p}$.

Driving $\boldsymbol{\xi} - \mathbf{p}$ to zero eliminates the divergent component of motion, leaving the convergent LIPM dynamics under

which $\mathbf{c}_{xy} \rightarrow \mathbf{p}$. The robot can regulate $\mathbf{p}$ by redistributing ground reaction forces (GRFs) across its foot contacts, including through ankle torques that shift pressure within a supporting foot (the ankle strategy) and load transfer between supporting feet. However, the set of achievable CoP locations is constrained by the support polygon $\mathcal{S}_{\text{support}}$. Thus, when $\boldsymbol{\xi} \in \mathcal{S}_{\text{support}}$, there exists an admissible CoP that can be driven toward the capture point to arrest the divergent motion. Conversely, when $\boldsymbol{\xi}$ lies outside $\mathcal{S}_{\text{support}}$, no CoP achievable under the current support configuration can coincide with the capture point.

This capture-point criterion describes stopping without stepping under the LIPM assumptions. Because the capture point includes CoM velocity, it also applies to dynamic walking, where subsequent foot placements change the support region. A capture point outside the current support polygon therefore does not by itself imply a fall.

This motivates using the distance between the capture point and the feasible support region in Eq. 9. Keeping this distance bounded maintains the robot near states in which its GRFs retain the ability to modulate the CoP toward the capture point and arrest divergent CoM motion. We next describe how CBF-RL uses this safety function to guide the wrist-force teacher during training.

B. CBF-RL Training and Reward Guidance

Safety filtering. The safe set is $\mathcal{S}_{\text{safe}} = \{\mathbf{s} \in \mathcal{S} : h_{\text{cp}}(\mathbf{s}) \geq 0\}$. For a valid CBF, continuously enforcing $\dot{h}_{\text{cp}} + \kappa_{\text{CBF}} h_{\text{cp}} \geq 0$, with $\kappa_{\text{CBF}} > 0$, gives $h_{\text{cp}}(t) \geq e^{-\kappa_{\text{CBF}} t} h_{\text{cp}}(0)$: a trajectory starting in the safe set stays there, provided the constraint remains feasible under the modeled dynamics [19], [20].

Our training filter uses a local approximation of this condition. Let $\mathbf{q}_\ell$ be the measured lower-body joint positions and $\mathbf{q}_\ell^\pi$ the joint targets generated by the policy action $\mathbf{a}_t$. With local barrier sensitivity $\mathbf{g}_{\text{cp}}$ and per-step gain $\alpha_{\text{CBF}} = \Delta t \kappa_{\text{CBF}}$, the CBF-QP finds the closest target satisfying the constraint:

$$\begin{aligned} \mathbf{q}_\ell^{\text{QP}} = & \arg \min_{\mathbf{q}_\ell^{\text{tgt}}} \frac{1}{2} \|\mathbf{q}_\ell^{\text{tgt}} - \mathbf{q}_\ell^\pi\|_2^2 \\ \text{s.t. } & \mathbf{g}_{\text{cp}}^\top (\mathbf{q}_\ell^{\text{tgt}} - \mathbf{q}_\ell) + \alpha_{\text{CBF}} h_{\text{cp}} \geq 0. \end{aligned} \quad (10)$$

An admissible proposal passes through unchanged. For nonzero sensitivity, a violating proposal is projected onto the constraint boundary in closed form, avoiding an iterative optimization solver in the training loop.

Reward guidance. During training, the simulator executes the filtered target $\tilde{\mathbf{q}}_\ell$, while the policy is optimized using its original sampled action. Following CBF-RL [21], we add

$$\begin{aligned} r_{\text{CBF}} = & \Delta t w_{\text{CBF}} \left(\min(\bar{\rho}_{\text{cp}}, 0) \right. \\ & \left. + \exp \left[-\frac{\|\tilde{\mathbf{q}}_\ell - \mathbf{q}_\ell^\pi\|_2^2}{\sigma_{\text{CBF}}^2} \right] - 1 \right), \end{aligned} \quad (11)$$

where the proposed barrier residual used for the reward is

$$\begin{aligned} \bar{\rho}_{\text{cp}} = & \mathbf{g}_{\text{cp}}^\top (\mathbf{q}_\ell^\pi - \mathbf{q}_\ell) \\ & + \alpha_{\text{CBF}} \max(h_{\text{cp}}, h_{\text{min}}). \end{aligned} \quad (12)$$

The first term penalizes the proposed violation; the second penalizes the correction. This encourages the policy to need less intervention. The filter and auxiliary reward are removed at deployment.

Implementation. We approximate $\mathbf{g}_{cp}$ by $\mathbf{J}_{CoM,xy}^\top \nabla_{\xi_f} h_{cp}$, using the horizontal CoM Jacobian restricted to lower-body joints and differentiating only the CoM-position contribution to the capture point. We cap the correction norm from the closed-form QP solution at $0.30 \max(\|\mathbf{q}_\ell^T - \mathbf{q}_\ell\|_2, 10^{-3} \text{ rad})$ to obtain $\tilde{\mathbf{q}}_\ell$ and regularize the sensitivity norm near zero. The reward uses $w_{CBF} = 150$, $\sigma_{CBF} = 0.5 \text{ rad}$, and $h_{min} = -1$; the QP uses the unclipped barrier value. The approximate, capped training filter and the unfiltered deployed policy do not inherit the ideal invariance guarantee. Section IV-B empirically evaluates the learned wrist-force teachers' tracking performance and disturbance recovery.

C. Wrist-Force Teacher MPC Optimization

We specialize Eqs. (1) and (6) to two foot wrenches and two wrist forces, with state $\mathbf{x} = (\mathbf{c}, \boldsymbol{\Theta}, \mathbf{l}, \boldsymbol{\kappa})$ and the input

$$\mathbf{u}_k = (\mathbf{f}_{L,k}, \boldsymbol{\tau}_{L,k}, \mathbf{f}_{R,k}, \boldsymbol{\tau}_{R,k}, \mathbf{f}_{L,k}^h, \mathbf{f}_{R,k}^h) \in \mathbb{R}^{18}. \quad (13)$$

All forces act on the robot in the world frame; $\mathbf{d}_h(t)$ is the prescribed force applied at wrist h , and $\mathbf{f}_{h,k}^h$ is the corresponding MPC decision variable. These prescribed forces represent external loading without explicitly modeling a payload's mass distribution or inertial dynamics.

For wrist sites $\mathcal{H} = \{L, R\}$ at time-varying positions $\mathbf{p}_h(t)$, the perturbed linear and angular momenta $\hat{\mathbf{l}}$ and $\hat{\boldsymbol{\kappa}}$ satisfy

$$\dot{\hat{\mathbf{l}}}(t) = \dot{\mathbf{l}}(t) + \sum_{h \in \mathcal{H}} \mathbf{d}_h(t), \quad (14)$$

$$\dot{\hat{\boldsymbol{\kappa}}}(t) = \dot{\boldsymbol{\kappa}}(t) + \sum_{h \in \mathcal{H}} (\mathbf{p}_h(t) - \mathbf{c}(t)) \times \mathbf{d}_h(t). \quad (15)$$

where the unhatted rates use the same robot state and foot wrenches without wrist perturbations.

The prediction uses application points from forward kinematics (FK) of the retargeted arm command:

$$\mathbf{p}_h^{\text{cmd}}(t) = \mathbf{p}_T(t) + \mathbf{R}_T(t) \text{FK}_h(\mathbf{q}_{\text{arm}}^{\text{cmd}}(t)), \quad (16)$$

where $(\mathbf{p}_T, \mathbf{R}_T)$ is the measured torso pose and FK_h returns torso-frame positions. Motion-frame sampling and blending follow Sec. E.

At solve time t_s , Eq. (1) uses the costs in Eqs. (2) and (4) and input reference

$$\mathbf{u}_k^{\text{ref}} = (\mathbf{0}_{12}, \mathbf{d}_L(t_s), \mathbf{d}_R(t_s)). \quad (17)$$

The guidance reward in Eq. (5) tracks

$$\begin{aligned} \Lambda_{\text{CoM}}(\mathbf{x}, \mathbf{u}) &= \mathbf{c}, & \Lambda_{\text{CoMvel}}(\mathbf{x}, \mathbf{u}) &= \mathbf{l}/m, \\ \Lambda_{\text{GRF},i}(\mathbf{x}, \mathbf{u}) &= \mathbf{f}_i, & \Lambda_{\text{resist},h}(\mathbf{x}, \mathbf{u}) &= \mathbf{f}_h^h. \end{aligned} \quad (18)$$

D. Grasp Teacher MPC Optimization

Across the training object set, the grasp teacher allocates contact wrenches at each established torso-and-arm grasp to balance the assigned object, then supplies their reaction to the locomotion MPC.

Object equilibrium. For the object assigned to each environment, let $\mathcal{G}$ be the active grasp contacts, $\mathbf{p}_j$ their world-frame positions, and $(m_o, \mathbf{c}_o)$ its mass and CoM. Define $\mathbf{g}_j$ and $\boldsymbol{\nu}_j$ as the force and contact moment exerted by the robot *on the object* at contact j . Under the quasi-static object model, equilibrium requires

$$\sum_{j \in \mathcal{G}} \mathbf{g}_j + m_o \mathbf{g} = \mathbf{0}, \quad (19)$$

$$\sum_{j \in \mathcal{G}} [(\mathbf{p}_j - \mathbf{c}_o) \times \mathbf{g}_j + \boldsymbol{\nu}_j] = \mathbf{0}. \quad (20)$$

The coupled friction constraint is

$$\|\boldsymbol{\eta}_j\|_2 \leq g_j^n, \quad 0 \leq g_j^n \leq g_j^{n,\max}, \quad (21)$$

where g_j^n is the inward normal force, $g_j^{n,\max}$ its cap, and $\boldsymbol{\eta}_j$ stacks the two tangential-force and three contact-moment components in the contact frame, divided by their respective sliding, torsional, and rolling friction coefficients. Zero-friction components are fixed to zero and omitted. We then apply these forces to the centroidal dynamics and integrate the equilibrium constraints into the MPC for the contact force decision variables to solve.

Grasp-force guidance through Λ . Let $\mathbf{z}^g$ collect grasp wrenches and $\mathbf{u}^f$ foot wrenches. The scale $\ell_\tau > 0$ converts moments to force units. The CoM and GRF maps in Eq. (18) are retained, while the grasp output for each torso or arm body b is

$$\Lambda_{\text{grasp},b}(\mathbf{x}, \mathbf{u}^f, \mathbf{z}^g) = (\mathbf{G}_b, \mathbf{N}_b/\ell_\tau), \quad (22)$$

where $\mathbf{G}_b = \sum_{j \in \mathcal{G}_b} \mathbf{g}_j$ and $\mathbf{N}_b = \sum_{j \in \mathcal{G}_b} [(\mathbf{p}_j - \mathbf{p}_b) \times \mathbf{g}_j + \boldsymbol{\nu}_j]$, with $\mathcal{G}_b$ the contacts on body b and $\mathbf{p}_b$ the common reference point for planned and measured wrenches. Their tracking error and reward are

$$\begin{aligned} E_g &= \sum_b \|\tilde{\Lambda}_{\text{grasp},b} - \Lambda_{\text{grasp},b}^{\text{meas}}\|_2^2, \\ r_{\text{grasp}} &= \begin{cases} (1 + E_g/\sigma_g^2)^{-1}, & \text{valid allocation,} \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \quad (23)$$

with $\sigma_g = 25 \text{ N}$ and tilde denoting the plan. Validity requires active contact, finite wrenches, and solver convergence with Eqs. (19) to (21) satisfied within numerical tolerances. Otherwise, measured contact reactions supply the locomotion MPC load.

E. Wrist-Force Teacher Training

Upper-body command curriculum. We uniformly sample individual frames, not sequences, from 48,233 retargeted frames across 252 motions, retaining 14 arm joint angles. For environment e , coverage $c_e \in [0.3, 1]$ contracts the sampled pose toward the nominal pose:

$$\mathbf{q}_e^{\text{ref}} = \mathbf{q}^0 + c_e(\mathbf{q}_e^{\text{sample}} - \mathbf{q}^0). \quad (24)$$

Joint-limited targets resample every 3–8 s, with a 1 s linear blend from the measured pose followed by a hold.

Starting at $c_e = 0.3$, coverage updates at reset using the completed episode length L_e in policy steps:

$$c_e^+ = \text{clip}_{[0.3,1]} (c_e + 0.01 \mathbf{1}[L_e > 200] - 0.01 \mathbf{1}[L_e < 40]). \quad (25)$$

At 50 Hz the thresholds are 4 s and 0.8 s.

Force curriculum. Each force-active environment independently updates a magnitude factor $f_e \in [0, 1]$, initialized to zero, at reset:

$$f_e^+ = \text{clip}_{[0,1]} (f_e + 0.02 \mathbf{1}[L_e > 200] - 0.02 \mathbf{1}[L_e < 200]). \quad (26)$$

A fixed 20% subset remains force-free with no curriculum advancement, preserving exposure to unloaded conditions throughout training and discouraging reliance on external forces for balance and command tracking.

For each side, heading-frame components are sampled independently from $[-40, 40]$, $[-40, 40]$, and $[-50, 5]$ N, scaled by f_e , independently zeroed with probability 0.25, and held for 3–5 s. During walking, horizontal forces oppose the planar velocity command:

$$\mathbf{u} = -\frac{\mathbf{v}_{xy}^{\text{cmd}}}{\|\mathbf{v}_{xy}^{\text{cmd}}\|_2}, \quad \mathbf{F}_{xy} \leftarrow |\mathbf{F}_{xy}^\top \mathbf{u}| \mathbf{u}. \quad (27)$$

Standing retains the sampled direction; vertical forces are unchanged. Current base yaw rotates the forces into the world frame.

F. Grasp Teacher Training

Initialization and commanded grasp. We initialize from a locomotion and static-arm checkpoint, not a grasp-force-trained policy. In the initialization study, only actor and critic network weights transfer; optimizer and curriculum state start afresh. The grasp bank spans multiple object geometries, with each environment assigned an object and a compatible grasp. Reset restores the authored arm posture and the assigned object’s transform together, clears velocity, and resets the relative object reference for slip and drop measurement (Fig. 6). Upper-body residuals about the grasp use a 0.05 rad scale; lower-body offsets use 0.5 rad. Motion-bank coverage is disabled to preserve the grasp geometry.

Per-environment payload curriculum. With an assigned object–grasp pair per environment, each environment maintains an independent upper mass bound b_e , initially 0.5 kg. At reset,

$$m_e \sim \mathcal{U}(0.5, b_e), \quad 0.5 \leq b_e \leq 5 \text{ kg}. \quad (28)$$

A full 20 s episode without terminal failure increases the bound; a terminal failure decreases it:

$$b_e^+ = \begin{cases} \min(5, 1.1b_e), & \text{full-episode success,} \\ \max(0.5, b_e/1.1), & \text{terminal failure,} \\ b_e, & \text{otherwise.} \end{cases} \quad (29)$$

Masses are in kilograms. Administrative resets and shortened timeouts do not increase b_e . Mass stays fixed within each



Fig. 6: **Example whole-body grasp configurations.** Authored torso-and-arm postures and object transforms are restored together at reset. These illustrate initialization geometry, not cross-object policy evaluation.

episode, with inertia scaled proportionally at reset. For resampled grasps, the bound follows the identity of the grasp and its object across resets. Duplicate grasps in a reset batch update at most once, with terminal failure taking precedence.

G. Teacher Training Curves

Fig. 7 reports rank-zero rolling returns over 100 completed episodes for one run of each teacher, both using seed 42. The wrist-force run spans 15,000 updates with 8 synchronized ranks; the grasp run spans 501 new updates with 32 synchronized ranks. Both use 4096 environments per rank and 24 steps per update. The grasp curve measures early adaptation, not convergence.

H. Teacher MDPs

The wrist-force and grasp teachers use separate Markov decision processes (MDPs), with shared actor observations listed in Tab. IV. Their rewards combine command tracking, MPC guidance, and action and joint regularization; the grasp teacher additionally rewards object retention (Sec. I). Critic networks receive privileged simulation observations (Tab. V).

Domain randomization varies foot friction over 0.3–1.2, proportional/derivative gains by $\times 0.8$ –1.2, rotor inertia by $\times 0.5$ –1.5, and base mass by -1 to $+3$ kg. Base CoM offsets span ± 0.025 m horizontally and ± 0.030 m vertically; pseudo-inertia parameter $\alpha \in [-0.112, 0.091]$ and encoder bias spans ± 0.015 rad. Action delay ramps to 20 ms over 2500 updates after a 1500-update warmup.

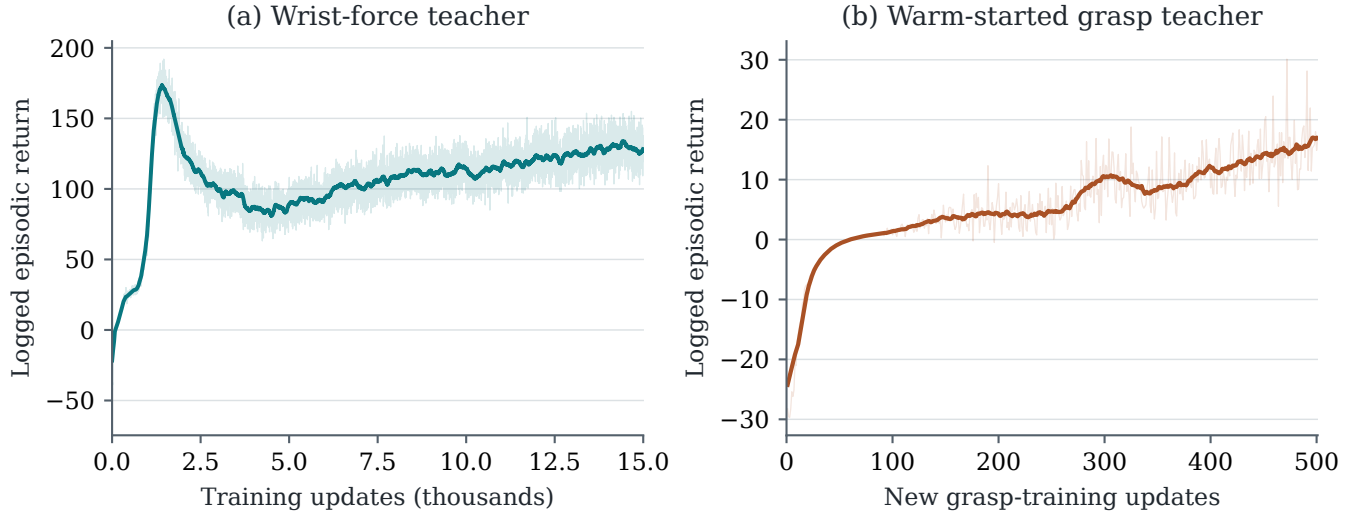


Fig. 7: **Recorded teacher training returns.** Both panels show the reward training curve during training for both teachers. The plots use a moving average filter on the actual training curve to produce a cleaner plot, but the original curve is shown in a lighter color. For the grasp teacher, we do not include the pre-training steps.

Observation	Values per sample
Base angular velocity	3
Projected gravity	3
Joint position / velocity	29 / 29
Previous action	29
Commanded planar velocity / yaw rate	3
Gait phase (sine and cosine)	2
Base-height command	1
Torso roll/pitch error	2
Arm joint target	14
Actor total, one sample	115
Actor total, five samples	575

TABLE IV: **Shared observations for wrist-force and grasp teacher actors.** Observation noise is uniform: angular velocity ± 0.2 rad/s, gravity components ± 0.05 , joint position ± 0.01 rad, and joint velocity ± 1.5 rad/s. Privileged critic observations are listed in Tab. V.

Wrist-force training uses the applied-force curriculum in Sec. E, with base-velocity perturbations every 4–6 s: $v_x, v_y \in [-2, 2]$ m/s, $v_z \in [-0.4, 0.4]$ m/s, roll/pitch rates within ± 0.52 rad/s, and yaw rate within ± 0.78 rad/s. Sustained pushes are attempted at the same interval with probability 0.5 and last 0.5–1.2 s; their magnitude ramps to 150 N over 4000 updates after a 300-update warmup. Grasp training disables these perturbations and uses loads from multiple objects, with initialization and payload sampling as in Sec. F. Its height and torso commands resample every 5 s, and action delay is sampled uniformly in simulation steps. Both teachers terminate on base tilt beyond 70° or non-foot ground contact. Grasp episodes also terminate on invalid simulation states or object drop: relative object drift above 0.15 m or height below 0.30 m, monitored from reset.

I. Reward Formulation

For body partition $g \in \{\text{lower}, \text{upper}\}$, the reward is $r_g = \Delta t \sum_{i \in \mathcal{I}_g} w_i f_i$, where $\mathcal{I}_g$ contains its reward terms, f_i

Critic-only observation	Values
Base linear velocity	3
Foot height / air time / contact flag	2 + 2 + 2
Foot contact-force vectors	6
MPC CoM reference	3
MPC angular-momentum reference	3
MPC tilt reference	3
Filtered yaw rate	1
Additional privileged values	25
Actor channels, one current sample	115
Total per critic	140

TABLE V: **Privileged observations for teacher critics.** Each critic receives 140 current-state values without added noise. Actors instead stack five samples of 115 values, spanning 80 ms at 50 Hz, without direct object-pose or contact-force observations.

is the term value, w_i its weight, and $\Delta t = 0.02$ s. Tabs. VII and VIII summarize the term formulations and weights for both teachers; the barrier reward is defined in Sec. B. We write $[x]_+ = \max(x, 0)$ and $\mathbf{1}[A]$ for the indicator of condition A . Motion gates use $s = \|\mathbf{v}_{xy}^{\text{cmd}}\|_2 + |\omega_z^{\text{cmd}}|$; gated terms are zero when inactive unless stated otherwise. The vector $\hat{\mathbf{g}}^B$ is unit gravity in frame B , and wrap maps angles to $[-\pi, \pi)$.

MPC errors $\mathbf{e}_{x,\ell} = \mathbf{x} - \mathbf{x}_\ell^{\text{MPC}}$ use lookahead index ℓ ; $\dot{\kappa}$ is estimated by finite differences. For foot i , c_i , c_i^{gait} , and d_i indicate measured contact, scheduled contact, and first touchdown; h_i, h_i^{peak} are sole and peak swing heights, and $\mathbf{v}_i, \mathbf{F}_i$ are velocity and contact force. Lengths and times are in meters and seconds.

J. Teacher Dual-Policy Training

Following [15], each teacher uses separate lower-body and upper-body actor-critic pairs. Both actors receive the same full-body history and commands, while both critics receive

Hyperparameter	Value
Rollout length per environment	24 steps (0.48 s)
Samples per GPU (4096 environments)	98304
Learning epochs / minibatches	5 / 4
Samples per minibatch per GPU	24576
Optimizer steps/partition/update	20
Discount γ / GAE λ	0.99 / 0.95
PPO clipping ratio	0.2
Entropy / value-loss coefficient	0.01 / 1.0
Clipped value loss	Enabled
Per-minibatch advantage normalization	Disabled
Initial learning rate	10^{-3}
Learning-rate schedule	Adaptive, per partition
Target KL divergence	0.01
Maximum gradient norm	1.0
Optimizers per teacher	Two Adam optimizers
Action standard deviation bounds	[0.01, 5]
Action clipping	None
Intrinsic reward	None

TABLE VI: **Shared PPO settings.** Generalized advantage estimation (GAE) uses $\lambda = 0.95$; the adaptive learning-rate schedule targets Kullback–Leibler (KL) divergence 0.01 separately for each body partition. Batch counts are per graphics processing unit (GPU), each simulating 4096 environments, with synchronized training across GPUs. The two body partitions share transitions but use separate rewards and updates.

the same privileged observations. Their Gaussian actions concatenate into 15 lower-body and 14 upper-body joint offsets. The pairs train on shared trajectories using separate reward streams (Tabs. VII and VIII) and Adam optimizers under Proximal Policy Optimization (PPO; Tab. VI).

Term	Formulation	Wrist-force	Grasp
<i>MPC guidance</i>			
CoM position	$\exp[-\sum_{\ell=1}^5 \alpha_{\ell}^c \ \mathbf{e}_{c,\ell}\ _2^2]$.	2	2
CoM velocity	$\exp[-8 \sum_{\ell=1}^3 \alpha_{\ell}^v \ \mathbf{e}_{v,\ell}\ _2^2]$.	4	4
Angular momentum	$-\sum_{\ell=1}^5 \alpha_{\ell}^a (\ \mathbf{e}_{\kappa,\ell}\ _2^2 + 0.01 \mathbf{e}_{\kappa,\ell}^{\top} \mathbf{Q}_{\kappa} \mathbf{e}_{\kappa,\ell})$.	0.01	0.01
Ground reaction forces	$\exp[-\frac{1}{ \mathcal{V} } \sum_{i \in \mathcal{V}} \ \mathbf{e}_{F,i}\ _2^2 / (150 \text{ N})^2]$; $\mathcal{V}$: valid loaded stance feet; 0 if empty.	0.002	0.002
Swing foothold	$\frac{1}{ \mathcal{S} } \sum_{i \in \mathcal{S}} \exp[-\ \mathbf{e}_{p,i,xy}\ _2^2 / (0.08 \text{ m})^2]$; $\mathcal{S}$: valid swing feet; 1 if empty or $s \leq 0.1$.	2	2
Base tilt	$\exp[-4 \sum_{\ell=1}^3 \alpha_{\ell}^{\theta} \ \mathbf{e}_{\theta,\ell}\ _2^2]$.	2	2
External-force resistance [†]	$\frac{1}{2} \sum_{i=1}^2 \exp[-\ \mathbf{F}_i^{e,\text{applied}} - \mathbf{F}_i^{e,\text{MPC}}\ _2^2 / (30 \text{ N})^2]$; inactive sides contribute 1.	1	—
<i>Command tracking</i>			
Linear velocity	$\exp[-(\ \mathbf{v}_{xy}^b - \mathbf{v}_{xy}^{\text{cmd}}\ _2^2 + (v_z^b)^2) / 0.5^2]$; base frame.	0.5	0.5
Yaw rate	$\exp[-(\bar{\omega}_z - \omega_z^{\text{cmd}})^2 / 0.2^2]$; 0.4 s yaw-rate filter.	3	3
Base height	$\exp(- z - z^{\text{cmd}} / 0.1)$.	3	3
Torso roll/pitch	$\exp(-\ \boldsymbol{\theta} - \boldsymbol{\theta}^{\text{cmd}}\ _2 / 0.15)$; $\boldsymbol{\theta}$: torso roll/pitch.	3	3
Arm joint targets [†]	$\exp[-\sum_{j \in \text{arms}} (q_j - q_j^{\text{cmd}})^2 / 0.5^2]$.	4	4
<i>Gait and swing shaping</i>			
Gait-phase contact match	$\frac{1}{2} \sum_{i=1}^2 \mathbf{1}[c_i = c_i^{\text{gait}}]$, for $s > 0.2$.	4	4
Air time	$\sum_{i=1}^2 \mathbf{1}[0.02 < t_i^{\text{air}} < 0.45]$, for $s > 0.5$.	4	4
Flat sole orientation	$\frac{1}{2} \sum_{i=1}^2 \exp[-2(Q_{i,x}^2 + Q_{i,y}^2) / 0.15]$; $\mathbf{Q}_i$: sole quaternion.	1	1
Center of pressure	$\frac{1}{2} \sum_{i=1}^2 \exp[-\tilde{e}_i^2 / 0.25^2]$, where $\tilde{e}_i = (\rho_i - 0.5)(1 + \mathbf{1}[\rho_i < 0.5])$ and ρ_i is toe-load fraction; unloaded feet contribute one.	1	1
Foot clearance	$\sum_{i=1}^2 [h_i - 0.12 - 0.02]_+$, for $s > 0.2$.	-5	-5
Sole swing height	$\sum_{i=1}^2 d_i (h_i^{\text{peak}} / 0.12 - 1)^2$, for $s > 0.2$.	-20	-20
Foot slip	$\sum_{i=1}^2 c_i \ \mathbf{v}_{i,xy}\ _2^2$, for $s > 0.05$.	-0.1	-0.1
Soft landing	$\sum_{i=1}^2 d_i \ \mathbf{F}_i\ _2$, for $s > 0.05$.	-10^{-5}	-10^{-5}
Support-mode mismatch	$\mathbf{1}[n_c \neq 1]$ for $s > 0.2$, otherwise $\mathbf{1}[n_c < 2]$, with $n_c = c_1 + c_2$.	-2	-2
Feet separation	$\max(0.12 - \ \mathbf{p}_{L,xy} - \mathbf{p}_{R,xy}\ _2, 0)$.	-0.5	-0.5
Sole orientation	$\sum_{i=1}^2 \ \hat{\mathbf{g}}_{xy}^{\text{sole}_i}\ _2$.	-2	-2
Ankle roll	$\sum_{j \in \text{ankle roll}} q_j $.	-2	-2
Feet heading	$\sum_{i=1}^2 \text{wrap}(\psi_i - \psi_{\text{base}} - 0.4 \omega_z^{\text{cmd}}) $, where ψ denotes heading.	-1	-1

TABLE VII: **Teacher tracking and gait rewards.** [†] marks upper-body terms; others enter the lower-body reward. A dash denotes an absent term.

Term	Formulation	Wrist-force	Grasp
<i>Posture and standstill</i>			
Pelvis upright	$\exp[-5\ \hat{\mathbf{g}}_{xy}^{\text{pelvis}}\ _2^2]$.	1	1
Torso upright	$\exp[-5\ \hat{\mathbf{g}}_{xy}^{\text{torso}}\ _2^2]$.	0	1
Nominal posture	$\exp[-\text{mean}_j((q_j - q_j^0)^2/\sigma_j^2)]$ over legs and waist, with speed-dependent joint tolerances.	1	1
Base-height error	$(1 + 4\mathbf{1}[s \leq 0.2])(z - z^{\text{cmd}})^2$.	-10	-10
Leg splay	$\sum_{j \in \text{hip roll/yaw}} q_j^2$.	-4	-4
Waist anchor	$(q_{\text{waist,roll}} - \theta_{\text{roll}}^{\text{cmd}})^2 + (q_{\text{waist,pitch}} - \theta_{\text{pitch}}^{\text{cmd}})^2$.	-1	-1
Waist yaw	$q_{\text{waist,yaw}}^2$.	-6	-2
Chest heading	$ \text{wrap}(\psi_{\text{torso}} - \psi_{\text{base}}) $, where ψ denotes heading.	-6	-2
Waist effort	$\sum_{j \in \text{waist}} [\max(\tau_j /\tau_j^{\text{max}} - 0.7, 0)]^2$.	-10	-10
Body angular velocity	$\omega_x^2 + \omega_y^2$ for the pelvis angular velocity in the world frame; yaw is excluded.	-0.05	-0.05
Standing joint deviation	$\sum_{j \in L \cup U} (q_j - q_j^0)^2$, when all task commands request rest.	-1	-1
Standing base drift	$\ \mathbf{v}_{\text{base},xy}\ _2^2$, when all task commands request rest.	-6	-6
<i>Joint regularization</i>			
All-action rate	$\ \mathbf{a}_t - \mathbf{a}_{t-1}\ _2^2$ over all raw joint actions.	-0.25	-0.25
Upper action rate [†]	$\ \mathbf{a}_t^U - \mathbf{a}_{t-1}^U\ _2^2$ over raw arm actions.	-0.1	-0.1
All-joint limits	$\sum_{j \in L \cup U} ([q_j^{\min} - q_j]_+ + [q_j - q_j^{\max}]_+)$.	-1	-1
Upper joint limits [†]	$\sum_{j \in U} ([q_j^{\min} - q_j]_+ + [q_j - q_j^{\max}]_+)$.	-1	-1
Lower joint velocity	$\sum_{j \in L} \dot{q}_j^2$.	-2.5×10^{-3}	-2.5×10^{-3}
Upper joint velocity [†]	$\sum_{j \in U} \dot{q}_j^2$.	-10^{-3}	-10^{-3}
Lower joint acceleration	$\sum_{j \in L} \ddot{q}_j^2$.	-7.5×10^{-7}	-7.5×10^{-7}
Upper joint acceleration [†]	$\sum_{j \in U} \ddot{q}_j^2$.	-1.2×10^{-8}	-1.2×10^{-8}
Lower joint torque	$\sum_{j \in L} \tau_j^2$.	-8.75×10^{-5}	-8.75×10^{-5}
Upper joint torque [†]	$\sum_{j \in U} \tau_j^2$.	-3.5×10^{-5}	-3.5×10^{-5}
<i>Object support and retention</i>			
Grasp-wrench tracking [†]	$[1 + E_g/(25\text{ N})^2]^{-1}$ for a valid allocation, otherwise zero; E_g is the squared wrench error in Eq. (23).	—	8
Grasp-force regularization [†]	$e/(e + 50\text{ N})$, where $e = \max(F - 30\text{ N}, 0)$ and F sums measured contact-region force magnitudes.	—	-0.5
Object drop, lower	$\mathbf{1}[\text{object-drop termination}]$	—	-250
Object drop, upper [†]	$\mathbf{1}[\text{object-drop termination}]$	—	-250
Object slip, lower	$\ \mathbf{v}_{\text{obj}/T}\ _2$, relative to a torso	—	-2
Object slip, upper [†]	$\ \mathbf{v}_{\text{obj}/T}\ _2$, relative to a torso	—	-2
Lower torque-limit excess	$\sum_{j \in L} \max(\tau_j - 0.9\tau_j^{\text{max}}, 0)$.	—	-0.1
Upper torque-limit excess [†]	$\sum_{j \in U} \max(\tau_j - 0.9\tau_j^{\text{max}}, 0)$.	—	-0.1

TABLE VIII: **Teacher posture, regularization, and grasp rewards.** L : leg/waist joints; U : arm joints; q_j^0 : default joint position; $q_j^{\min}, q_j^{\max}$: soft limits. Markers follow Tab. VII.