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

Anonymous Authors

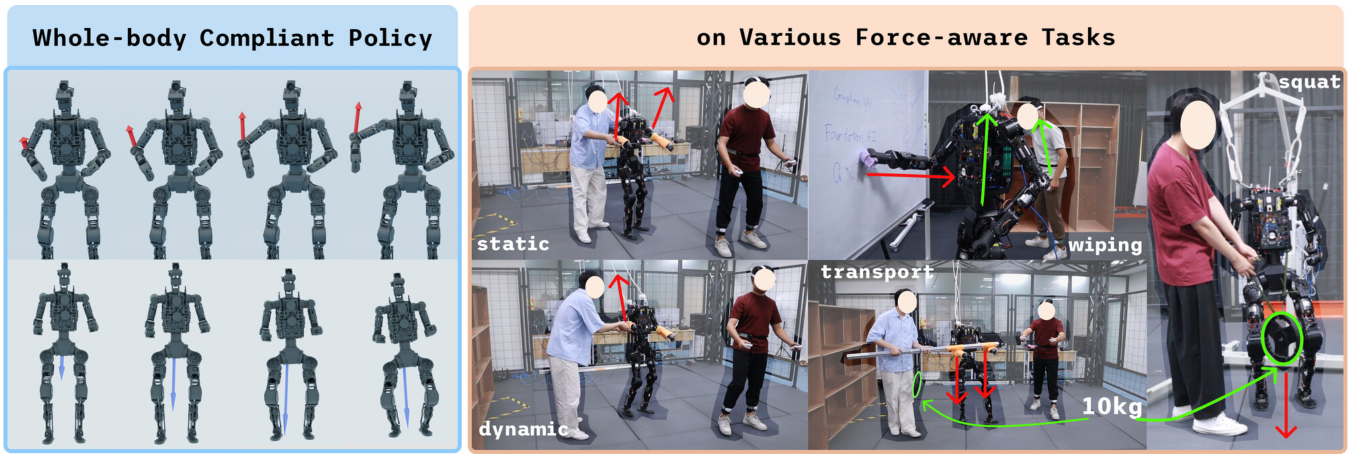

Fig. 1: COMPLIANTWBC pairs a force-aware base policy with a bounded residual on the impedance equilibrium, trained under a Phong-weighted force-origin sampler that extends compliance to the whole body. We demonstrate it on a range of force-aware tasks on a real heavy humanoid.

***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 *base policy* trained with RL to maximize compliance-fidelity reward, guided by a multi-site whole-body impedance *reference controller*, extending classical Cartesian impedance to any controlled link; (2) A bounded *residual policy* that edits the per-link impedance equilibrium over a frozen base, correcting the coarse but structured wrench estimate supplied by a *force encoder* co-trained behind a gradient barrier; (3) A *Phong-weighted force-origin sampler* with an axis-decoupled pelvis anchor induces lower-body-inclusive compliance curriculum training via two interpretable parameters. We evaluate COMPLIANTWBC in simulation against both compliant and stiff baselines, achieving best compliant fidelity of 2.58cm deviation from analytical solutions, 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/**

I. 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]–[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, previous works restrict compliance to the end-effector or to the upper-body chain [8]–[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 $\boldsymbol{f}_{\text{ext}}$ is unobserved on hardware) and must output torques that *yield* compliantly to $\boldsymbol{f}_{\text{ext}}$ at arbitrary contact sites while sustaining balance and command tracking. Transferring classical compliance into a learning pipeline with reinforcement learning (RL) therefore faces two main 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 appropriate force perturbations, and the learned policy must *encode* compliance up to a reasonable force tolerance limit rather than override it aggressively to maintain good tracking. 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]–[19] expose no impedance interface during training, so compliance must be injected separately while running the full analytical pipeline at deployment. This leads to significantly higher 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 (Section III) 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 can avoid the QP's runtime cost and contact-mode fragility. For the second, a Phong-weighted force-origin sampler (Section IV-B) 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. To better introduce external force signal to the policy, we jointly train a variational *force encoder* with the base policy behind a gradient barrier from PPO and shaped by wrench reconstruction, contrastive clustering [21], a KL bottleneck, and temporal smoothness. This *force encoder* supplies a class-clustered, wrench-grounded estimate of the unobserved contact state [22] and serves as a coarse yet structured signal for the policy. Finally, based on this signal, a bounded residual on the frozen base edits the impedance *equilibrium* rather than motor commands or gains [23], [24]. We study these behaviors on a heavy humanoid platform (*70 kg*, versus the *35 kg* Unitree G1 [25] used in most prior learning-based compliant humanoid work) to tackle a wider range of contact-rich tasks. Our contributions can be summarized as follows:

- An **analytical multi-site whole-body compliance reference controller** (Section III) composing centroidal-momentum balance with per-link Cartesian impedance at arbitrary sites, whose virtual targets guide the compliance-fidelity reward of the base policy.
- A **bounded residual on the per-link impedance equilibrium** (Section IV-A) over a frozen base, which corrects the direction- and magnitude-structured error of a proprioceptive wrench estimate, recovering most of the compliance fidelity of an oracle wrench.
- A **Phong-weighted force-origin sampler with an axis-decoupled pelvis anchor** (Section IV-B) that induces multi-contact, lower-body-inclusive compliance to our policy, evaluated extensively in simulation and demonstrated on a real heavy humanoid (Section V).

II. Related Work

A. Compliance Control for Humanoids

Classical compliance strategies, including task-space impedance and admittance control [1]–[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 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 [26], [27]. Recent whole-body tracking controllers [28]–[30] may achieve agile loco-manipulation via motion imitation but do not expose an impedance interface, so compliance must be injected separately.

TABLE I: Humanoid whole-body-control landscape. *LB*: lower-body compliance (✓), quadruped only (quad.), or none (–). *Sites*: where compliance is realized—end-effector (EE), upper body, CoM reference, or arbitrary links (whole). *Residual*: where a learned correction is applied.

Method	LB	Sites	Residual	Platform
FALCON [8]	–	EE	–	G1/Booster
GentleHumanoid [11]	–	upper	–	G1
FACET [10]	quad.	CoM	impedance ref.	Go2/G1
SoftMimic [12]	per-motion	whole	–	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

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]. 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. Some [12] manage to extend the compliant behavior to the lower body, but limits to a single motion and requires complicated per-motion augmentation to implicitly introduce desired interaction rather than targeting an arbitrary robot configuration within its workspace. A parallel line targets heavy-payload loco-manipulation with multi-policy or trajectory-optimized references [31], [32] 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 I).

B. Residual Policies and Latent Force Estimation

Residual reinforcement learning [23], [33], [34] 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. [35], 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 [9], [13], [14]. 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 [21], [36]–[39]. Recent work specializes such latents to estimate external wrenches from proprioception without dedicated force sensors [40]–[42]. We adopt this proprioceptive estimation approach rather than improve on it: the estimate it supplies is deliberately coarse, and our contribution is the *bounded residual on the per-link impedance equilibrium* that it guides. Because the link stiffness $\boldsymbol{K}_\ell$ is held fixed, the residual's effect is upper-bounded by an explicit virtual-target shift times the stiffness.

III. Whole-Body Impedance Heuristic

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

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 $\boldsymbol{f}_{\text{ext}}$ applied at body site p,

$$\boldsymbol{x}_\ell(\boldsymbol{f}_\ell) = \boldsymbol{x}_\ell^{\text{ref}} + \boldsymbol{K}_\ell^{-1}\, \boldsymbol{f}_\ell, \qquad \boldsymbol{f}_\ell = \mathrm{Ad}_{p\to\ell}^{-\top}\, \boldsymbol{f}_{\text{ext}}, \tag{1}$$

where $\boldsymbol{f}_\ell$ is the external wrench transported to link ℓ's frame and $\boldsymbol{K}_\ell^{-1}$ is the anisotropic Cartesian compliance, so the correction has units of displacement. We use the argument throughout to mark a *commanded* target and to name the wrench it was built from: $\boldsymbol{x}_\ell(\boldsymbol{f})$ uses the ground-truth wrench and is the reward anchor, $\boldsymbol{x}_\ell(\hat{\boldsymbol{f}})$ uses the estimate $\hat{\boldsymbol{f}}_{\text{ext}} = \psi(\boldsymbol{z}_F)$ and is what the policy is commanded with, and $\boldsymbol{x}_\ell(\boldsymbol{0}) = \boldsymbol{x}_\ell^{\text{ref}}$ is the nominal reference. Written without an argument, $\boldsymbol{x}_\ell$ is always the *realized* link position. This generalizes GentleHumanoid's scalar spring to arbitrary sites and anisotropic stiffness: restricting $\mathcal{L}$ to upper-body links with $\boldsymbol{K}_\ell = K_p\boldsymbol{I}_3,\ K_p \sim \mathcal{U}(5, 250)$ N/m and $\boldsymbol{D}_\ell = 2\sqrt{m_v K_p}\boldsymbol{I}_3$ ($m_v = 0.1$ 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.

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

A. Pipeline Architecture

The analytical controller of Section III assumes rigid contacts, perfect torque control, accurate dynamics, and a measured external wrench, none of which a deployed policy on a heavy humanoid can rely on in the rewal world. We therefore train a two-stage pipeline (Figure 2): (i) a *force-aware base policy* that jointly trains with a force 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 estimation and the true external wrench.

Base policy is a force-aware whole-body policy $\pi_{\text{base}}(\mathbf{prop}, \boldsymbol{q}_d, \{\boldsymbol{x}_\ell(\boldsymbol{0})\}, \hat{\boldsymbol{f}}_{\text{ext}}) \to \boldsymbol{q}^{\text{des}}$ that takes as input the proprioception $\mathbf{prop}$, the retargeted joint targets $\boldsymbol{q}_d$, the *nominal* per-link targets $\{\boldsymbol{x}_\ell(\boldsymbol{0})\}_{\ell\in\mathcal{L}}$, and the wrench estimate $\hat{\boldsymbol{f}}_{\text{ext}}$ decoded from the force encoder, and outputs *joint position targets* $\boldsymbol{q}^{\text{des}}$. A fixed-gain joint-space PD law $\boldsymbol{\tau} = \boldsymbol{K}_p(\boldsymbol{q}^{\text{des}} - \boldsymbol{q}) - \boldsymbol{D}_p\dot{\boldsymbol{q}}$ is the only torque law executed online; the analytical controller of Section III and its balancing QP never run at deployment. The policy is trained end-to-end with PPO [43] against a compliance-fidelity reward

$$r_t = r_{\text{track}} \;-\; w_c\,\|\boldsymbol{x}_\ell - \boldsymbol{x}_\ell(\boldsymbol{f})\|^2 \;-\; w_e\,\|\boldsymbol{\tau}\|^2, \tag{2}$$

whose term $w_c\|\cdot\|^2$ drives the controlled links toward the analytical controller's virtual-target deformation $\boldsymbol{x}_\ell(\boldsymbol{f})$ (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.

Force encoder follows a variational encoder architecture that maps an observation history window (proprioceptions, velocities, last torques, and motion command) to a force latent $\boldsymbol{z}_F \in \mathbb{R}^{d_F}$. 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 $\hat{\boldsymbol{f}}_{\text{ext}} = \psi(\boldsymbol{z}_F)$ 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 $\boldsymbol{z}_F$ in the true contact wrench through ψ; a supervised contrastive loss [21] over perturbation-class labels (which limb, which direction); a KL/VIB bottleneck [44] suppressing dimensions that carry neither wrench- nor class-relevant information; and a temporal-smoothness term on consecutive posterior means:

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

Together they make $\boldsymbol{z}_F$ a wrench-grounded, class-clustered, low-bandwidth embedding of the privileged contact state, which provides a coarse yet structured signal for both the base policy and the residual policy.

Residual policy is a small MLP added on top of the frozen base policy and frozen force encoder. It reads $(\mathbf{prop}_t, \boldsymbol{q}_d, \{\boldsymbol{x}_\ell(\boldsymbol{0})\}, \hat{\boldsymbol{f}}_{\text{ext}})$, with the frozen wrench estimate $\hat{\boldsymbol{f}}_{\text{ext}} = \psi(\boldsymbol{z}_F)$, and outputs a bounded delta on the impedance

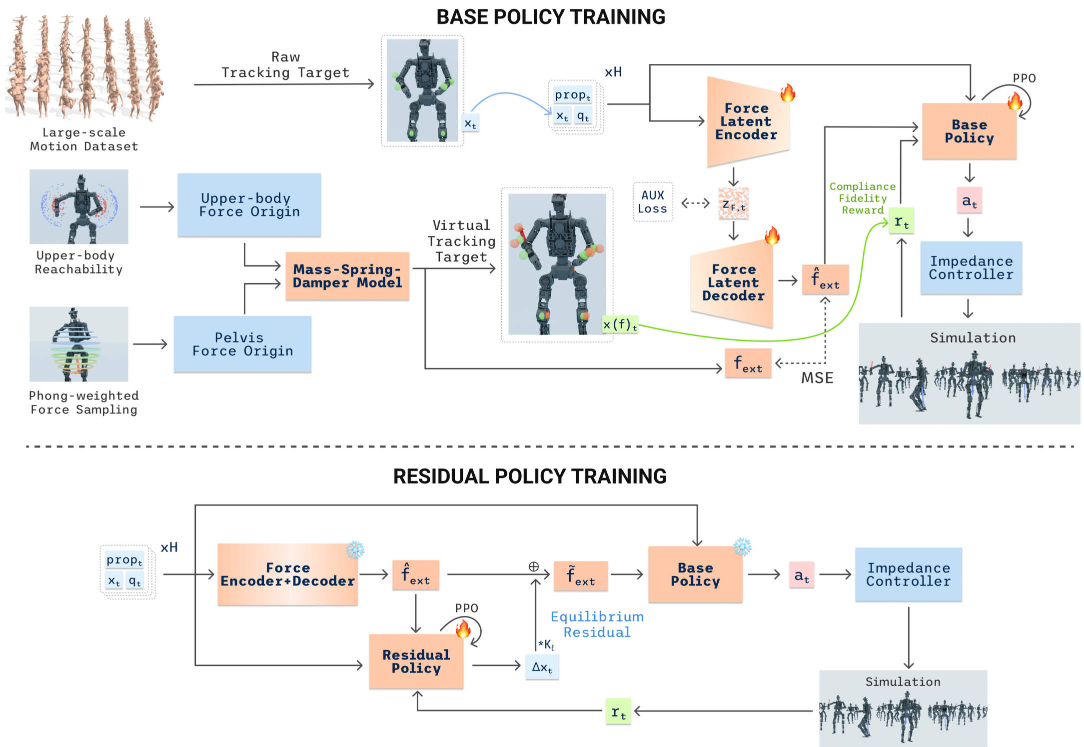

Fig. 2: Two-stage training of COMPLIANTWBC: (i) a base policy jointly trained with the force encoder to induce whole-body compliance; (ii) a bounded residual on the impedance equilibrium that compensates for wrench-estimation error.

equilibrium. The deployed control graph is then

$$\tilde{\boldsymbol{x}}_\ell(\hat{\boldsymbol{f}}) = \boldsymbol{x}_\ell(\hat{\boldsymbol{f}}) + \Delta\boldsymbol{x}_\ell, \quad \Delta\boldsymbol{x}_\ell = \epsilon_x \tanh(\boldsymbol{u}_t), \tag{4}$$

$$\tilde{\boldsymbol{f}}_\ell = \hat{\boldsymbol{f}}_\ell + \boldsymbol{K}_\ell \Delta\boldsymbol{x}_\ell, \tag{5}$$

$$\boldsymbol{q}_t^{\text{des}} = \pi_{\text{base}}(\textbf{prop}_t, \boldsymbol{q}_d, \{\boldsymbol{x}_\ell(\boldsymbol{0})\}, \tilde{\boldsymbol{f}}), \tag{6}$$

$$\boldsymbol{\tau}_t = \boldsymbol{K}_p(\boldsymbol{q}_t^{\text{des}} - \boldsymbol{q}_t) - \boldsymbol{D}_p\dot{\boldsymbol{q}}_t, \tag{7}$$

where $\boldsymbol{u}_t = \pi_{\text{res}}(\cdot)$ is the raw residual output and $\Delta\boldsymbol{x}_\ell$ the applied edit with $\|\Delta\boldsymbol{x}_\ell\| \le \epsilon_x = 5$ cm. Because Equation (1) is affine in the wrench, $\boldsymbol{x}_\ell(\hat{\boldsymbol{f}} + \boldsymbol{K}_\ell\Delta\boldsymbol{x}_\ell) = \tilde{\boldsymbol{x}}_\ell(\hat{\boldsymbol{f}})$, the equilibrium edit of Equation (4) and the wrench correction of Equation (5) are the same residual in two coordinates and therefore can be defined on the equilibrium and transmitted on $\hat{\boldsymbol{f}}_{\text{ext}}$ as policy input.

When $\hat{\boldsymbol{f}}_{\text{ext}}$ 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. Because $\tilde{\boldsymbol{f}}$ occupies the same input slot π_{base} was trained on in Stage 1, the edit also stays on the frozen policy's input manifold. Both stages are rewarded against $\boldsymbol{x}_\ell(\boldsymbol{f})$, the analytical target under the *ground-truth* wrench and the reward anchor is never the edited $\tilde{\boldsymbol{x}}_\ell(\hat{\boldsymbol{f}})$, so π_{res} cannot earn reward by moving its own target. We train the residual in a separate stage rather than jointly with the base, as under a single PPO loss the residual would absorb compliance the base could itself learn, which would turn its bound into a constraint on the joint system rather than serve an interpretable discrepancy compensator as in our framework.

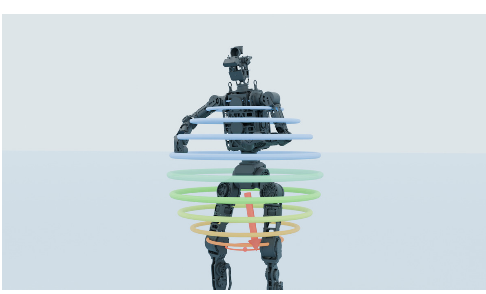

Fig. 3: Phong-weighted sampling generates downward-biased force origins, decoupled by axis.

B. 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 $\hat{\boldsymbol{n}}$ following a Phong lobe [45] concentrated on the downward axis $-\hat{\boldsymbol{e}}_z$ with exponent $n = 2$:

$$p(\hat{\boldsymbol{n}}) = (1-\epsilon)\,\frac{n+1}{2\pi}\,\max\!\big(-\hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{e}}_z,\,0\big)^n + \epsilon\,\frac{1}{4\pi}, \tag{8}$$

with isotropic mixing weight $\epsilon = 0.1$, sampled in closed form by inverse CDF to give the offset $\boldsymbol{\delta} = r(\sin\theta\cos\phi,\, \sin\theta\sin\phi,\, -\cos\theta)$. 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 $\boldsymbol{\delta}$ 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 $\boldsymbol{a}_{\text{pelvis}}$:

$$\boldsymbol{a}_{\text{pelvis}}(t) = \underbrace{\boldsymbol{x}^{xy}_{\text{pelvis}} + \boldsymbol{\delta}^{xy}}_{\text{base-relative (horizontal)}} + \underbrace{\left(h^* + \delta^z\right)\hat{\boldsymbol{e}}_z}_{\text{gravity-anchored (vertical)}}, \quad (9)$$

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}: 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.

V. Experiments

Evaluation goal. We ask three questions: (i) does *whole-body* compliance from multi-site (wrist, elbow, torso, pelvis, hip, knee) force sampling improve over aggressive tracking and end-effector-only compliance, especially for forces away from the hands? (ii) how accurate is the learned wrench estimate, and what does the bounded residual on virtual targets actually correct? and (iii) how does CompliantWBC behave in the real world on contact-rich tasks? We address (i)–(ii) in Section V-B and (iii) in Section V-C.

A. Settings and Metrics

We compare against two external baselines—a stiff whole-body tracker **TWIST2** [15] (nominal command $\boldsymbol{x}^{\text{ref}}$, 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 encoder, no Stage-2 residual). **Ours (full)** adds the bounded residual on virtual targets with per-link stiffness fixed by the curriculum. All variants share an identical training recipe, environment budget, and paired random seeds.

Metrics. Apart from comparing **Task success** S and **Tracking performance** $E^{\text{free}}_{\text{cmd}}$, we also report metrics that characterize the compliant response to 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. We further report **Compliance fidelity** E_{imp}, the distance between the realized pose of the *perturbed* link and the analytical impedance response anchored to the *ground-truth* wrench. As the compliant controller deviates from the original tracking target under external forces, E_{imp} compares how close the robot to the analytical controller induced targets so policies that are too stiff or too soft both score badly.

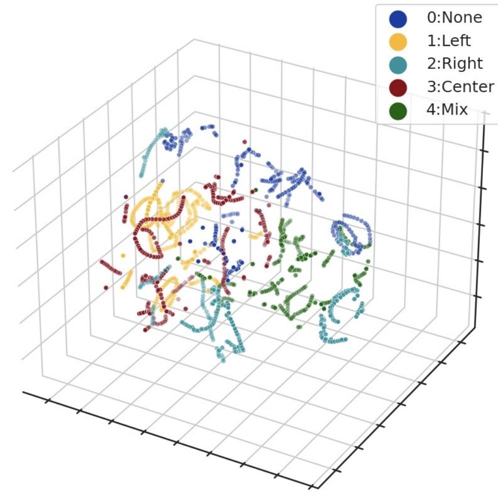

Fig. 4: Force latent embedding projections via t-SNE of different force profiles across contact sites.

B. Simulation Experiments

How compliant is the policy? Table II lays out the tracking–compliance trade-off. TWIST2 tracks best in free space but succeeds on only 59% of trials, fighting every perturbation. This is also demonstrated by the worst impedance fidelity score $E_{\text{imp}} = 6.79$. 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. It also attains the best compliance fidelity ($E_{\text{imp}} = 2.58$, against ≥ 3.81 for every baseline and ablation). 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 highlight the distinct behavior in the lower body between the whole-body compliant and non-compliant policy under a pelvis perturbation in Figure 5. The compliant policy keeps lower-body torques much smaller and stays stable through

TABLE II: Performance of COMPLIANTWBC versus both *compliant* and *non-compliant* baselines over 100 paired rollouts. Arrows indicate the preferred direction; bracketed columns are in units of 10^{-2}. Metrics are defined in Section V-A.

Controller	Tracking quality		WB compliance and robustness		
	$E_{\text{cmd}}^{\text{free}}\downarrow$ [$\times10^{-2}$]	$E_{\text{imp}}\downarrow$ [$\times10^{-2}$]	$\rho_\tau\downarrow$ [$\times10^{-2}$]	$R_{\text{LB}}\uparrow$	$S\uparrow$
TWIST2	$\mathbf{2.18 \pm 0.31}$	6.79 ± 0.72	2.74 ± 0.38	0.16 ± 0.04	0.59
GENTLEHUMANOID	5.22 ± 0.47	3.92 ± 0.35	2.12 ± 0.29	0.14 ± 0.03	0.91
OURS w/o pelvis force sampling	4.12 ± 0.38	4.22 ± 0.21	1.87 ± 0.24	0.16 ± 0.03	0.93
OURS w/o residual policy	4.29 ± 0.35	3.81 ± 0.29	1.56 ± 0.21	0.20 ± 0.03	0.94
OURS (full)	4.65 ± 0.41	$\mathbf{2.58 \pm 0.24}$	$\mathbf{0.93 \pm 0.15}$	$\mathbf{0.31 \pm 0.04}$	**0.98**

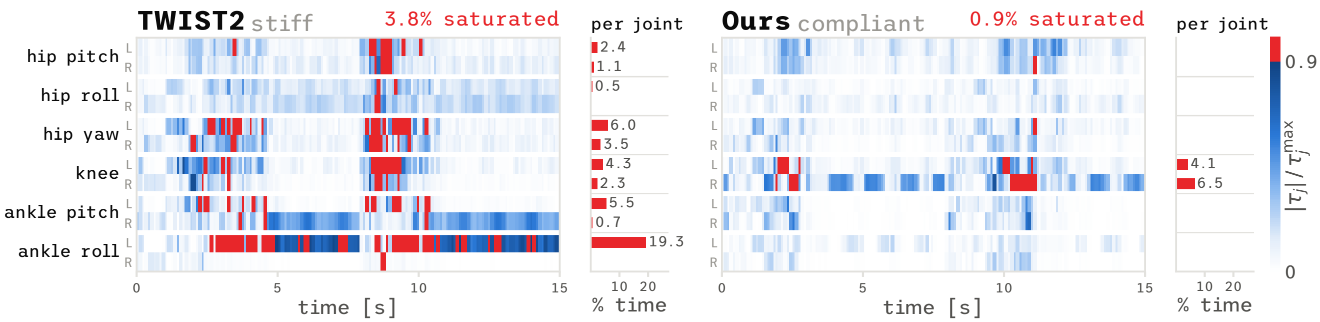

Fig. 5: Lower-body joint effort $|\tau_j|/\tau_j^{\max}$ under two trapezoidal pelvis pushes (up 1–2 s and 9–10 s, down 7–8 s); L/R rows per joint, 0.1 s bins. Red: above $0.9\,\tau_j^{\max}$; bars: fraction of time above it.

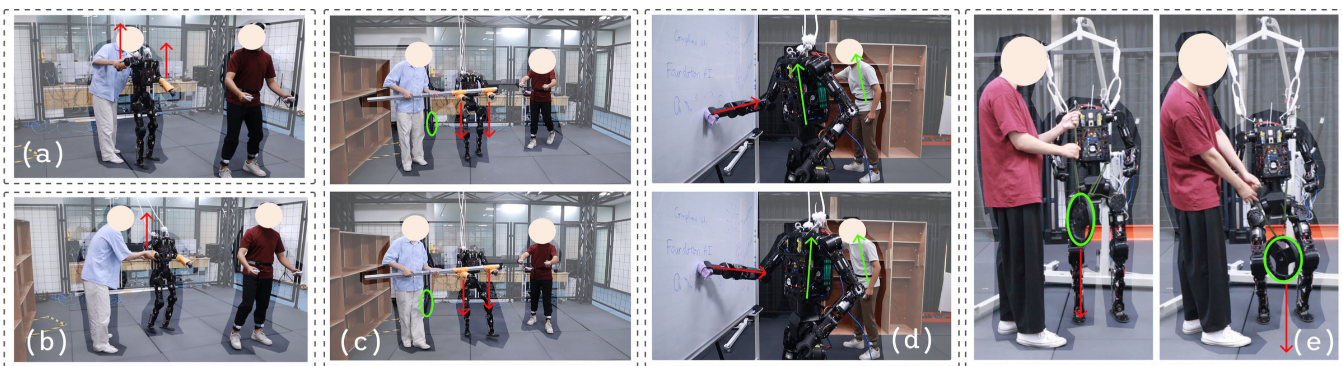

Fig. 6: We deploy COMPLIANTWBC in the real world on five force-aware tasks: (a) Static Force Reaction; (b) Dynamic Force Reaction; (c) Cooperative Payload Transport; (d) Board Wiping; and (e) Squat Under Payload.

TABLE III: Residual ablation: equilibrium source × residual on/off, over the same 100 paired rollouts. $\hat{\boldsymbol{f}}$ is the estimated wrench, $\boldsymbol{f}$ the simulator ground truth (oracle); D is OURS (full) and A is OURS w/o residual. $\|\Delta\boldsymbol{x}\|$ is the mean magnitude of the residual's equilibrium edit.

Variant	$\|\boldsymbol{f}-\hat{\boldsymbol{f}}\|$ [N]	$E_{\text{imp}}\downarrow$ [cm]	$\rho_\tau\downarrow$ [$\times10^{-2}$]	$\|\Delta\boldsymbol{x}\|$ [cm]	$S\uparrow$
A ($\hat{\boldsymbol{f}}$, off)	10.1 ± 2.1	3.81 ± 0.29	1.56 ± 0.21	–	0.94
B ($\boldsymbol{f}$, off)	0	2.44 ± 0.22	1.02 ± 0.17	–	0.97
C ($\boldsymbol{f}$, on)	0	$\mathbf{2.31 \pm 0.21}$	$\mathbf{0.89 \pm 0.14}$	0.4	0.98
D ($\hat{\boldsymbol{f}}$, on)	10.1 ± 2.1	2.58 ± 0.24	0.93 ± 0.15	2.1	0.98

the contact, especially when the force is saturated at its peak in the trapezoid, consistent with its lower ρ_τ in Table II and concretizing the resist-and-fail behavior behind TWIST2's low success rate.

How accurate is the wrench estimate? Figure 7 evaluates the frozen decoder $\hat{\boldsymbol{f}}_{\text{ext}} = \psi(\boldsymbol{z}_F)$ against the ground-truth wrench over 242k evaluation samples, separated into *direction* and *amplitude*. From the the figure, direction sharpens monotonically with load: the median $\angle(\hat{\boldsymbol{f}}, \boldsymbol{f})$ falls from 33° in the 1–5 N bin to 19° above 100 N. This suggests that larger force induces sharper signal to the policy, guiding the robot to which direction it should yield to. Amplitude instead carries a systematic, monotone bias — the median ratio $\|\hat{\boldsymbol{f}}\|/\|\boldsymbol{f}\|$ runs $2.35 \to 0.97 \to 0.80$ across the same range, crossing calibration in the 10–25 N bin and under-estimating heavy loads thereafter. The force estimation analysis along the direction and magnitude axis together characterizes the residual's job: the estimate is largely *collinear* with the true wrench, so the policy yields in the right direction but by the wrong distance, leaving an equilibrium displaced along a *known* axis by an unknown amount. This is exactly the error structure a bounded edit on the equilibrium can absorb. We provide further qualitative insight into the learned behavior of the force encoder and the compliant response of the policy. Figure 4 shows the force 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 site-dependent reactions.

What does the residual correct? Table III crosses the equilibrium source (estimated $\hat{\boldsymbol{f}}$ vs. oracle $\boldsymbol{f}$) with the residual (off/on), which ablates the contribution of our residual policy introduced in Stage 2. If the residual compensates *wrench-estimation error*, it should help only when the equilibrium is built from $\hat{\boldsymbol{f}}$; if it is generic adaptation, it should help equally either way. The data support the estimation error compensation nature of the Residual: enabling it is worth $1.23\,\mathrm{cm}$ of E_{imp} and 0.63 of ρ_τ under the estimated wrench (A $\rightarrow$ D), but only 0.13 on each under the oracle (B $\rightarrow$ C). Its own output shrinks accordingly — the mean equilibrium edit $\|\Delta \boldsymbol{x}\|$ falls $5.3\times$, from $2.1\,\mathrm{cm}$ (D) to $0.4\,\mathrm{cm}$ (C) — so with nothing left to correct the residual largely switches itself off. In absolute terms it recovers 90% of the oracle gap on E_{imp} (A $3.81 \rightarrow$ D 2.58, oracle B 2.44) and closes it entirely on ρ_τ ($1.56 \rightarrow 0.93$, against B 1.02). Access to the true wrench is therefore worth little once the residual is present, which is why the coarse estimator of Figure 7 provides sufficient information for the residual policy to react.

C. Real-World Experiments

We deploy our trained policy on a real humanoid via teleoperation [46]–[48] 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 (Section IV).

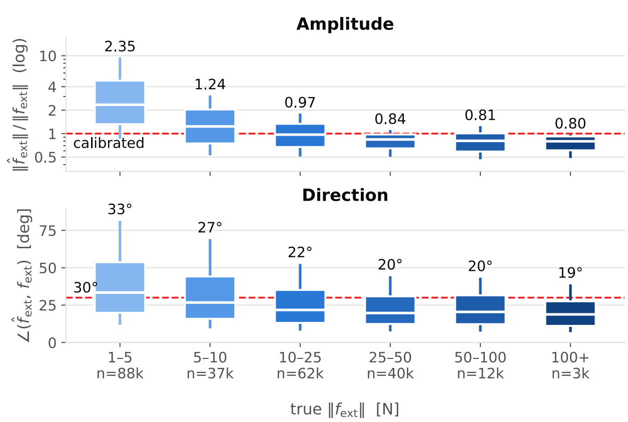

Fig. 7: Wrench-estimation quality of the frozen decoder $\psi(\boldsymbol{z}_F)$ against the ground-truth wrench, binned by true load ($n = 242\mathrm{k}$ samples). Direction sharpens monotonically with load, while amplitude carries a systematic bias that crosses calibration in the 10–$25\,\mathrm{N}$ bin.

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 $100\,\mathrm{N}$ 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—a response an upper-body-only policy is not trained for, since no perturbation reaches its lower body in training. These demonstrations are qualitatively consistent with the simulation results.

VI. CONCLUSIONS

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 encoder and a force-aware base policy against this compliance-fidelity 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. We validate the result on heavy-payload loco-manipulation tasks that require lower-body and multi-contact compliance, observing compliance behavior with better whole-body compliance with a significantly higher lower-body participation both in simulation and in the real world.

REFERENCES

[1] N. Hogan, "Impedance control: An approach to manipulation: Parts I–III," *Journal of Dynamic Systems, Measurement, and Control*, vol. 107, pp. 1–24, 1985.

[2] C. Ott, A. Albu-Schäffer, A. Kugi, and G. Hirzinger, "On the passivity-based impedance control of flexible joint robots," *IEEE Transactions on Robotics*, vol. 24, no. 2, pp. 416–429, 2008.

[3] O. Khatib, "A unified approach for motion and force control of robot manipulators: The operational space formulation," *IEEE Journal of Robotics and Automation*, vol. 3, no. 1, pp. 43–53, 1987.

[4] L. Sentis and O. Khatib, "Synthesis of whole-body behaviors through hierarchical control of behavioral primitives," *International Journal of Humanoid Robotics*, vol. 2, no. 4, pp. 505–518, 2005.

[5] B. Henze, M. A. Roa, and C. Ott, "Passivity-based whole-body balancing for torque-controlled humanoid robots in multi-contact scenarios," *The International Journal of Robotics Research*, vol. 35, no. 12, pp. 1522–1543, 2016.

[6] A. Dietrich, *Whole-Body Impedance Control of Wheeled Humanoid Robots*, ser. Springer Tracts in Advanced Robotics. Springer, 2016, vol. 116.

[7] A. Albu-Schäffer, C. Ott, and G. Hirzinger, "A unified passivity-based control framework for position, torque and impedance control of flexible joint robots," *The International Journal of Robotics Research*, vol. 26, no. 1, pp. 23–39, 2007.

[8] Y. Zhang, Y. Yuan, P. Gurunath, I. Gupta, S. Omidshafiei, A.-a. Aghamohammadi, M. Vazquez-Chanlatte, L. Pedersen, T. He, and G. Shi, "FALCON: Learning force-adaptive humanoid loco-manipulation," *arXiv preprint arXiv:2505.06776*, 2025.

[9] S. Chen, Z.-A. Cao, Z. Luo, F. Castañeda, C. Li, T. Wang, Y. Yuan, L. Fan, C. K. Liu, and Y. Zhu, "CHIP: Adaptive compliance for humanoid control through hindsight perturbation," *arXiv preprint arXiv:2512.14689*, 2025.

[10] B. Xu, H. Weng, Q. Lu, Y. Gao, and H. Xu, "FACET: Force-adaptive control via impedance reference tracking for legged robots," *arXiv preprint arXiv:2505.06883*, 2025.

[11] Q. Lu, Y. Feng, B. Shi, M. Piseno, Z. Bao, and C. K. Liu, "Gentle-Humanoid: Learning upper-body compliance for contact-rich human and object interaction," *arXiv preprint arXiv:2511.04679*, 2025.

[12] G. B. Margolis, M. Wang, N. Fey, and P. Agrawal, "SoftMimic: Learning compliant whole-body control from examples," *arXiv preprint arXiv:2510.17792*, 2025.

[13] T. He, J. Gao, W. Xiao, Y. Zhang, Z. Wang, J. Wang, Z. Luo, G. He, N. Sobanbab, C. Pan *et al.*, "Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills," *arXiv preprint arXiv:2502.01143*, 2025.

[14] S. Zhao, Y. Ze, Y. Wang, C. K. Liu, P. Abbeel, G. Shi, and R. Duan, "ResMimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning," *arXiv preprint arXiv:2510.05070*, 2025.

[15] Y. Ze, S. Zhao, W. Wang, A. Kanazawa, R. Duan, P. Abbeel, G. Shi, J. Wu, and C. K. Liu, "Twist2: Scalable, portable, and holistic humanoid data collection system," *arXiv preprint arXiv:2511.02832*, 2025.

[16] Y. Li, Y. Lin, J. Cui, T. Liu, W. Liang, Y. Zhu, and S. Huang, "Clone: Closed-loop whole-body humanoid teleoperation for long-horizon tasks," arXiv preprint arXiv:2506.08931, 2025.

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

[18] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "Humanplus: Humanoid shadowing and imitation from humans," arXiv preprint arXiv:2406.10454, 2024. [Online]. Available: https://arxiv.org/abs/2406.10454

[19] Y. Ze, Z. Chen, J. P. Araújo, Z. ang Cao, X. B. Peng, J. Wu, and C. K. Liu, "Twist: Teleoperated whole-body imitation system," arXiv preprint arXiv:2505.02833, 2025. [Online]. Available: https://arxiv.org/abs/2505.02833

[20] Z. He, C. Chen, D. Azizov, and Y. Nakamura, "Cotap: Compliant task pipeline and reinforcement learning of its controller with compliance modulation," arXiv preprint arXiv:2509.25443, 2025. [Online]. Available: https://arxiv.org/abs/2509.25443

[21] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[22] A. Demont, M. Benallegue, A. Benallegue, P. Gergondet, A. Dallard, R. Cisneros, M. Murooka, and F. Kanehiro, "The kinetics observer: A tightly coupled estimator for legged robots," arXiv preprint arXiv:2406.13267, 2024. [Online]. Available: https://arxiv.org/abs/2406.13267

[23] T. Johannink, S. Bahl, A. Nair, J. Luo, A. Kumar, M. Loskyll, J. A. Ojea, E. Solowjow, and S. Levine, "Residual reinforcement learning for robot control," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019, pp. 6023–6029.

[24] M. J. Balas and S. 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*, 2011, pp. 2338–2343.

[25] Unitree Robotics, "Unitree G1 Humanoid Robot," https://www.unitree.com/cn/g1/, 2024, accessed: 2026-05-29.

[26] F. Ferraguti, C. Secchi, and C. Fantuzzi, "A tank-based approach to impedance control with variable stiffness," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2013, pp. 4948–4953.

[27] K. Kronander and A. Billard, "Stability considerations for variable impedance control," *IEEE Transactions on Robotics*, vol. 32, no. 5, pp. 1298–1305, 2016.

[28] 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," arXiv preprint arXiv:2410.21229, 2025. [Online]. Available: https://arxiv.org/abs/2410.21229

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

[30] Z. Luo, J. Cao, K. Kitani, W. Xu *et al.*, "Perpetual humanoid control for real-time simulated avatars," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 10 895–10 904.

[31] K. Xiao, Z. Xu, Z. Cheng, C. Liu, and Q. Chen, "Kinematics-aware multi-policy reinforcement learning for force-capable humanoid loco-manipulation," arXiv preprint arXiv:2511.21169, 2025. [Online]. Available: https://arxiv.org/abs/2511.21169

[32] H. Zhang, Y. Tseng, D. Zhao, and H. E. Tseng, "Interaction-aware whole-body control for compliant object transport," arXiv preprint arXiv:2603.03751, 2026. [Online]. Available: https://arxiv.org/abs/2603.03751

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

[34] X. Zhang, C. Wang, L. Sun, Z. Wu, X. Zhu, and M. Tomizuka, "Efficient sim-to-real transfer of contact-rich manipulation skills with online admittance residual learning," in *7th Annual Conference on Robot Learning*, 2023. [Online]. Available: https://openreview.net/forum?id=gFXVysXh48K

[35] J. Buchli, F. Stulp, E. Theodorou, and S. Schaal, "Learning variable impedance control," *I. J. Robotic Res.*, vol. 30, pp. 820–833, 06 2011.

[36] A. Kumar, Z. Fu, D. Pathak, and J. Malik, "RMA: Rapid motor adaptation for legged robots," in *Robotics: Science and Systems (RSS)*, 2021.

[37] A. Kumar, Z. Li, J. Zeng, D. Pathak, K. Sreenath, and J. Malik, "Adapting rapid motor adaptation for bipedal robots," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 1161–1168.

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

[39] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[40] D. Lim, M.-J. Kim, J. Cha, D. Kim, and J. Park, "Proprioceptive external torque learning for floating base robot and its applications to humanoid locomotion," arXiv preprint arXiv:2309.04138, 2023. [Online]. Available: https://arxiv.org/abs/2309.04138

[41] P. Zhi, P. Li, J. Yin, B. Jia, and S. Huang, "Learning unified force and position control for legged loco-manipulation," arXiv preprint arXiv:2505.20829, 2025. [Online]. Available: https://arxiv.org/abs/2505.20829

[42] H. Shi, S. Hu, Y. Hou, W. Wang, K. Liu, and S. Song, "Minimalist compliance control," *arXiv preprint arXiv:2603.00913*, 2026.

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

[44] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *International Conference on Learning Representations (ICLR)*, 2017.

[45] B. T. Phong, "Illumination for computer generated pictures," *Commun. ACM*, vol. 18, no. 6, p. 311–317, Jun. 1975. [Online]. Available: https://doi.org/10.1145/360825.360839

[46] Z. Zhao, L. Yu, K. Jing, and N. Yang, "Xrobotoolkit: A cross-platform framework for robot teleoperation," arXiv preprint arXiv:2508.00097, 2025. [Online]. Available: https://arxiv.org/abs/2508.00097

[47] J. P. Araujo, Y. Ze, P. Xu, J. Wu, and C. K. Liu, "Retargeting matters: General motion retargeting for humanoid motion tracking," arXiv preprint arXiv:2510.02252, 2025. [Online]. Available: https://arxiv.org/abs/2510.02252

[48] PICO Immersive Pte. Ltd., "PICO Motion Tracker," https://www.picoxr.com/global/products/pico-motion-tracker, 2023, accessed: 2026-05-29.