

COMPLIANTWBC: Whole-Body Compliance for Heavy Humanoids via Force Latent Estimation and Residual Impedance Targets

Anonymous Author(s)

Affiliation

Address

email

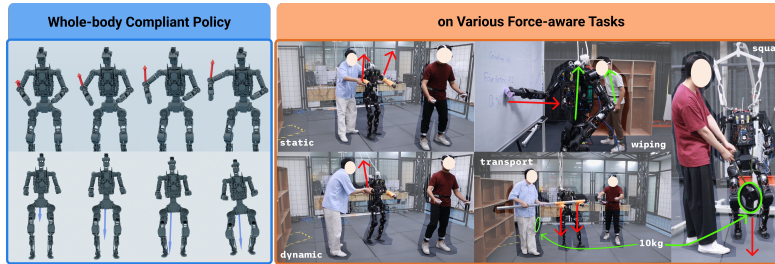


Figure 1: COMPLIANTWBC pairs a force-aware base policy with a bounded residual on the impedance equilibrium, trained under a Phong-weighted force-origin sampler that extends compliance to the whole body. We demonstrate it on a range of force-aware tasks on a real heavy humanoid.

Abstract: Whole-body compliant control is essential for deploying heavy humanoids under high payload in human-centric environments. Most prior force-aware learning-based pipelines focus on end-effector resistance, per-link upper-body springs, or end-effector stiffness modulation, leaving *arbitrary-site perturbations on heavy platforms with lower-body engagement* largely unaddressed. We close this gap with COMPLIANTWBC comprising: (1) A multi-site whole-body impedance *reference controller* generalizes classical Cartesian impedance to any controlled link via null-space projection and a multi-contact balancing quadratic programming (QP) and supplies per-link virtual targets that supervise an RL policy through a compliance-fidelity reward. (2) A *Force Latent Encoder*, co-trained with the base policy behind a gradient barrier and shaped by wrench reconstruction, supervised-contrastive clustering, a KL bottleneck, and temporal smoothness, yields a wrench-grounded, class-clustered latent; a bounded *Residual Policy* then edits the impedance equilibrium to absorb wrench-estimation error, preserving *local* classical-impedance passivity under fixed stiffness. (3) A *Phong-weighted force-origin sampler* with an axis-decoupled pelvis anchor induces lower-body-inclusive compliance via two interpretable parameters instead of a hand-crafted schedule. We evaluate COMPLIANTWBC in simulation against compliant and stiff baselines and demonstrate it on a real heavy humanoid across static/dynamic force reaction, board wiping, squat under payload, and cooperative payload transport. Project website: <https://compliantwbc.github.io/>

Keywords: Whole-Body Compliant Control, Latent Force Modeling, Residual Policy Learning

1 Introduction

Heavy humanoid loco-manipulation requires whole-body compliance. For instance, pelvis bracing against a wall, hip translation against a pulled cart, or double-support redistribution while lifting cannot be expressed by upper-body springs alone. Classical operational-space methods [1, 2, 3, 4, 5, Under submission. Do not distribute.

6, 7] cover this regime analytically, but require accurate dynamics, hand-tuned task hierarchies, and contact-mode reasoning that are brittle on heavy platforms. Learning-based pipelines relax these requirements; however, the closest predecessors restrict compliance to the end-effector or to the upper-body chain [8, 9, 10, 11, 12], leaving arbitrary-site multi-contact compliance with lower-body impedance participation largely unaddressed.

In the learning-based context, the policy observes only proprioception and a motion command (the external wrench $\mathbf{f}_{\text{ext}}$ is unobserved on hardware) and must output torques that *yield* compliantly to $\mathbf{f}_{\text{ext}}$ at arbitrary contact sites while sustaining balance and command tracking. Transferring classical compliance into a learning pipeline with reinforcement learning (RL) then faces two challenges. First, the classical impedance formulation considers a single end-effector, so a principled *whole-body* heuristic must compose centroidal-momentum balance with per-link impedance at arbitrary sites and distribute the resulting wrench to a multi-contact support set. Second, training must actually exercise the whole body (not just the end-effectors) under force perturbations, and the learned policy must *encode* compliance rather than override it. Consequently, residual schemes that only edit motor commands or tracked motions [13, 14] lose the analytical controller’s physical bounds. On the other hand, some whole-body trackers and teleoperation pipelines [15, 16, 17, 18, 19] expose no impedance interface during training, so compliance must be injected separately while running the full analytical pipeline at deployment inherits its runtime cost and contact-mode fragility on heavy platforms.

We address both challenges with COMPLIANTWBC (Figure 1). For the first, we derive an analytical multi-site reference controller (§3) composing centroidal-momentum balance with per-link Cartesian impedance at arbitrary sites. Rather than cloning its torque, we use its per-link virtual targets to supervise a base RL policy through a compliance-fidelity reward [12, 20], so the deployed policy avoids the QP’s runtime cost and contact-mode fragility. For the second, a Phong-weighted force-origin sampler (§4.2) drives perturbations across the whole body—including pelvis and lower-body sites that upper-body-only curricula never excite—so the policy *encodes* compliance rather than overriding it. A variational *Force Latent Encoder*, jointly trained with the base policy behind a gradient barrier from PPO and shaped by wrench reconstruction, contrastive clustering [21], a KL bottleneck, and temporal smoothness, supplies a class-clustered, wrench-grounded estimate of the unobserved contact state [22]. Finally, a bounded residual on the frozen base edits the impedance *equilibrium* rather than motor commands or gains [23, 24], so its effect is upper-bounded by the fixed stiffness and the closed loop is *locally* passive by the classical Cartesian-impedance argument (§F) [1, 25]. We study these behaviors on a heavy humanoid platform (70 kg, versus the 35 kg Unitree G1 [26] used in most prior learning-based compliant humanoid work; specifications in §B) to tackle a wider range of contact-rich tasks.

Our contributions can be summarized as follows: (1) An **analytical multi-site whole-body compliance reference controller** (§3, §A) composing centroidal-momentum balance with per-link Cartesian impedance at arbitrary sites, whose per-link virtual targets serve as an RL compliance-fidelity reward rather than a behavior-cloning target. (2) A **Force Latent Encoder with residual-on-equilibrium architecture** (§4.1) whose latent is grounded in privileged wrench supervision and clustered by perturbation class, and whose bounded residual edits the impedance equilibrium, preserving analytic compliance under sim-to-real adaptation. (3) A **Phong-weighted force-origin sampler with an axis-decoupled pelvis anchor** (§4.2) that induces multi-contact, lower-body-inclusive compliance, replacing a hand-crafted per-site frequency/magnitude schedule, validated on a real heavy humanoid (§5).

2 Related Work

2.1 Compliance Control for Humanoids

Classical compliance strategies, including task-space impedance and admittance control [1, 2, 3, 4, 5, 6, 7], regulate interaction forces via virtual mass–spring–damper dynamics, with operational-space extensions that handle multiple frames through hierarchical task-priority projection and a

Table 1: Humanoid whole-body-control landscape. *Lower-body compliance*: yields with the legs/pelvis (✓), only in a quadruped setting (quadruped), or not at all (–). *Multi-site*: where compliance is realized—end-effector only (EE), upper body (upper only), center-of-mass reference (CoM), or arbitrary body links (whole-body). *Force latent*: an explicit, policy-conditioning force/contact embedding (✓), an implicit one (implicit), or none (–). *Residual architecture*: where a learned correction is applied.

Method	Lower-body Compliance	Multi-site	Force latent	Residual architecture	Humanoid platform
FALCON [8]	–	EE	–	–	G1/Booster
GentleHumanoid [11]	–	upper only	implicit	–	G1
FACET [10]	quadruped	CoM	–	impedance ref.	Go2/G1
SoftMimic [12]	✓	whole-body	–	–	G1
CHIP [9]	–	EE	–	hindsight goal	G1
ResMimic [14]	–	–	–	on motion	G1
ASAP [13]	–	–	–	on action	G1
CompliantWBC (ours)	✓	✓	✓	on equilibrium	in-house, 70 kg

balancing QP that distributes the centroidal wrench to support contacts. These pipelines provide rigorous stability guarantees but require accurate dynamics, hand-crafted task hierarchies, and explicit contact-mode reasoning that is brittle on heavy humanoid platforms; time-varying stiffness additionally breaks passivity unless an explicit storage mechanism is enforced [27, 28]. Recent whole-body tracking controllers [29, 30, 31] may achieve agile loco-manipulation via motion imitation but do not expose an impedance interface, so compliance must be injected separately.

To relax these requirements, recent learning-based methods train policies to exhibit compliant behavior through reward shaping, perturbation curricula, per-link virtual springs, or end-effector stiffness modulation [8, 11, 10, 9]. Most restrict compliance to the end-effector or the upper-body chain, leaving lower-body bracing, hip translation, and double-support redistribution—the regimes that dominate heavy loco-manipulation—unaddressed. SoftMimic [12] is a notable exception that achieves genuine whole-body compliance, but tracks a single motion rather than an arbitrary configuration. A parallel line targets heavy-payload loco-manipulation with multi-policy or trajectory-optimized references [32, 33] but similarly does not expose a per-link impedance interface. Our work composes centroidal-momentum balance and per-link Cartesian impedance at arbitrary sites inside an RL loop, admitting end-effector-only and upper-body-only formulations as special cases (Table 1).

2.2 Residual Policies and Latent Force Estimation

Residual reinforcement learning [23, 34, 35] composes a learned correction on top of a hand-crafted base controller, retaining analytical stability while absorbing unmodeled dynamics. The idea of learning residuals on controller parameters with RL was pioneered by Buchli et al. [36], who optimized both reference trajectories and gain schedules on hardware. Our work operates in the same spirit but residualizes on the impedance *equilibrium* rather than on gains. Recent humanoid pipelines instantiate this residual idea differently, focusing on motor commands, on tracked motions, or on goals via hindsight relabeling [13, 14, 9]. However, none of these approaches edit a per-link impedance interface with an explicit force bound. Consequentially, the residual’s effect on induced contact force is implicit rather than physically bounded. In contrast, we compose per-link impedance at arbitrary sites and place a *bounded* residual on the per-link *equilibrium* over a *frozen* base.

A complementary line learns compact latent embeddings of unobserved context from proprioceptive history, with supervised or contrastive objectives shown to prevent collapse to motion-phase memorization [37, 38, 39, 40, 21]. Recent work specializes such latents to estimate external wrenches from proprioception without dedicated force sensors [41, 42, 43]. Relative to this RMA-style lineage, we infer a wrench-grounded, class-clustered latent from history observations behind a gradient barrier, which drives a *bounded residual on the per-link impedance equilibrium*. Because the link stiffness $\mathbf{K}_\ell$ is held fixed, the residual’s effect is upper-bounded by an explicit virtual-target shift times the stiffness, so the closed loop is *locally* passive by the classical Cartesian-impedance argument (§F).

3 Whole-Body Impedance Heuristic

A humanoid has configuration $\mathbf{q} \in \text{SE}(3) \times \mathbb{R}^{n_j}$, generalized velocity $\boldsymbol{\nu}$, and floating-base dynamics $M\dot{\boldsymbol{\nu}} + C\boldsymbol{\nu} + \mathbf{g} = \mathbf{S}^\top \boldsymbol{\tau} + \sum_c \mathbf{J}_c^\top \boldsymbol{\lambda}_c + \sum_i \mathbf{J}_{\ell_i}^\top \mathbf{f}_{\text{ext},i}$ with support-contact reaction wrenches $\boldsymbol{\lambda}_c$ and external wrenches $\mathbf{f}_{\text{ext},i}$ at links ℓ_i . Classical Cartesian impedance at a single end-effector [1] prescribes $M_d\ddot{\mathbf{x}} + D_d\dot{\mathbf{x}} + K_d(\mathbf{x} - \mathbf{x}_d) = \mathbf{f}_{\text{ext}}$, with $\mathbf{x} \in \mathbb{R}^3$. We extend this to a hierarchy of frames coupled by balance.

Reference controller (summary). For controlled links $\mathcal{L}$ (hands, elbows, knees, pelvis, torso) and support contacts $\mathcal{C}_s$, the analytical reference controller composes (i) a centroidal-momentum balance task whose wrench is distributed to the support contacts by a balancing QP over friction cones and centers of pressure [5], (ii) per-link Cartesian impedance at arbitrary sites projected into the balance null space [3], and (iii) a joint-space posture task. It produces a reference torque $\boldsymbol{\tau}_c$ and per-link virtual targets $\mathbf{x}_{\ell,d}^c$. We give the full controller, including the energy-tank passivity treatment for time-varying stiffness, in §A, and use it strictly as a *reference*: of its outputs, *only* the per-link virtual targets $\mathbf{x}_{\ell,d}^c$ enter the policy reward (§4). The QP-distributed torque $\boldsymbol{\tau}_c$ is never cloned or rewarded, so the deployed policy inherits neither the QP’s runtime cost nor its contact-mode fragility on heavy platforms. Admitting knees and the pelvis as controlled and support sites is the mechanism by which the heuristic extends two-footed, end-effector-only settings to multi-contact, lower-body-inclusive postures.

Force-aware virtual targets. The single signal carried into training is the per-link virtual target. For a perturbed link $\ell \in \mathcal{L}$ under an external wrench $\mathbf{f}_{\text{ext}}$ applied at body site p ,

$$\mathbf{x}_{\ell,d} = \mathbf{x}_{\ell}^{\text{ref}}(t) + \mathbf{K}_{\ell}^{-1} \mathbf{f}_{\ell}(t), \quad \mathbf{f}_{\ell} = \text{Ad}_{p \rightarrow \ell}^{-\top} \mathbf{f}_{\text{ext}}, \quad (1)$$

where $\mathbf{f}_{\ell}$ is the external wrench transported to link ℓ ’s frame (equal to $\mathbf{f}_{\text{ext}}$ when applied at ℓ) and $\mathbf{K}_{\ell}^{-1}$ is the anisotropic Cartesian compliance, so the correction has units of displacement. This generalizes GentleHumanoid’s scalar spring to arbitrary sites and anisotropic stiffness: restricting $\mathcal{L}$ to upper-body links with $\mathbf{K}_{\ell} = K_p \mathbf{I}_3$, $K_p \sim \mathcal{U}(5, 250)$ N/m and $\mathbf{D}_{\ell} = 2\sqrt{m_v K_p} \mathbf{I}_3$ ($m_v = 0.1$ kg) recovers GentleHumanoid’s upper-body Cartesian-spring response [11] (up to its motion-data-driven guiding contacts), and restricting the perturbed set to the end-effectors reduces to FALCON’s end-effector perturbation model [8] as an analytical special case. Time-varying $\mathbf{K}_{\ell}$ would break passivity [28]; we hold $\mathbf{K}_{\ell}$ fixed (§4.1), which makes the closed loop locally passive by the classical Cartesian-impedance argument (§F), and defer the energy-tank treatment of the general time-varying case to §A.

4 Method: Force-Aware Base Policy and Residual on Impedance Targets

4.1 Pipeline Architecture

The analytical controller of §3 assumes rigid contacts, perfect torque control, and accurate dynamics, and assumes the external wrench $\mathbf{f}_{\text{ext}}$ is observed; none holds on a heavy humanoid. We therefore train a two-stage pipeline (Figure 2): (i) a *force-aware base policy* that jointly trains with a Force Latent Encoder to induce compliance behavior, and (ii) a *residual policy* on top of the frozen base that compensates for the discrepancy between the base’s wrench estimate and the true external wrench.

Base policy. We first train a force-aware whole-body policy $\pi_{\text{base}}(\mathbf{o}, \mathbf{c}, \mathbf{z}_F) \rightarrow \boldsymbol{\tau}$ that takes as input a proprioceptive observation $\mathbf{o}$, a motion command $\mathbf{c}$, and the latent wrench embedding $\mathbf{z}_F$ produced by the Force Latent Encoder, and outputs the joint torque. The policy is trained end-to-end with PPO [44] against a compliance-fidelity reward

$$r_t = r_{\text{track}} - w_c \|\mathbf{x}_{\ell} - \mathbf{x}_{\ell,d}^c\|^2 - w_e \|\boldsymbol{\tau}\|^2, \quad (2)$$

whose term $w_c \|\cdot\|^2$ drives the controlled links toward the analytical controller’s virtual-target deformation $\mathbf{x}_{\ell,d}^c$ (Eq. (1)) under the ground-truth perturbation. This incentivizes the policy to learn multi-site compliance without inheriting the balancing QP’s runtime cost at deployment. The full reward formulations (Table 3) and the two-stage training loop (Algorithm 1) are given in §B.

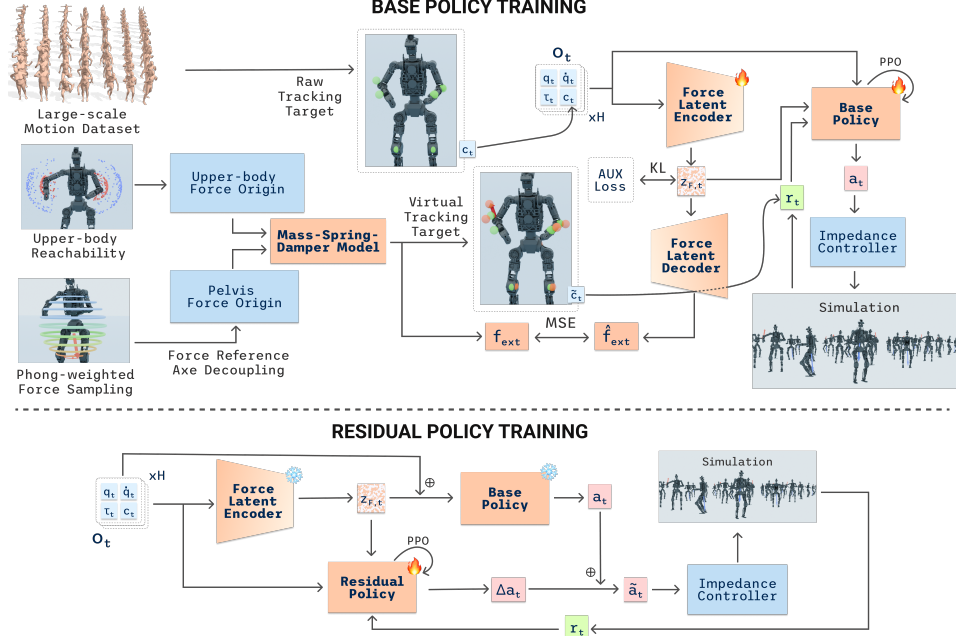


Figure 2: Two-stage training of COMPLIANTWBC: (i) a base policy jointly trained with the Force Latent Encoder to induce whole-body compliance; (ii) a bounded residual on the impedance equilibrium that compensates for physics discrepancies.

Force Latent Encoder follows a variational encoder architecture that maps an observation history window (proprioceptions, velocities, last torques, and motion command) to a force latent $z_F \in \mathbb{R}^{d_F}$. The encoder is paired with two auxiliary heads — a *wrench decoder* ψ that regresses the per-site external wrench on all controlled links, and a *projection head* g on the unit sphere dedicated to the contrastive loss. The encoder is shaped *only* by the auxiliary objective in Eq. (3) while the posterior mean is detached before it is fed to π_{base} and π_{res} . This forms a *gradient barrier* preventing PPO gradients from flowing through the encoder, so that the latent cannot be co-opted away from its wrench-grounded objective toward reward-hacking features.

The auxiliary objective $\mathcal{L}_\phi$ combines four terms: a per-site wrench-reconstruction loss (the main term) grounding z_F in the true contact wrench through ψ ; a supervised contrastive loss [21] over perturbation-class labels (which limb, which direction); a KL/VIB bottleneck [45] suppressing dimensions that carry neither wrench- nor class-relevant information; and a temporal-smoothness term on consecutive posterior means:

$$\mathcal{L}_\phi = \lambda_w \mathcal{L}_{\text{wrench}} + \lambda_s \mathcal{L}_{\text{supcon}} + \lambda_k \mathcal{L}_{\text{kl}} + \lambda_m \mathcal{L}_{\text{smooth}}. \quad (3)$$

Together they make z_F a wrench-grounded, class-clustered, low-bandwidth embedding of the privileged contact state (full forms in §B).

Residual policy is a small MLP added on top of the frozen base policy and frozen Force Latent Encoder. The residual reads $(o_t, c_t, z_F, \hat{f}_{\text{ext}})$, with the frozen wrench estimate $\hat{f}_{\text{ext}} = \psi(z_F)$, and outputs a bounded delta on the impedance equilibrium:

$$\tilde{x}_{\ell,d} = x_{\ell,d} + \epsilon_x \tanh(\Delta x_{\ell,d}), \quad (4)$$

with positional bound $\epsilon_x = 5$ cm. When the frozen wrench estimate $\hat{f}_{\text{ext}} = \psi(z_F)$ is biased or noisy with respect to the true external wrench, the residual edits the impedance *equilibrium* to absorb the discrepancy without changing the base policy and without modulating gains. Editing the equilibrium rather than the torque preserves the base controller’s analytic compliance: a 5 cm shift in $x_{\ell,d}$ induces at most $K_\ell \cdot 5$ cm of force, an explicit physical bound the residual cannot violate. We train the residual in a separate stage rather than jointly with the base: under a single PPO loss the residual would absorb compliance the base could itself learn, turning its bound into a constraint on the joint system rather than an interpretable discrepancy compensator. More details in §B.

4.2 Compliant virtual targets via external-force sampling

Phong-weighted sampling. For perturbed upper-body links (wrists, elbows) we sample force origins from a kinematic reachability cloud and drag the link toward them [11]. For the pelvis—which endures much larger forces over a limited range of motion—we instead draw a force origin from a continuous *Phong-weighted* distribution inside a ball of radius $R_a = 0.5$ m centered on the pelvis, with direction $\hat{n}$ following a Phong lobe [46] concentrated on the downward axis $-\hat{e}_z$ with exponent $n=2$:

$$p(\hat{n}) = (1 - \epsilon) \frac{n+1}{2\pi} \max(-\hat{n} \cdot \hat{e}_z, 0)^n + \epsilon \frac{1}{4\pi}, \quad (5)$$

with isotropic mixing weight $\epsilon = 0.1$, sampled in closed form by inverse CDF (§C) to give the offset $\delta = r(\sin \theta \cos \phi, \sin \theta \sin \phi, -\cos \theta)$. The downward bias emphasizes gravity on a carried load while lateral and upward samples cover reaching and lifting. Unlike a fixed cloud the distribution is gap-free, needs no forward-kinematics precomputation, and exposes its bias through two interpretable parameters (n, ϵ) (Figure 3).

Axis-decoupled pelvis anchor and compliant height target.

Anchoring the offset δ rigidly to the moving pelvis yields a self-chasing constant load, whereas a world-fixed anchor lets the force grow without bound as the robot walks away. We instead decouple the anchor by axis: the horizontal component co-moves with the floating base (leaving locomotion a bounded, body-relative load), while the vertical component—the axis we want the policy to yield along—is pinned to the *commanded* base height h^* , giving the pelvis force-origin anchor $\mathbf{a}_{\text{pelvis}}$:

$$\mathbf{a}_{\text{pelvis}}(t) = \underbrace{\mathbf{x}_{\text{pelvis}}^{xy}(t) + \delta^{xy}}_{\text{base-relative (horizontal)}} + \underbrace{(h^* + \delta^z) \hat{e}_z}_{\text{gravity-anchored (vertical)}}, \quad (6)$$

so the anchor height $z_a = h^* + \delta^z$ is a per-chunk gravity-aligned datum that does not sink with the robot. We then evaluate the base-height term of r_{track} against a compliance-aware target $z_{\text{virt}}^*(t) = h^* + \alpha_z(z_a(t) - h^*)$, where the gain $\alpha_z = k/(k + k_{\text{leg}})$ is set by a two-spring static balance between the pelvis-anchor stiffness k and the effective vertical leg stiffness k_{leg} (§C): a stiff anchor pulls the comfortable pose toward the load, a stiff stance resists it. Penalizing $(z_{\text{pelvis}} - z_{\text{virt}}^*)^2$ thus interpolates from rigid tracking ($\alpha_z = 0$) to full anchor-following ($\alpha_z = 1$), so compliant lower-body sinking emerges without an explicit compliance offset.

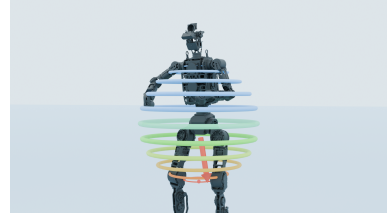


Figure 3: Phong-weighted sampling generates downward-biased force origins, decoupled by axis.

5 Experiments

Evaluation goal. We ask three questions: (i) does *whole-body* compliance from multi-site (pelvis, torso, elbow, wrist) force sampling improve over aggressive tracking and end-effector-only compliance, especially for forces away from the hands? (ii) does the residual on virtual targets improve force-aware tracking without sacrificing yielding? and (iii) how does COMPLIANTWBC behave in the real world on contact-rich tasks? We address (i)–(ii) in §5.2 and (iii) in §5.3.

5.1 Settings and Metrics

We compare against two external baselines—a stiff whole-body tracker **TWIST2** [15] (nominal command $\mathbf{x}^{\text{ref}}$, no force-aware targets) and an embodiment-specific re-implementation of upper-body-only **GentleHumanoid** [11]—and two ablations: **Ours w/o pelvis force sampling** (upper-body perturbations only, approximating GentleHumanoid-style compliance) and **Ours w/o residual** (Stage-1 base policy with the Force Latent Encoder, no Stage-2 residual). **Ours (full)** adds the bounded residual on virtual targets (Eq. (4)) with per-link stiffness fixed by the curriculum. All variants share an identical training recipe, environment budget, and paired random seeds.

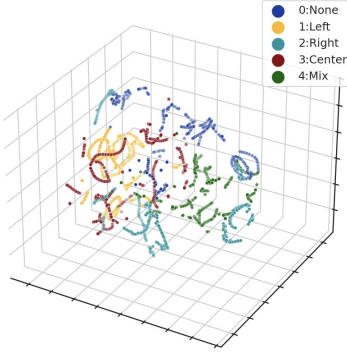


Figure 4: Force latent embedding projections via t-SNE of different force profiles across contact sites.

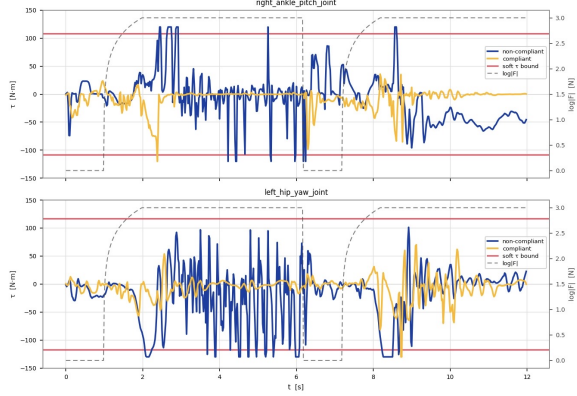


Figure 5: Compliant and Non-compliant controller reaction under pelvis perturbations.

Table 2: Performance of COMPLIANTWBC versus compliant and non-compliant baselines over 100 paired rollouts. Arrows indicate the preferred direction; bracketed columns are reported in units of 10^{-2} . $E_{\text{cmd}}^{\text{free}}/E_{\text{cmd}}^{\text{force}}$: wrist command-tracking RMSE without/with external force; ρ_{τ} : fraction of near-saturated joints (lower is more compliant); R_{LB} : fraction of torque deviation borne by the lower body; S : success (no fall) rate.

Controller	Tracking quality		WB compliance and robustness		
	$E_{\text{cmd}}^{\text{free}} \downarrow$ [$\times 10^{-2}$]	$E_{\text{cmd}}^{\text{force}} \downarrow$ [$\times 10^{-2}$]	$\rho_{\tau} \downarrow$ [$\times 10^{-2}$]	$R_{\text{LB}} \uparrow$	$S \uparrow$
TWIST2	2.18 ± 0.31	3.14 ± 0.42	2.74 ± 0.38	0.16 ± 0.04	0.59
GENTLEHUMANOID	5.22 ± 0.47	7.86 ± 0.61	2.12 ± 0.29	0.14 ± 0.03	0.91
OURS w/o pelvis force sampling	4.12 ± 0.38	5.89 ± 0.51	1.87 ± 0.24	0.16 ± 0.03	0.93
OURS w/o residual policy	4.29 ± 0.35	5.43 ± 0.47	1.56 ± 0.21	0.20 ± 0.03	0.94
OURS (full)	4.65 ± 0.41	6.57 ± 0.53	0.93 ± 0.15	0.31 ± 0.04	0.98

Metrics. Apart from comparing **Task success** S and **Tracking performance** $E_{\text{cmd}}^{\text{free}}$, we also report metrics that characterize the compliant response to external forces, including **Torque reaction** ρ_{τ} , **Lower-body participation** R_{LB} , and **Force-aware tracking** $E_{\text{cmd}}^{\text{force}}$ upon external forces. **Torque reaction** ρ_{τ} reflects the fraction of near-saturated joints. A stiff controller fights the force (high ρ_{τ}); a compliant one yields, keeping it low. **Lower-body participation** R_{LB} is the fraction of the total perturbation-induced torque deviation (relative to a matched no-perturbation rollout) borne by the lower-body joints; we treat it as a behavioral *diagnostic* of whether the policy engages the legs, not as a standalone success criterion. The metric formulations are given in the appendix.

5.2 Simulation Experiments

Protocol. Each controller runs 100 paired rollouts (a force trapezoid matching training) over a fixed site pool (wrist, elbow, torso, pelvis, hip, knee), split evenly between static (standing) and dynamic (stepping, locomotion) motions, plus a perturbation-free pass per (*controller, motion*) pair for the free-space and lower-body-participation metrics.

Results. Table 2 lays out the tracking–compliance trade-off. TWIST2 tracks best in free space but succeeds on only 59% of trials, fighting every perturbation; GENTLEHUMANOID reaches $S = 0.91$ by relaxing the upper-body limbs but shows the lowest whole-body engagement ($R_{\text{LB}} = 0.14$). OURS (FULL) sits on the favorable frontier, with the highest lower-body participation and the lowest joint saturation ρ_{τ} at the highest success rate, at only a modest free-tracking gap to TWIST2. The ablations isolate each component: removing pelvis sampling collapses R_{LB} toward the GENTLEHUMANOID regime, while removing the residual recovers slight tracking precision at the cost of compliance and robustness, demonstrated by worst joint saturation, lower-body participation, success rate. We provide further qualitative insight into the learned behavior of the Force Latent Encoder and the compliant response of the policy. Figure 4 shows the Force Latent Encoder organizes its embeddings into distinct per-site clusters (left/right limb, pelvis, mixed, and no-perturbation)—a force-origin-aware representation rather than a collapsed force/no-force mode, a prerequisite for

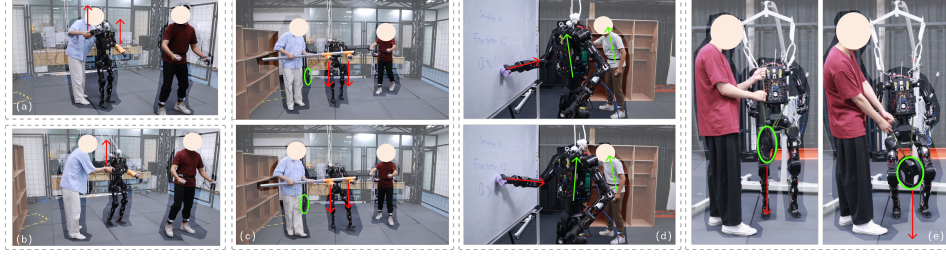


Figure 6: We deploy COMPLIANTWBC in the real world on five force-aware tasks: (a) Static Force Reaction; (b) Dynamic Force Reaction; (c) Cooperative Payload Transport; (d) Board Wiping; and (e) Squat Under Payload.

site-dependent reactions. Figure 5 contrasts compliant and non-compliant joint torques under a pelvis perturbation: the compliant policy keeps lower-body torques much smaller and stays stable through the contact, consistent with its lower ρ_τ in Table 2 and concretizing the resist-and-fail behavior behind TWIST2’s low success rate.

5.3 Real-World Experiments

We deploy our trained policy on a real humanoid via teleoperation [47, 48, 49] on five tasks: *Static Force Reaction*, *Dynamic Force Reaction*, *Cooperative Payload Transport*, *Board Wiping*, and *Squat Under Payload* (Figure 6). The first two mirror the simulation perturbation protocol; the remaining three apply sustained, structured contacts that no single-site end-effector formulation can absorb, directly stressing the multi-site, whole-body assumptions of our pipeline (§4). We report these as *qualitative* demonstrations; quantitative on-robot evaluation is left to future work.

Under large pushes while tracking a dynamic walking motion (Figure 6b,c) the robot stays balanced—the regime where the stiff TWIST2 collapses in simulation ($S = 0.59$). In *Cooperative Payload Transport* the operator hands over a 100 N bar; the robot yields at the wrists, redistributes the load through torso and hips, and re-establishes a stable gait—the yield-then-reconcile behavior of the impedance equilibrium (Eq. (1)). *Squat Under Payload* stresses the same mechanism vertically, spreading sustained pelvis load across the chain rather than saturating the hip and knee. *Board Wiping* is the clearest whole-body case: as the operator presses a board against the robot’s hand, the humanoid leans its trunk back and shifts its CoM over the support polygon—which an upper-body-only baseline cannot do, since no perturbation reaches its lower body in training. These demonstrations are qualitatively consistent with the simulation ablation.

6 Conclusion

We presented COMPLIANTWBC, a two-stage RL pipeline that defines a multi-site whole-body impedance reference controller and uses it as a compliance-fidelity reward signal. Stage 1 jointly trains the Force Latent Encoder and a force-aware base policy against this reward while stage 2 trains a bounded residual on the impedance equilibrium that compensates for the discrepancy between the frozen wrench estimator and the true external wrench, with per-link stiffness held fixed so that passivity follows from the classical Cartesian-impedance argument. We validate the result on heavy-payload loco-manipulation tasks that require lower-body and multi-contact compliance and observe compliance behavior with high lower-body participation both in simulation and in the real world.

Limitations & Future Work. Our passivity result is *local* (fixed stiffness, near the equilibrium, assuming the base realizes the analytical torque; §F), and the balancing QP (§A) can become infeasible under large simultaneous perturbations. The Force Latent Encoder is trained on simulated ground-truth wrenches, so hardware relies on proprioceptive generalization; a force–torque or current-based estimator [43, 41] would help. Finally, experiments use a single heavy platform and the real-world study is qualitative, so isolating the heaviness advantage and adding quantitative on-robot metrics remain key next steps, alongside bimanual sustained-contact manipulation and end-to-end stiffness learning under a differentiable energy-tank gate.

References

- [1] Neville Hogan. Impedance control: An approach to manipulation: Parts I–III. *Journal of Dynamic Systems, Measurement, and Control*, 107:1–24, 1985.
- [2] Christian Ott, Alin Albu-Schäffer, Andreas Kugi, and Gerd Hirzinger. On the passivity-based impedance control of flexible joint robots. *IEEE Transactions on Robotics*, 24(2):416–429, 2008.
- [3] Oussama Khatib. A unified approach for motion and force control of robot manipulators: The operational space formulation. *IEEE Journal of Robotics and Automation*, 3(1):43–53, 1987.
- [4] Luis Sentis and Oussama Khatib. Synthesis of whole-body behaviors through hierarchical control of behavioral primitives. *International Journal of Humanoid Robotics*, 2(4):505–518, 2005.
- [5] Bernd Henze, Máximo A. Roa, and Christian Ott. Passivity-based whole-body balancing for torque-controlled humanoid robots in multi-contact scenarios. *The International Journal of Robotics Research*, 35(12):1522–1543, 2016.
- [6] Alexander Dietrich. *Whole-Body Impedance Control of Wheeled Humanoid Robots*, volume 116 of *Springer Tracts in Advanced Robotics*. Springer, 2016. doi:10.1007/978-3-319-40557-5.
- [7] Alin Albu-Schäffer, Christian Ott, and Gerd Hirzinger. A unified passivity-based control framework for position, torque and impedance control of flexible joint robots. *The International Journal of Robotics Research*, 26(1):23–39, 2007. doi:10.1177/0278364907073776.
- [8] Yuanhang Zhang, Yifu Yuan, Prajwal Gurunath, Ishita Gupta, Shayegan Omidshafiei, Aliakbar Agha-mohammadi, Marcell Vazquez-Chanlatte, Liam Pedersen, Tairan He, and Guanya Shi. FALCON: Learning force-adaptive humanoid loco-manipulation. *arXiv preprint arXiv:2505.06776*, 2025.
- [9] Sirui Chen, Zi-Ang Cao, Zhengyi Luo, Fernando Castañeda, Chenran Li, Tingwu Wang, Ye Yuan, Linxi Fan, C. Karen Liu, and Yuke Zhu. CHIP: Adaptive compliance for humanoid control through hindsight perturbation. *arXiv preprint arXiv:2512.14689*, 2025.
- [10] Botian Xu, Haoyang Weng, Qingzhou Lu, Yang Gao, and Huazhe Xu. FACET: Force-adaptive control via impedance reference tracking for legged robots. *arXiv preprint arXiv:2505.06883*, 2025.
- [11] Qingzhou Lu, Yao Feng, Baiyu Shi, Michael Pisen, Zhenan Bao, and C. Karen Liu. GentleHumanoid: Learning upper-body compliance for contact-rich human and object interaction. *arXiv preprint arXiv:2511.04679*, 2025.
- [12] Gabriel B. Margolis, Michelle Wang, Nolan Fey, and Pulkit Agrawal. SoftMimic: Learning compliant whole-body control from examples. *arXiv preprint arXiv:2510.17792*, 2025.
- [13] Tairan He, Jiawei Gao, Wenli Xiao, Yuanhang Zhang, Zi Wang, Jiashun Wang, Zhengyi Luo, Guanqi He, Nikhil Sobanbab, Chaoyi Pan, et al. Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. *arXiv preprint arXiv:2502.01143*, 2025.
- [14] Siheng Zhao, Yanjie Ze, Yue Wang, C. Karen Liu, Pieter Abbeel, Guanya Shi, and Rocky Duan. ResMimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning. *arXiv preprint arXiv:2510.05070*, 2025.
- [15] Yanjie Ze, Siheng Zhao, Weizhuo Wang, Angjoo Kanazawa, Rocky Duan, Pieter Abbeel, Guanya Shi, Jiajun Wu, and C. Karen Liu. Twist2: Scalable, portable, and holistic humanoid data collection system. *arXiv preprint arXiv:2511.02832*, 2025.

- [16] Yixuan Li, Yutang Lin, Jieming Cui, Tengyu Liu, Wei Liang, Yixin Zhu, and Siyuan Huang. Clone: Closed-loop whole-body humanoid teleoperation for long-horizon tasks, 2025.
- [17] Chenhao Lu, Xuxin Cheng, Jialong Li, Shiqi Yang, Mazeyu Ji, Chengjing Yuan, Ge Yang, Sha Yi, and Xiaolong Wang. Mobile-television: Predictive motion priors for humanoid whole-body control, 2025. URL <https://arxiv.org/abs/2412.07773>.
- [18] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans, 2024. URL <https://arxiv.org/abs/2406.10454>.
- [19] Yanjie Ze, Zixuan Chen, João Pedro Araújo, Zi ang Cao, Xue Bin Peng, Jiajun Wu, and C. Karen Liu. Twist: Teleoperated whole-body imitation system, 2025. URL <https://arxiv.org/abs/2505.02833>.
- [20] Zewen He, Chenyuan Chen, Dilshod Azizov, and Yoshihiko Nakamura. Cotap: Compliant task pipeline and reinforcement learning of its controller with compliance modulation, 2025. URL <https://arxiv.org/abs/2509.25443>.
- [21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [22] Arnaud Demont, Mehdi Benallegue, Abdelaziz Benallegue, Pierre Gergondet, Antonin Daldard, Rafael Cisneros, Masaki Murooka, and Fumio Kanehiro. The kinetics observer: A tightly coupled estimator for legged robots, 2024. URL <https://arxiv.org/abs/2406.13267>.
- [23] Tobias Johannink, Shikhar Bahl, Ashvin Nair, Jianlan Luo, Avinash Kumar, Matthias Loskyll, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. Residual reinforcement learning for robot control. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 6023–6029, 2019. doi:10.1109/ICRA.2019.8794127.
- [24] Mark J. Balas and Susan A. Frost. Adaptive control of linear modal systems using residual mode filters and a simple disturbance estimator. In *Proceedings of the 2011 American Control Conference*, pages 2338–2343, 2011. doi:10.1109/ACC.2011.5991208.
- [25] Yu Zhang, Long Cheng, Xiuze Xia, and Haoyu Zhang. Learning variable impedance skills from demonstrations with passivity guarantee, 2024. URL <https://arxiv.org/abs/2306.11308>.
- [26] Unitree Robotics. Unitree G1 Humanoid Robot. <https://www.unitree.com/cn/g1/>, 2024. Accessed: 2026-05-29.
- [27] Federica Ferraguti, Cristian Secchi, and Cesare Fantuzzi. A tank-based approach to impedance control with variable stiffness. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4948–4953, 2013. doi:10.1109/ICRA.2013.6631284.
- [28] Klas Kronander and Aude Billard. Stability considerations for variable impedance control. *IEEE Transactions on Robotics*, 32(5):1298–1305, 2016. doi:10.1109/TRO.2016.2593492.
- [29] Tairan He, Wenli Xiao, Toru Lin, Zhengyi Luo, Zhenjia Xu, Zhenyu Jiang, Jan Kautz, Changliu Liu, Guanya Shi, Xiaolong Wang, Linxi Fan, and Yuke Zhu. Hover: Versatile neural whole-body controller for humanoid robots, 2025. URL <https://arxiv.org/abs/2410.21229>.
- [30] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *arXiv preprint arXiv:2406.08858*, 2024.

- [31] Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10895–10904, 2023.
- [32] Kaiyan Xiao, Zihan Xu, Zhe Cheng, Chengju Liu, and Qijun Chen. Kinematics-aware multi-policy reinforcement learning for force-capable humanoid loco-manipulation, 2025. URL <https://arxiv.org/abs/2511.21169>.
- [33] Hao Zhang, Yves Tseng, Ding Zhao, and H. Eric Tseng. Interaction-aware whole-body control for compliant object transport, 2026. URL <https://arxiv.org/abs/2603.03751>.
- [34] Tom Silver, Kelsey Allen, Josh Tenenbaum, and Leslie Kaelbling. Residual policy learning. *arXiv preprint arXiv:1812.06298*, 2018.
- [35] Xiang Zhang, Changhao Wang, Lingfeng Sun, Zheng Wu, Xinghao Zhu, and Masayoshi Tomizuka. Efficient sim-to-real transfer of contact-rich manipulation skills with online admittance residual learning. In *7th Annual Conference on Robot Learning*, 2023. URL <https://openreview.net/forum?id=gFXVysXh48K>.
- [36] Jonas Buchli, Freek Stulp, Evangelos Theodorou, and Stefan Schaal. Learning variable impedance control. *I. J. Robotic Res.*, 30:820–833, 06 2011. doi:10.1177/0278364911402527.
- [37] Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. RMA: Rapid motor adaptation for legged robots. In *Robotics: Science and Systems (RSS)*, 2021.
- [38] Ashish Kumar, Zhongyu Li, Jun Zeng, Deepak Pathak, Koushil Sreenath, and Jitendra Malik. Adapting rapid motor adaptation for bipedal robots. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1161–1168. IEEE, 2022.
- [39] Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science Robotics*, 5(47):eabc5986, 2020.
- [40] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [41] Daegyu Lim, Myeong-Ju Kim, Junhyeok Cha, Donghyeon Kim, and Jaeheung Park. Proprioceptive external torque learning for floating base robot and its applications to humanoid locomotion, 2023. URL <https://arxiv.org/abs/2309.04138>.
- [42] Peiyuan Zhi, Peiyang Li, Jianqin Yin, Baoxiong Jia, and Siyuan Huang. Learning unified force and position control for legged loco-manipulation, 2025. URL <https://arxiv.org/abs/2505.20829>.
- [43] Haochen Shi, Songbo Hu, Yifan Hou, Weizhuo Wang, Karen Liu, and Shuran Song. Minimalist compliance control. *arXiv preprint arXiv:2603.00913*, 2026.
- [44] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [45] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2017.
- [46] Bui Tuong Phong. Illumination for computer generated pictures. *Commun. ACM*, 18(6): 311–317, June 1975. ISSN 0001-0782. doi:10.1145/360825.360839. URL <https://doi.org/10.1145/360825.360839>.
- [47] Zhigen Zhao, Liuchuan Yu, Ke Jing, and Ning Yang. Xrobotoolkit: A cross-platform framework for robot teleoperation, 2025. URL <https://arxiv.org/abs/2508.00097>.

- 423 [48] Joao Pedro Araujo, Yanjie Ze, Pei Xu, Jiajun Wu, and C. Karen Liu. Retargeting matters:
424 General motion retargeting for humanoid motion tracking, 2025. URL [https://arxiv.org/](https://arxiv.org/abs/2510.02252)
425 [abs/2510.02252](https://arxiv.org/abs/2510.02252).
- 426 [49] PICO Immersive Pte. Ltd. PICO Motion Tracker. [https://www.picoxr.com/global/](https://www.picoxr.com/global/products/pico-motion-tracker)
427 [products/pico-motion-tracker](https://www.picoxr.com/global/products/pico-motion-tracker), 2023. Accessed: 2026-05-29.
- 428 [50] Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based
429 control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*,
430 pages 5026–5033, 2012. doi:10.1109/IROS.2012.6386109.

431 A Analytical whole-body reference controller

432 This appendix gives the full analytical compliance controller summarized in §3. It is used *only*
 433 as a reference that generates the per-link virtual targets $\mathbf{x}_{\ell,d}^c$ of Eq. (1) used in the policy reward;
 434 the reference torque $\boldsymbol{\tau}_c$ and the energy tank below are derivation/reference quantities and are *not*
 435 executed at deployment.

436 A humanoid has configuration $\mathbf{q} \in \text{SE}(3) \times \mathbb{R}^{n_j}$, generalized velocity $\boldsymbol{\nu}$, and floating-base dynamics
 437 $\mathbf{M}\dot{\boldsymbol{\nu}} + \mathbf{C}\boldsymbol{\nu} + \mathbf{g} = \mathbf{S}^\top \boldsymbol{\tau} + \sum_c \mathbf{J}_c^\top \boldsymbol{\lambda}_c + \sum_i \mathbf{J}_{\ell_i}^\top \mathbf{f}_{\text{ext},i}$ with support-contact reaction wrenches $\boldsymbol{\lambda}_c$
 438 and external wrenches $\mathbf{f}_{\text{ext},i}$ at links ℓ_i . Classical Cartesian impedance at a single end-effector [1]
 439 prescribes $\mathbf{M}_d \ddot{\mathbf{x}} + \mathbf{D}_d \dot{\mathbf{x}} + \mathbf{K}_d(\mathbf{x} - \mathbf{x}_d) = \mathbf{f}_{\text{ext}}$; we lift this to a hierarchy of frames coupled by
 440 balance.

441 **Reference compliant torque.** For controlled links $\mathcal{L}$ (hands, elbows, knees, pelvis, torso), support
 442 contacts $\mathcal{C}_s$, and centroidal momentum $\mathbf{h}$, the reference torque is

$$\boldsymbol{\tau}_c = \underbrace{\mathbf{J}_h^\top \mathbf{w}_h^{\text{imp}}}_{\text{centroidal balance}} + \underbrace{\sum_{\ell \in \mathcal{L}} \mathbf{N}_{\text{bal}}^\top \mathbf{J}_\ell^\top [\mathbf{K}_\ell(\mathbf{x}_{\ell,d} - \mathbf{x}_\ell) + \mathbf{D}_\ell(\dot{\mathbf{x}}_{\ell,d} - \dot{\mathbf{x}}_\ell)]}_{\text{per-link Cartesian impedance}} + \underbrace{\mathbf{N}_{\text{task}}^\top [\mathbf{K}_q(\mathbf{q} - \mathbf{q}_{\text{nom}}) + \mathbf{D}_q \dot{\mathbf{q}}]}_{\text{joint-space posture}} \quad (7)$$

443 where the centroidal-impedance wrench is $\mathbf{w}_h^{\text{imp}} = \mathbf{K}_h(\mathbf{h}_d - \mathbf{h}) + \mathbf{D}_h(\dot{\mathbf{h}}_d - \dot{\mathbf{h}}) + m\mathbf{g}$, $\mathbf{N}_{\text{bal}} =$
 444 $\mathbf{I} - \mathbf{J}_h^\top (\mathbf{J}_h \mathbf{M}^{-1} \mathbf{J}_h^\top)^{-1} \mathbf{J}_h \mathbf{M}^{-1}$ is the dynamically consistent null-space projector of the balance
 445 task [3], and $\mathbf{N}_{\text{task}}$ projects orthogonal to both balance and per-link tasks. Each per-link Cartesian
 446 wrench is first mapped to joint torque by $\mathbf{J}_\ell^\top$ and then projected by $\mathbf{N}_{\text{bal}}^\top$, so every term acts in the
 447 n_j -dimensional generalized-force space. The desired centroidal wrench is realized by the support
 448 contacts via a balancing QP in the spirit of [5]:

$$\min_{\{\boldsymbol{\lambda}_c\}} \left\| \mathbf{w}_h^{\text{imp}} - \sum_c \mathbf{G}_c \boldsymbol{\lambda}_c - \sum_i \mathbf{G}_{\text{ext},i} \mathbf{f}_{\text{ext},i} \right\|^2 + \gamma_{\text{reg}} \sum_c \|\boldsymbol{\lambda}_c\|^2 \quad \text{s.t.} \quad \boldsymbol{\lambda}_c \in \mathcal{F}_{\mu_c}, \text{CoP}_c \in \mathcal{S}_c \quad (8)$$

449 with grasp maps $\mathbf{G}_c$, linearized friction cones $\mathcal{F}_{\mu_c}$, support polygons $\mathcal{S}_c$, and regularizer weight γ_{reg} .
 450 Admitting knees into $\mathcal{C}_s$ adds support sites and generalizes the two-footed setting to multi-contact,
 451 lower-body-involved postures.

452 **Energy tank (general time-varying-stiffness case).** Time-varying $\mathbf{K}_\ell(t)$ with null-space projec-
 453 tion injects power and breaks passivity [27, 28]. For that general case one may add a scalar tank
 454 energy $T(t) \in [T_{\min}, T_{\max}]$,

$$\dot{T} = \sum_\ell \dot{\mathbf{x}}_\ell^\top \mathbf{D}_\ell \dot{\mathbf{x}}_\ell - P_{\text{inject}}(\dot{\mathbf{K}}_\ell, \mathbf{N}_{\text{bal}}, \mathbf{N}_{\text{task}}), \quad (9)$$

455 filled by dissipated power and drained by non-passive control, freezing gains as $T \rightarrow T_{\min}$. In our
 456 deployed pipeline $\mathbf{K}_\ell$ is held fixed and the residual edits only the equilibrium, so this tank is inert
 457 at runtime and is retained only as scaffolding for the time-varying-stiffness extension (cf. §F).

458 B Architecture details

459 **Base policy.** 3-layer MLP, hidden width 512, trained with PPO [44]. Observation $\mathbf{o}$ comprises
 460 joint configurations, velocities, projected gravity, the last action, and the motion command $\mathbf{c}$. The
 461 latent $\mathbf{z}_F$ enters as an additional input.

462 **Force Latent Encoder.** The encoder takes a history window $\mathbf{h}_t =$
 463 $(\mathbf{q}_{t-H:t}, \boldsymbol{\nu}_{t-H:t}, \boldsymbol{\tau}_{t-H:t-1}, \mathbf{c}_{t-H:t})$ of $H = 16$ steps (160 ms at 100 Hz), flattens it, and
 464 passes it through a 3-layer MLP (widths $512 \rightarrow 512 \rightarrow 256$, LayerNorm on the last hidden
 465 layer) emitting $(\boldsymbol{\mu}_t, \log \boldsymbol{\sigma}_t^2)$ over a latent of dimension $d_F = 32$. A reparameterized sample
 466 $\mathbf{z}_t = \boldsymbol{\mu}_t + \boldsymbol{\sigma}_t \odot \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, drives the wrench head $\psi : \mathbf{z}_F \rightarrow \hat{\mathbf{f}}_{\text{ext}} \in \mathbb{R}^{3|\mathcal{L}|}$ on all $|\mathcal{L}|$ controlled
 467 links and the projection head $g : \mathbf{z}_F \rightarrow \mathbb{S}^{d_g-1}$ with $d_g = 64$.

Auxiliary losses.

$$\mathcal{L}_{\text{wrench}} = \|\psi(\mathbf{z}_t) - \mathbf{f}_{\text{ext}}\|_2^2, \quad (10)$$

$$\mathcal{L}_{\text{supcon}} = \mathcal{L}^{\text{sup}}(g(\mathbf{z}_t), y_t), \quad (11)$$

$$\mathcal{L}_{\text{kl}} = \text{D}_{\text{KL}}(\mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t^2) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I})), \quad (12)$$

$$\mathcal{L}_{\text{smooth}} = \|\boldsymbol{\mu}_t - \boldsymbol{\mu}_{t-1}\|_2^2. \quad (13)$$

Weights: $(\lambda_w, \lambda_s, \lambda_k, \lambda_m) = (1.0, 0.3, 10^{-3}, 0.1)$. $\mathcal{L}^{\text{sup}}$ is the supervised contrastive loss of [21] with perturbation-class labels y_t indicating which limb is perturbed and in which direction.

Residual policy. 3-layer MLP, hidden width 256, trained with PPO against the same reward of Eq. (2). Output is bounded as in Eq. (4) with $\epsilon_x = 5$ cm; the per-link stiffness $\mathbf{K}_\ell$ is held fixed by the curriculum throughout Stage 2.

C Force sampling details

Phong-weighted force-origin sampling. The inverse-CDF expressions for the Phong lobe of Eq. (5) are

$$\cos \theta = u^{1/(n+1)}, \quad \phi \sim \mathcal{U}[0, 2\pi), \quad r = R_a v^{1/3}, \quad (14)$$

with $u, v \sim \mathcal{U}[0, 1]$, θ measured from $-\hat{\mathbf{e}}_z$. The cube-root radial warp yields uniform density in volume; with probability ϵ the direction is drawn isotropically instead.

The pelvis stiffness (k, c) is set so the spring force at maximum displacement R_a reaches a fixed fraction of the robot’s nominal weight: the stiffness term yields a coherent load that grows as the policy drifts from the anchor and admits a well-defined rest position against the leg stiffness, while the damping term reproduces the velocity-dependent tug of an elastically attached inertia. Per chunk we draw hold time $T_{\text{hold}} \sim \mathcal{U}(0.2, 2.0)$ s, ramp-down $\mathcal{U}(0.25, 1.0)$ s, and upper-body stiffness $K_p \sim \mathcal{U}(5, 250)$ N/m, with upper-body type sampled from {none, single-arm, both arms}; knees and feet are never perturbed directly. The shared downward bias of Eq. (5) makes the dominant carried-load regime coherent without an explicit pelvis-to-upper-body direction-inheritance heuristic.

Force law. Upper-body anchors apply a unilateral spring that auto-ramps as the body approaches and releases [11], while the pelvis is driven by a mass–spring–damper coupling to the sampled anchor to model a heavy object held through a compliant grasp. The two laws share the same anchor-based form and are gated on link type:

$$\mathbf{f}_\ell(t) = \begin{cases} k(\mathbf{a}_{\text{pelvis}}(t) - \mathbf{x}_{\text{pelvis}}(t)) - c \dot{\mathbf{x}}_{\text{pelvis}}(t), & \ell = \text{pelvis}, \\ K_p(\mathbf{a}_\ell - \mathbf{x}_\ell(t)) \mathbf{1}[\mathbf{a}_\ell \text{ on active side}], & \ell \in \mathcal{L}_{\text{upper}}. \end{cases} \quad (15)$$

D Algorithm

Reward. The trained objective is the compliance-fidelity reward of Eq. (2), where $\mathbf{x}_{\ell,d}^c$ is the analytical controller’s impedance-implied trajectory under the realized (ground-truth) perturbation, available in simulation. The term $w_c \|\cdot\|^2$ rewards following this compliant deformation rather than the original motion command—generalizing GentleHumanoid’s [11] reference-dynamics reward to the whole body. The encoder’s wrench-reconstruction MSE does *not* appear in r_t : it shapes ϕ only through $\mathcal{L}_\phi$ (Eq. (3)), behind the gradient barrier, so the policy reward never confuses “latent fits the wrench” with “robot behaves compliantly.”

Hyperparameters. $N_1 = 5 \times 10^8$ environment-steps for Stage 1; $N_2 = 5 \times 10^8$ environment-steps for Stage 2; 8192 parallel environments in MjWarp (MuJoCo-based [50]); PPO with clip 0.2, discount $\gamma = 0.99$, GAE $\lambda = 0.95$; control rate 100 Hz; reward weights $w_c = 0.5$, $w_e = 10^{-3}$; force-latent auxiliary weights $(\lambda_w, \lambda_s, \lambda_k, \lambda_m)$ as in Eq. (3).

Algorithm 1 COMPLIANTWBC training pipeline

Require: Motion dataset $\mathcal{M}$, force-sampling distribution $\mathcal{C}$ (§4.2), simulator $\mathcal{S}$

Stage 1: Force-aware base policy (PPO) + Force Latent Encoder (aux).

- 1: Initialize π_{base} , encoder ϕ , wrench head ψ , projection head g randomly.
- 2: **for** $i = 1, \dots, N_1$ **do**
- 3: Sample $(m, \mathbf{f}_{\text{ext}}, y, \mathcal{C}_s) \sim \mathcal{M} \times \mathcal{C}$ (y : perturbation class).
- 4: **for** $t = 1, \dots, T$ **do**
- 5: $(\boldsymbol{\mu}_t, \log \sigma_t^2) \leftarrow \phi(\mathbf{q}_{t-H:t}, \boldsymbol{\nu}_{t-H:t}, \boldsymbol{\tau}_{t-H:t-1}, \mathbf{c}_{t-H:t}); \mathbf{z}_F \leftarrow \boldsymbol{\mu}_t.\text{detach}()$.
- 6: $\boldsymbol{\tau}_t \leftarrow \pi_{\text{base}}(\mathbf{o}_t, \mathbf{c}_t, \mathbf{z}_F)$; step $\mathcal{S}$.
- 7: Roll the analytical controller on the same state to obtain $\mathbf{x}_{\ell,d}^c$; compute reward (2).
- 8: **end for**
- 9: PPO update on π_{base} [44].
- 10: Aux update on (ϕ, ψ, g) : $\nabla \mathcal{L}_\phi$ from Eq. (3) using $(\mathbf{f}_{\text{ext}}, y)$.
- 11: **end for**

Stage 2: Residual on impedance targets (PPO).

- 12: Freeze $\pi_{\text{base}}, \phi, \psi$. Initialize the residual policy π_{res} randomly.
- 13: **for** $i = 1, \dots, N_2$ **do**
- 14: Sample $(m, \mathbf{f}_{\text{ext}}, \mathcal{C}_s) \sim \mathcal{M} \times \mathcal{C}$.
- 15: **for** $t = 1, \dots, T$ **do**
- 16: $\mathbf{z}_F \leftarrow \phi(\cdot); \hat{\mathbf{f}}_{\text{ext}} \leftarrow \psi(\mathbf{z}_F)$.
- 17: Form analytical target $\mathbf{x}_{\ell,d}$ via Eq. (1) using $\hat{\mathbf{f}}_{\text{ext}}$.
- 18: $\Delta \mathbf{x}_{\ell,d} \leftarrow \pi_{\text{res}}(\mathbf{o}_t, \mathbf{c}_t, \mathbf{z}_F, \hat{\mathbf{f}}_{\text{ext}})$.
- 19: Compose $\tilde{\mathbf{x}}_{\ell,d}$ via (4); $\mathbf{K}_\ell$ held fixed by curriculum.
- 20: $\boldsymbol{\tau}_t \leftarrow \pi_{\text{base}}(\mathbf{o}_t, \mathbf{c}_t, \mathbf{z}_F)$ executed against $(\tilde{\mathbf{x}}_{\ell,d}, \mathbf{K}_\ell)$; step $\mathcal{S}$; compute reward (2).
- 21: **end for**
- 22: PPO update on π_{res} [44].
- 23: **end for**
- 24: **return** $\pi_{\text{base}}, \phi, \psi, \pi_{\text{res}}$.

E Metrics Formulation

Task success. A trial is successful if the robot remains valid throughout the force perturbation roll-out without triggering the evaluation termination conditions. In our evaluation, these terminations are restricted to fall/instability events: root linear velocity exceeding 5 m/s, root height dropping below 0.5 m, or root height rising above 1.0 m. Let $f_i \in \{0, 1\}$ denote whether any such termination is triggered during trial i . The success indicator is

$$s_i = \mathbf{1}[f_i = 0], \quad (16)$$

and the reported success rate is

$$S = \frac{1}{N} \sum_{i=1}^N s_i. \quad (17)$$

Command tracking error. For the commanded body frames $\mathcal{K}_{\text{cmd}}$, currently the left and right wrist frames, we measure Cartesian tracking error relative to the reference motion. Let $\mathcal{T}_{\text{free}}$ denote timesteps for which the external-force envelope is inactive, and let $\mathcal{T}_{\text{force}}$ denote timesteps for which it is active, i.e., the normalized applied-force envelope $a_t > 0.05$ (the trapezoid amplitude of §4.2). For a time set $\mathcal{T}$, the command tracking RMSE is

$$E_{\text{cmd}}(\mathcal{T}) = \sqrt{\frac{1}{|\mathcal{T}| |\mathcal{K}_{\text{cmd}}|} \sum_{t \in \mathcal{T}} \sum_{k \in \mathcal{K}_{\text{cmd}}} \|\mathbf{p}_{k,t} - \mathbf{p}_{k,t}^{\text{ref}}\|_2^2}. \quad (18)$$

We report the free-space error $E_{\text{cmd}}^{\text{free}} = E_{\text{cmd}}(\mathcal{T}_{\text{free}})$ from the no-force pass and the force-window error $E_{\text{cmd}}^{\text{force}} = E_{\text{cmd}}(\mathcal{T}_{\text{force}})$ over the active perturbation window, which quantifies tracking relaxation under external contact.

Table 3: Full reward bundle for COMPLIANTWBC (both stages), grouped by purpose with per-term weights; the compliance-fidelity term of Eq. (2) corresponds to the *Compliance* group.

Term	Form	Weight
<i>Whole-body tracking</i>		
tracking_joint_dof	$\exp(-0.15 \sum_j (q_j - q_{j,d})^2)$	+2.0
tracking_joint_vel	$\exp(-0.01 \sum_j (\dot{q}_j - \dot{q}_{j,d})^2)$	+0.2
tracking_root_rotation	$\exp(-5 \theta_q^2)$	+1.0
tracking_root_linear_vel	$\exp(-\ \mathbf{v}_b - \mathbf{v}_{b,d}\ ^2)$	+1.0
tracking_root_angular_vel	$\exp(-\ \boldsymbol{\omega}_b - \boldsymbol{\omega}_{b,d}\ ^2)$	+1.0
tracking_keybody_pos	$\exp(-10 \sum_k \ \Delta_b^k - \Delta_{b,d}^k\ ^2)$	+2.0
tracking_keybody_pos_global	$\exp(-10 \sum_k \ \mathbf{x}_k - \mathbf{x}_{k,d}\ ^2)$	+0.1
<i>Compliance — admittance force & virtual target</i>		
force_reward	$\exp(-0.0625 \langle \ \mathbf{F}_\ell^{\text{app}} - \mathbf{F}_\ell^{\text{exp}}\ \rangle_\ell) \mathbb{K}[\text{not exc.}]$	+2.0
force_target_tracking	$\exp(-25 \langle \ \mathbf{x}_\ell - \tilde{\mathbf{x}}_\ell\ \rangle_\ell^2)$	+2.0
force_target_vel_tracking	$\exp(-\langle \ \dot{\mathbf{x}}_\ell - \dot{\tilde{\mathbf{x}}}_\ell\ \rangle_\ell^2)$	+1.0
force_ext_penalty	$\langle \mathbb{K}[\ \mathbf{F}_\ell^{\text{app}}\ > F_{\text{max},\ell} \wedge \ \mathbf{F}_\ell^{\text{app}}\ > \ \mathbf{F}_\ell^{\text{exp}}\ + \frac{\delta}{2}] \rangle_\ell$	-6.0
keypoint_tracking_imp	$\exp(-10 \sum_{k \in \mathcal{K}} \ \mathbf{x}_k - \tilde{\mathbf{x}}_k\ ^2), \tilde{\mathbf{x}}_k = \mathbf{x}_\ell^{\text{virt}} \text{ on } \mathcal{L}$	+2.0
<i>Lower-body yielding under load</i>		
lower_keypoint_tracking	$\exp(-10 \sum_{k \in \mathcal{K}_{\text{lo}}} \ \mathbf{x}_k - \tilde{\mathbf{x}}_{k,\text{lo}}\ ^2)$	+2.0
tracking_compliant_root_translation_z	$\exp(-5 (z_{\text{root}} - z_{\text{virt}})^2)$	+1.0
squat_reward	$\exp(-25 (\Delta z - c_z \sum_\ell \ \mathbf{F}_\ell\)^2), c_z = 3 \times 10^{-4} \text{ m/N}$	+1.0
<i>Safety — joint, torque, contact, collision</i>		
alive	$\mathbb{K}[\text{alive}]$	+0.5
dof_pos_limits	$\sum_j [\text{ReLU}(q_j - q_j^{\text{max}}) + \text{ReLU}(q_j^{\text{min}} - q_j)]$	-5.0
joint_limit	same form as dof_pos_limits (additional)	-10.0
dof_torque_limits	$\sum_j \text{ReLU}(\tau_j /\tau_{j,\text{max}} - 0.95)$	-1.0
self_collisions	$\mathbb{K}[\ \mathbf{F}^{\text{self}}\ > 10 \text{ N}]$ summed over contact pts	-10.0
feet_stumble	$\mathbb{K}[\exists f : \ \mathbf{F}_f^{xy}\ > 4 F_f^z]$	-1.25
feet_contact_forces	$\sum_f \text{ReLU}(F_f^z - 500 \text{ N})$	-5×10^{-4}
feet_slip	$\sum_f \sqrt{\ \dot{\mathbf{x}}_f^{xy}\ } \mathbb{K}[F_f^z > 5]$	-0.1
<i>Regularization — smoothness & locomotion</i>		
dof_vel	$\sum_j \dot{q}_j^2$	-10^{-4}
dof_acc	$\sum_j \ddot{q}_j^2$	-5×10^{-8}
ankle_dof_vel	$\sum_{j \in \mathcal{J}_{\text{ank}}} \dot{q}_j^2$	-2×10^{-4}
ankle_dof_acc	$\sum_{j \in \mathcal{J}_{\text{ank}}} \ddot{q}_j^2$	-10^{-7}
action_rate_12	$\ \mathbf{a}_t - \mathbf{a}_{t-1}\ ^2$	-0.1
ang_vel_xy	$\omega_{b,x}^2 + \omega_{b,y}^2$	-0.01
feet_air_time	$\sum_f \min(0, t_f^{\text{air}} - 0.5\text{s}) \mathbb{K}[\text{first contact}] \mathbb{K}[\ \mathbf{v}_{b,d}^{xy}\ > 0.05]$	+5.0

517 **Torque reaction.** We report the fraction of near-saturated joints, ρ_τ :

$$\rho_\tau = \frac{1}{T n_j} \sum_{t=1}^T \sum_{j=1}^{n_j} \mathbf{1}[|\tau_{j,t}| > 0.9 \tau_j^{\text{max}}], \quad (19)$$

518 where τ_j^{max} is the torque limit of joint j . A stiff controller fights the force (high ρ_τ); a compliant
519 one yields, keeping it low.

520 **Lower-body participation.** To verify that the controller actually uses the lower body under multi-
521 site contacts, we compare the torque deviation from a matched no-perturbation rollout:

$$R_{\text{LB}} = \frac{\sum_t \sum_{j \in \mathcal{J}_{\text{lower}}} |\tau_{j,t} - \tau_{j,t}^0|}{\sum_t \sum_{j \in \mathcal{J}_{\text{all}}} |\tau_{j,t} - \tau_{j,t}^0| + \epsilon}. \quad (20)$$

522 Here τ^0 is the torque sequence produced by the same controller on the same command trajectory
523 without external perturbation. A high R_{LB} during hip, pelvis, torso, or knee perturbations indicates
524 that the policy redistributes effort through the legs instead of treating force response as an upper-
525 body-only problem.

526 F Local passivity by construction

527 **Proposition 1** (Bounded interaction force and local passivity). *Let the closed-loop system satisfy:*
528 *(i) the base policy realizes τ_c of Eq. (7) at the equilibrium $(\mathbf{q}^*, \mathbf{0})$; (ii) the residual output is bounded*

529 as in Eq. (4); (iii) the per-link stiffness $\mathbf{K}_\ell$ is constant in time. Then (a) the residual shifts each
 530 link equilibrium by at most ϵ_x , so the induced restoring force is bounded by $\|\mathbf{K}_\ell\| \epsilon_x$; and (b) if
 531 additionally (iv) the residual-edited equilibrium $\tilde{\mathbf{x}}_{\ell,d}$ varies slowly enough that its power injection is
 532 dominated by the link damping $\mathbf{D}_\ell$, the storage function $S = \frac{1}{2} \boldsymbol{\nu}^\top \mathbf{M} \boldsymbol{\nu} + \frac{1}{2} \sum_\ell (\mathbf{x}_\ell - \tilde{\mathbf{x}}_{\ell,d})^\top \mathbf{K}_\ell (\mathbf{x}_\ell -$
 533 $\tilde{\mathbf{x}}_{\ell,d})$ satisfies $\dot{S} \leq \sum_i \mathbf{f}_{\text{ext},i}^\top \dot{\mathbf{x}}_{\text{ext},i}$, where $\mathbf{x}_{\text{ext},i}$ is the displacement of contact point i ; i.e. the closed
 534 loop is passive with respect to the external-wrench port.

535 *Sketch.* Bound (a) is immediate from Eq. (4): the residual moves the equilibrium by at most ϵ_x , so
 536 with fixed $\mathbf{K}_\ell$ the induced restoring force changes by at most $\|\mathbf{K}_\ell\| \epsilon_x$. For (b), differentiating S and
 537 substituting the closed-loop dynamics gives $\dot{S} = \sum_i \mathbf{f}_{\text{ext},i}^\top \dot{\mathbf{x}}_{\text{ext},i} - (\text{damping dissipation}) - \sum_\ell (\mathbf{x}_\ell -$
 538 $\tilde{\mathbf{x}}_{\ell,d})^\top \mathbf{K}_\ell \dot{\tilde{\mathbf{x}}}_{\ell,d}$. With $\mathbf{K}_\ell$ constant the $\dot{\tilde{\mathbf{x}}}_{\ell,d}$ term that would otherwise inject non-passive power van-
 539 ishes identically; the remaining moving-equilibrium term is dominated by the damping dissipation
 540 under assumption (iv), yielding the stated inequality. For the general time-varying-stiffness case the
 541 energy tank of §A would absorb the $\dot{\tilde{\mathbf{x}}}_{\ell,d}$ term; in our pipeline stiffness is fixed and the residual does
 542 not modulate gains, so the tank is reference scaffolding rather than a runtime mechanism.

543 **Scope and what this does and does not buy.** The result is *local* (around the equilibrium) and
 544 *by construction*: it follows from the residual bound and fixed stiffness, not from the learned com-
 545 ponents. Global stability under adversarial external forces is not claimed and remains open. Two
 546 consequences are worth stating explicitly. (i) Action-residual baselines (ASAP, ResMimic) edit the
 547 torque directly and do not admit this storage function; their effect on induced contact force is not
 548 physically bounded by construction. (ii) The proposition is a safety property for human–robot con-
 549 tact, not a performance property: it does not guarantee tracking, success rate, or sim-to-real transfer.
 550 Those are settled empirically (§5).