SoftMimic: Learning Compliant Whole-body Control from Examples

Gabriel B. Margolis* Michelle Wang* Nolan Fey Pulkit Agrawal

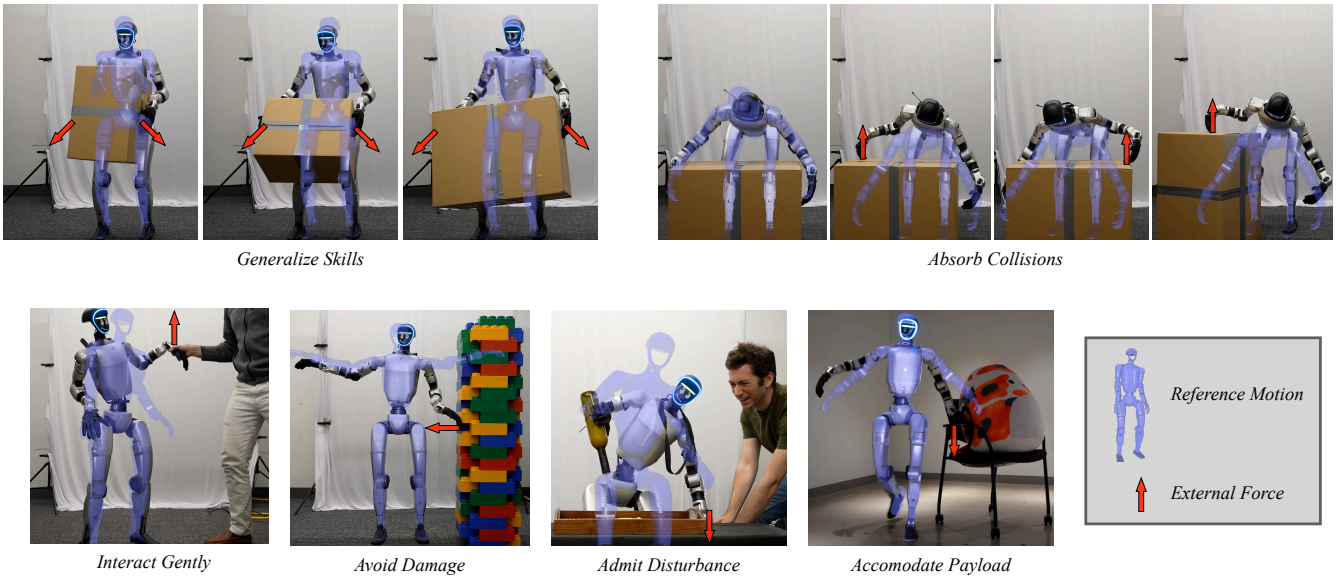


Fig. 1: **SoftMimic for Compliant Motion Tracking.** We train humanoid policies that compliantly respond to external forces while tracking a reference motion. The desired force-displacement relationship is modulated by a 'stiffness' input at deployment time, and a single policy learns to realize a wide range of stiffnesses. In diverse real-world experiments, SoftMimic benefits generalization and safety. In the images, the reference motion is visualized in blue, and the approximate external force on the robot is illustrated by the red arrows.

***Abstract*— Imitating human motions with reinforcement learning allows humanoids to quickly learn new skills. However, existing formulations reward "stiff" motion tracking that aggressively corrects errors, leading to brittle and unsafe behavior when the robot makes unmodeled contacts. We argue that for safe and generalizable interaction, a robot responding to external forces should instead *depart* from its reference motion following a user-specified stiffness. This presents a significant control challenge, as complying with a force on one limb requires sophisticated whole-body coordination to maintain balance and posture. To address this, we present a novel framework for learning compliant whole-body control. Our key idea is a learning-from-examples approach: we use an inverse kinematics solver to procedurally generate a large-scale dataset of feasible and stylistically desirable compliant motions. We then train a policy with reinforcement learning to reproduce these behaviors. By observing the original, non-compliant motion but being rewarded for matching the corresponding compliant response, the policy learns to track motions softly while leveraging the standard machinery of RL motion imitation. Our experiments show this approach enables a humanoid to controllably absorb disturbances, generalize a single motion clip to varied tasks like picking different-sized boxes, and safely handle unexpected contacts, all while preserving good motion tracking quality. We demonstrate these capabilities in simulation and on real hardware.**

All authors are with the Improbable AI Lab, Massachusetts Institute of Technology, USA. Correspondence to: {gmargo, wangmj}@mit.edu
* indicates co-first authors.
Website: https://gmargo11.github.io/softmimic

I. INTRODUCTION

A major goal in humanoid robotics is to build agents capable of performing a vast range of tasks humans execute in everyday environments. A promising avenue towards this goal is to leverage large-scale human motion capture data, enabling robots to learn human-like behaviors through imitation [1]. Recent work has successfully trained policies for tracking single motions, diverse motion datasets, and even real-time teleoperation on humanoid hardware [2]–[6]. These methods produce impressive, dynamic motions.

However, motion tracking is usually insufficient for safe and effective deployment in the real world, where sensing uncertainty and frequent, unplanned physical (i.e., contact-rich) interactions are commonplace. Policies trained to rigidly track a reference motion treat any deviation in the robot's motion as an error to be corrected aggressively. Consequently, when the robot makes an unexpected contact, such as brushing against a table, misjudging an object's location, or interacting with a person, the controller attempts to correct the motion error caused by the contact with large, uncontrolled forces, resulting in brittle and potentially dangerous behavior. This lack of compliance is also a fundamental barrier to deploying humanoids alongside people, leading to the current state of humanoids operating in isolation.

To address these shortcomings and pave the path for real-world humanoid deployment, we propose a framework for

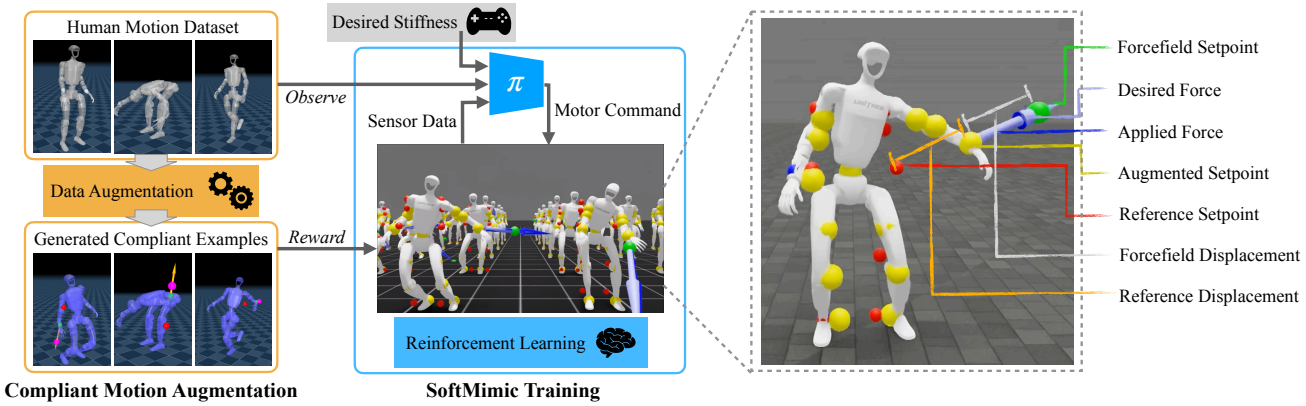


Fig. 2: **Soft Whole-body Control via Compliant Motion Augmentation.** ***Left***: Given an original reference motion (q_{ref}) and a specified interaction (external wrench W_{ext} and stiffness K_{robot}), our offline data generation stage uses an IK solver to generate a kinematically feasible and stylistically consistent compliant motion (q_{aug}). ***Right***: During online training, a policy learns to reproduce this behavior. It observes the robot's proprioceptive state and the original reference (q_{ref}), but is rewarded for matching the augmented target (q_{aug}). This forces the policy to implicitly infer the external wrench and react appropriately, resulting in a robot that can controllably comply with generalized unanticipated perturbations.

compliant whole-body motion tracking called *SoftMimic*. The objective of *SoftMimic* is not to blindly minimize tracking error, but to controllably *depart* from the reference motion in response to external forces according to a user-specified stiffness. A lower stiffness setting allows the robot to comply more and thereby deviate more from the reference trajectory, given the same force disturbance. Achieving compliant behavior on a high-DoF humanoid is challenging because complying with a force on a single limb requires coordinated, full-body adjustments to maintain balance and preserve the overall posture and style of the motion.

Directly learning compliant behavior with reinforcement learning (RL) poses significant exploration challenges, as a stiff, non-compliant policy is often a strong local optimum. In scenarios where the robot must comply substantially, reward terms like tracking keypoints and joint angles, which typically reinforce each other, can come into conflict, making it difficult to balance them across a large potential solution space [7]. Furthermore, many desired compliant responses are kinematically or dynamically infeasible depending on the robot's configuration, resulting in training for impossible tasks, which hinders learning [8].

To overcome these challenges, we adopt a learning-from-examples strategy. Instead of asking RL to discover compliant behaviors through intricate reward shaping, we first generate a dataset of kinematic references for compliant behaviors. We use an inverse kinematics (IK) solver to author a large-scale dataset of feasible and stylistically-consistent compliant trajectories for a wide range of interaction scenarios. This offline process allows us to filter out impossible tasks and precisely define the desired whole-body coordination. We then train a policy using RL, where the agent observes the robot's state and the original, non-compliant reference motion, but is rewarded for tracking the corresponding pre-computed compliant trajectory from our augmented dataset. This formulation forces the policy to learn to infer the external forces from proprioceptive sensing and react with the demonstrated compliant behavior.

Our experiments demonstrate that this approach yields a policy capable of tracking reference motions while exhibiting predictable, controllable compliance. Our compliant controller is more robust to disturbances, can generalize a single motion clip to handle variations in a task (e.g., picking up boxes of different sizes), and safely manages unexpected collisions. Crucially, these benefits are achieved while preserving good motion tracking performance in the absence of external forces. We validate our framework in simulation and on a real Unitree G1 humanoid robot.

II. RELATED WORK

A. Learning Humanoid Whole-Body Control

The convergence of recent progress in articulated rigid-body simulation [9]–[11], sim-to-real transfer techniques [12]–[14], and advanced legged hardware [15]–[18], combined with reinforcement learning paradigms like DeepMimic [1], [19], has enabled impressive performance in humanoid robot motion imitation [2], [5], [6] and real-time teleoperation [3], [4], [20]. Building on these components, some works use such whole-body controllers to teleoperate new tasks and train visuomotor policies on the resulting demonstrations, establishing a mapping from image observations to whole-body motion references [21]–[24].

During imitation or teleoperation, a common scenario is that the robot contacts an object while the teleoperator or motion reference does not. Light objects may be pushed out of the way or lifted, but heavier objects or fixtures may impede the robot and inhibit it from matching the reference. A natural question is what posture the robot should adopt when it is forced away from the reference motion, and how much force the robot should exert against the environment when attempting to reduce tracking error. Traditional factory arms that are purely position-controlled are extremely stiff, and

consequently may damage themselves or the environment with large and unpredictable forces when impeded, making them dangerous and brittle to small environmental variations. Modern quasi-direct-drive (QDD) actuators support torque sensing and control, which allows them to realize different stiffnesses through software emulation. A prevalent strategy within learning-based whole-body control frameworks is for a neural network policy to actuate the robot's QDD motors by sending target setpoints to a PD controller in each joint. Such policies can modulate position targets to intentionally incur position error, regulating applied forces during dynamic maneuvers [13]. Furthermore, PD gains can be tuned to shape the torque and position distributions excited by Gaussian policy exploration [25]. A natural misconception might be that lower gains or even direct torque control will always result in a compliant or 'soft' robot policy. In reality, as we show, the stiffness of a policy's interactions is dictated foremost by its high-level incentives, i.e. its reward function and training data. We also find neural network policies trained with constant low-level gains are capable of representing a wide range of stiff and compliant behaviors in task-space.

Works that combine motion tracking rewards with random external forces or pushes instruct the robot to follow the same trajectory regardless of the interaction force [20], [26], [27]. This encourages the policy to apply arbitrary resistive forces to maintain minimal tracking error, essentially acting as stiff as possible.

B. Classical Approaches to Compliant Control

Hybrid position/force control [28] and task-space impedance–admittance control [29] are longstanding formulations for compliant manipulation in robotics, which prescribe a motion–wrench relationship (e.g. a mass–spring–damper at the end effector) and synthesize joint torques that realize a specified apparent stiffness and damping. Extending these ideas from fixed-base robot arms to humanoids requires additionally modeling a floating base, intermittent contacts, and the need to coordinate interaction objectives with posture and balance. The operational-space formulation provides an approach to control the robot while balancing multiple tasks such as force interaction and posture control [30]. Whole-body operational-space control extended this to floating-base systems under contact constraints, organizing interaction tasks alongside posture and balance through contact-consistent projections [7], [31]–[33]. To implement compliant performance on robot hardware, research at the German Aerospace Center (DLR) has developed controllers for torque-controlled manipulators with elastic transmissions, utilizing cascaded control based on theory that accounts for compliance and modeling error [34], [35]. These analytical approaches have yielded impressive demonstrations of precise torque sensing, backdrivability, and safe physical interaction on robot arms as well as on the DLR Torque-controlled humanoid Robot (TORO) [36].

Drawing inspiration from the above literature, we develop a compliant approach to RL motion imitation (SoftMimic) which incorporates an explicit task-space interaction law. Our approach adopts the classical goal of making the robot behave like a spring in response to generalized disturbances, but replaces hand-engineered controllers with a learned policy trained on procedurally generated compliant trajectories. The policy observes proprioception and the original (non-compliant) reference, and is rewarded for reproducing the authored compliant deviation. This allows user-specified stiffness to be realized while maintaining high-fidelity whole-body motion tracking.

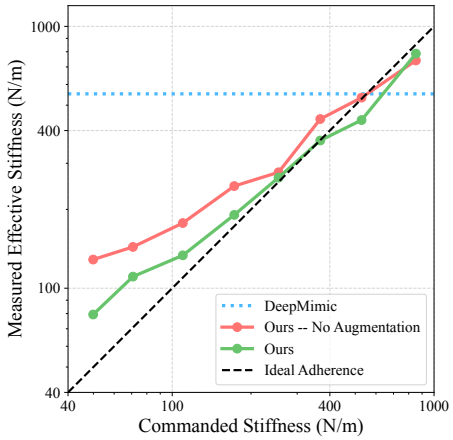


Fig. 3: **Stiffness adherence.** The humanoid's effective translational stiffness tracks the commanded stiffness over a wide range. We apply external forces to the hands of a standing robot in simulation and report the median force–displacement ratio for a single *SoftMimic* policy across stiffness commands (log–log scale). A stiff motion-tracking baseline maintains an approximately constant effective stiffness of $550\,\mathrm{N\,m^{-1}}$ under identical conditions. Data augmentation further improves adherence, especially at low stiffness (Fig. 8).

C. Data-Driven Compliant Control

Recently, reinforcement learning has been used to directly learn compliant behaviors. Initial explorations, such as Deep Compliant Control [37], demonstrated success with simulated characters but relied on perfect state information, sidestepping the real-world challenges of force estimation and model uncertainty. Portela et al. [38] made a key step toward hardware applications by showing that an end-to-end policy trained in simulation can learn to apply accurate task-space forces on a real legged manipulator using only proprioceptive actuators, and demonstrated that this facilitates impedance control of the robot's end-effector. Other work has trained locomotion policies to directly mimic a specific dynamic model, such as a spring-mass-damper template [39], a concept extended by FACET [40] to various embodiments. UniFP [41] demonstrated that explicit force information obtained from such policies can benefit imitation learning for downstream tasks. These prior approaches focus on controlling the force interaction while satisfying a simple locomotion task. Especially in the case of humanoid robots learning from human teleoperation and/or demonstrations, it is often critical to reconcile interaction objectives with

whole-body motion tracking in order to complete a task. A unified framework that combines wide-range impedance control with high-fidelity motion mimicry on real hardware remains an open challenge. Our work addresses this gap by training a single policy to imitate reference motions while achieving a user-specified stiffness, enabling both soft compliance and stiff resistance (Figure 3)

III. Method

Our goal is to train a policy that enables a humanoid robot to track a whole-body reference motion while compliantly responding to external forces with a user-specified stiffness. A naive approach could involve a standard motion imitation setup [1] with an additional task reward for compliant responses. Directly optimizing this objective with RL is not ideal for several reasons. First, exploration is brittle: a purely stiff tracker is a strong local optimum that suppresses compliant responses when these rewards are in conflict. Second, the humanoid's large postural null-space makes reward design–balancing interaction forces with whole-body style and stability–nontrivial. Third, the robot's feasible compliance is highly dependent on its configuration due to kinematic constraints and sensing limitations, yielding many tasks infeasible. Fourth, the desired deviation from reference motions is incompatible with the use of *early termination* and *reference state initialization* strategies commonly used to stabilize and accelerate training.

Our solution to these problems is to generate an *augmented* dataset with reference motions that specify how the robot should comply to different external forces. This sets up a motion tracking problem where the robot observes the original motion target but is rewarded for inferring the force interaction and matching the applicable augmented target. We show that this approach enables fine-grained control of the compliant response. A key challenge is how to generate a dataset of feasible complying motions that preserve desired components of the original motion style. In this work, we use differential inverse kinematics to ensure kinematically feasible and stylistically desirable reference motions, and an analysis of force and position sensing noise to specify feasible force/compliance tasks.

A. Compliant Motion Tracking (CMT)

Given an original motion reference, represented by the joint configuration q_{ref}, and an external wrench W_i applied to link i, the objective of compliant whole-body control is to achieve a corrected configuration $\hat{q}$. Let $f_i(q)$ be the forward kinematics function that computes the Cartesian pose of link i from the joint configuration q. The optimal configuration $\hat{q}$ is found by solving the following optimization problem where K_{robot} is the desired robot stiffness

$$\hat{q} = \arg\min_{\hat{q}} d(\hat{q}, q_{\text{ref}}) \quad \text{s.t.} \quad W_i = K_{\text{robot}}(f_i(\hat{q}) - f_i(q_{\text{ref}}))$$

This formulation dictates that link i, the link experiencing the external wrench, must behave like a spring (as defined by the constraint), while the rest of the robot's posture remains as close as possible to the original reference configuration q_{ref},

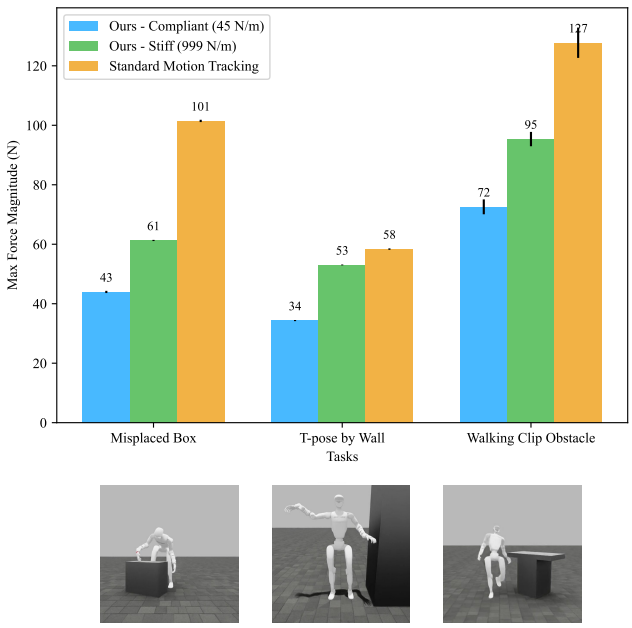


Fig. 4: **SoftMimic reduces collision forces across various motions in unseen environments.** The bar chart compares the maximum contact force generated by our policy (at low and high stiffness) and the stiff baseline across three challenging scenarios involving unexpected contact. In all cases, the compliant policy operating at a low stiffness significantly reduces interaction forces, enhancing safety.

as measured by some distance metric d. When the robot's stiffness is low or the external force is large, the optimal configuration $\hat{q}$ can deviate significantly from the reference q_{ref}. In such cases, the choice of the metric d (e.g., a simple joint-space error like $\|\hat{q} - q_{\text{ref}}\|^2$ versus a task-space error on other keypoints) has a large impact on the resulting behavior. This contrasts with typical motion tracking scenarios where the optimal solution remains close to the reference, and all errors are near zero. In this work, we choose to define d using a mixture of keypoint error, joint position error, foot placement consistency, and center-of-pressure maintenance as described in Section III-C.

B. SoftMimic: Reinforcement Learning for CMT

Observation, Reward, Action Space. We formulate compliant whole-body control as a reinforcement learning problem. The policy observes a state containing the robot's proprioceptive information $[q_t, \dot{q}_t]$, base state $[g_t^b, \omega_t^b]$, previous action a_{t-1}, and reference posture q_t^{ref}. The agent is rewarded with a sum of a DeepMimic-style reference motion tracking reward, $r_{\text{ref}} + r_{\text{smooth}}$, and a spring-like compliance reward, $r_{\text{spring}} = r_{\text{force}} + r_{\text{torque}} + r_{\text{pos}} + r_{\text{rot}}$, which depends on the current external wrench W_i. The policy outputs joint-space position targets for a PD controller with moderate gains, enabling torque control by modulating the position error.

Observation Content. The policy can implicitly learn admittance-style (estimate wrench, command pose), impedance-style (estimate pose, command wrench), or

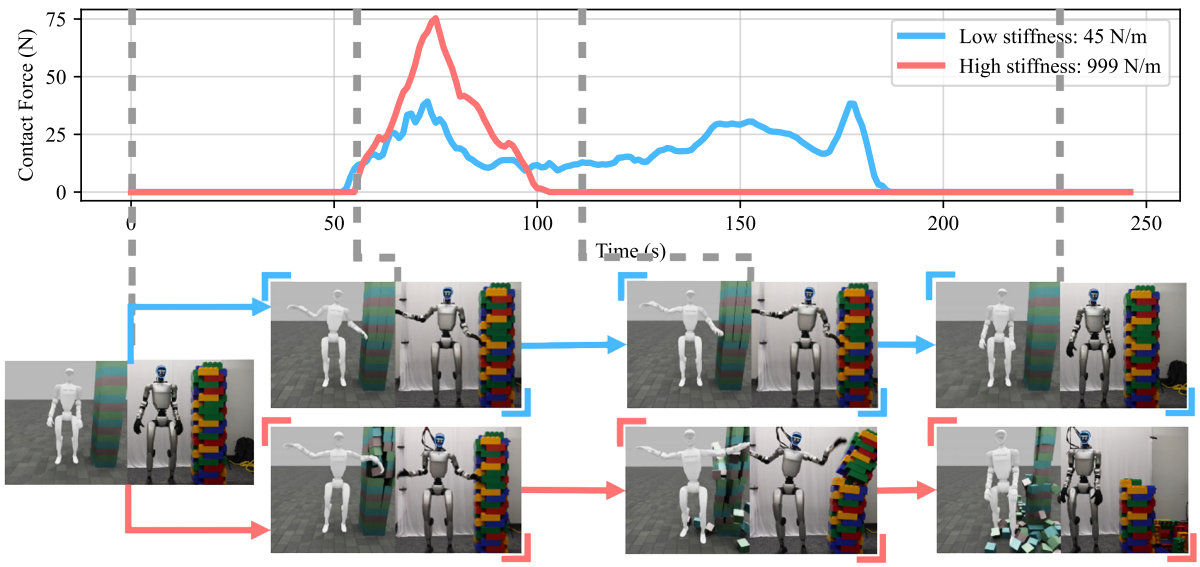
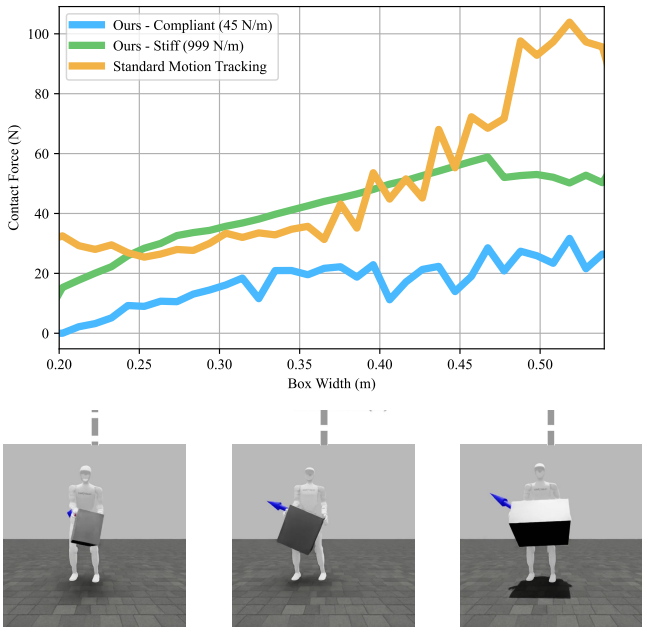


Fig. 5: **Stiffness modulation controls collision forces.** The plot shows the contact force over time as the robot's hand collides with a tower of blocks. By commanding different stiffness levels, our policy can produce low, controlled forces (blue) or high, potentially destructive forces (red), showcasing a direct trade-off between safety and posture tracking accuracy.

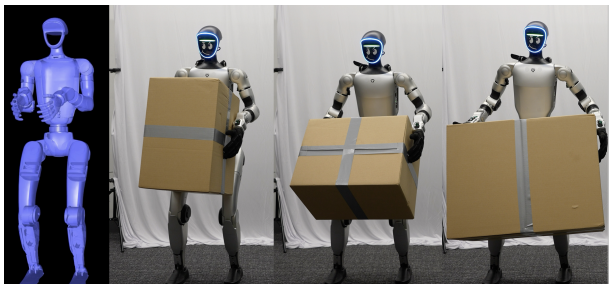
hybrid strategies depending on the desired stiffness and external force profile. Note that the policy directly observes neither the wrench nor displacement information, but can make inferences about them based on proprioceptive sensing. For an impedance strategy, the end-effector pose can be inferred from the joint positions q_t and root orientation g_t^b via forward kinematics. For an admittance strategy, the external wrench can be inferred from the robot's dynamics, using observations of previous joint position q_{t-1}, joint position target a_{t-1}, joint velocity $\dot{q}_{t-1}$, and joint accelerations (derived from $\dot{q}_t, \dot{q}_{t-1}, \omega_t^b, \omega_{t-1}^b$). To ensure this temporal information is available, the policy observes a history of the past 3 observation steps. As is standard in legged systems, the full root state and contact states are not directly observed; instead, the policy may partially infer them as needed, leveraging the associations between historical observations, commands, and simulation outcomes.

Command Sampling. Training episodes consist of sampling a motion clip, a desired robot stiffness, and an external force profile. The external force is implemented as a 'force field' [38] which pulls a selected link of the robot towards a moving setpoint with a distance-proportional force according to a randomized environment stiffness K_{env}. $K_{\text{env}} \to 0$ corresponds to a constant-force source (an admittance-like environment) and $K_{\text{env}} \to \infty$ corresponds to an immovable object (an impedance-like environment) [42]. Additional details of the force sampling parameters are provided in the appendix.

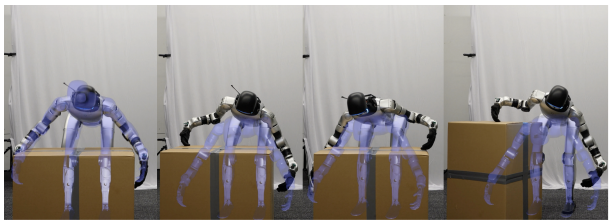
- *Stiffness Bounds:* When sensing and dynamics are noisy and the robot's state is only partially observable, inferences about pose and wrench also become noisy. This noise makes realizing highly sensitive responses, including extremely low or high stiffnesses, infeasible. To address this, we first train a state estimator to establish the approximate noise floor of the pose and wrench estimates. We then use these noise values in an idealized analysis to guide our stiffness sampling range. We define requirements of $10\,\text{N}$ force accuracy and $10\,\text{cm}$ position accuracy, and empirically observe that the learned force estimator has average noise of $4\,\text{N}$. Thus, an admittance control strategy should be able to achieve the positioning target only for $K > \frac{4\,\text{N}}{0.10\,\text{m}} = 40\,\text{N/m}$. Likewise, with a position estimation noise of $1\,\text{cm}$, an impedance controller can achieve the desired force accuracy only if $K < \frac{10\,\text{N}}{0.01\,\text{m}} = 1000\,\text{N/m}$. This analysis establishes approximate upper and lower feasible bounds for training.
- *Log-Uniform Sampling:* We aim to realize behaviors across a wide range of stiffness and compliance values. Since compliance is the inverse of stiffness, uniformly sampling stiffness would heavily bias the dataset towards high-stiffness, low-compliance behaviors, and vice versa. A change in stiffness from 1040 to $1080\,\text{N/m}$ is a minor tweak to a stiff behavior, while a change from 40 to $80\,\text{N/m}$ is a significant change for a compliant one. To ensure we explore these different regimes equally, we sample both the robot and environment stiffness from a log-uniform distribution.
- *Velocity-based Event Sampling:* Suppose that every point in space has the same probability of containing a stationary collision surface. Then if the robot is moving with no information about its surroundings, its probability of some point on the robot colliding is proportional



(a) **Controllable interaction force:** Simulated normal contact force vs. box width using one object-agnostic reference motion. Softmimic force increases predictably with box width; a stiff tracker produces large spikes resulting in box damage or torque limit violations on real hardware.



(b) **Generalization of a single reference:** Real-world deployment with the same reference motion and stiffness: successful picks across multiple box widths with a gentle, consistent squeeze.



(c) **Zero-shot robustness to misalignment:** picking a large box with nominal alignment (left) and under lateral/rotational misinitializations (middle–right). All behavior is achieved without simulating boxes or defining a prior over their location; robustness comes from SoftMimic training with generalized external forces.

Fig. 6: **SoftMimic enables generalization to unseen objects and disturbance scenarios.** Using a single motion reference designed for a 20cm box, our policy can successfully pick up boxes of increasing width. By commanding a low stiffness, the robot applies a consistent, gentle squeezing force across all sizes. In contrast, the stiff baseline generates large and uncontrolled force spikes, risking damage to the object or robot.

to the point's velocity. Therefore, we sample force event onsets for each link with probability proportional to its velocity, with a small constant additional probability for colliding while stationary.

Early Termination and Reference State Initialization. It is common practice to exploit early termination and reference state initialization to accelerate and stabilize motion imitation [1]. A key advantage of our data augmentation approach (Section III-C) is that we can appropriately terminate and initialize episodes while the robot is under load by using the augmented compliant posture q_{aug} as the reference. Without this augmented data, there would be no way to initialize the robot consistently with active wrenches,

C. Compliant Motion Augmentation (CMA)

A key challenge with the RL problem posed in Section III-B is that the final compliant posture arises from a competition between the motion tracking and spring-behavior rewards. This complicates exploration and makes the resulting behavior difficult to tune and predict.

Our proposed solution is to pre-generate an augmented motion dataset, D_{aug}, that explicitly contains desired whole-body responses to force events. This offline process enables two key advantages: 1) we can reject infeasible commands before RL training using simple kinematic and dynamic checks, and 2) we can precisely specify the desired compliant behavior through a structured optimization. We generate D_{aug} using a differential inverse kinematics (IK) solver.

Task Hierarchy: The IK solver optimizes for the following objectives:

1) **Compliant Interaction (high priority; $w = 5.0$).** For the interacting link, the target pose is defined to yield like a spring with Cartesian stiffness (K_{robot}, k_r):

$$\mathbf{p}_{\text{target}} = \mathbf{p}_{\text{ref}} + \frac{\mathbf{F}_{\text{ext}}}{K_{\text{robot}}}, \qquad \mathbf{R}_{\text{target}} = \mathbf{R}_{\text{ref}} \exp\left([\boldsymbol{\tau}_{\text{ext}}/k_r]_\times\right). \tag{1}$$

   In this work, we only sample wrenches at the hand links and only perturb a single link at a time.
2) **Foot Placement (high priority; $w = 2.5$).** High-weight link pose tasks ensure that stance feet remain consistent with the reference contact schedule.
3) **CoM Stabilization (medium priority; $w = 0.1$).** A Center of Pressure (CoP)-aware Center of Mass (CoM) task provides moment compensation while allowing necessary body shifts.
4) **Keypoint Posture (low priority; $w = 0.01$).** Moderate-weight pose tasks on key links (e.g., elbows, shoulders, torso) preserve the original motion's style.
5) **Joint Posture (very low priority; $w = 10^{-4}$).** A regularization task tethers all degrees of freedom towards the reference configuration $\mathbf{q}_{\text{ref}}$ to resolve redundancy.

This hierarchy yields a continuous and feasible adapted joint trajectory, $\mathbf{q}_{\text{adapted}}(t)$, that embodies the desired compliant

response across various interaction scenarios. When the IK solver fails to find a solution for a given wrench, we rewind the motion clip and iteratively scale down the wrench, rejecting the event entirely if the wrench falls below the sensing noise floor.

D. Motion Data, Training Details, Baselines

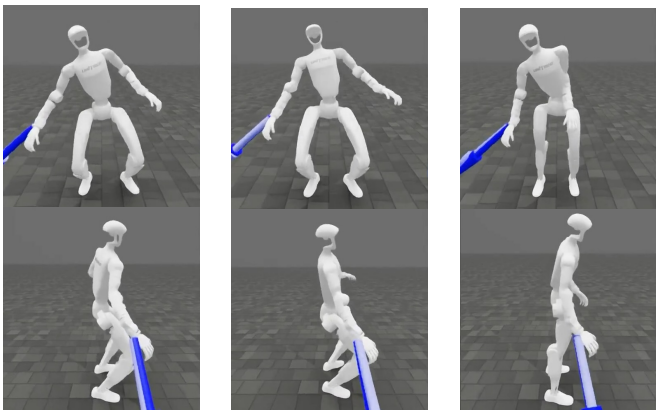
Motion Data. We trained and deployed compliant whole-body control policies on a Unitree G1 humanoid—one policy for each motion clip: standing, T-pose-move, walk, box-pick, pour, and dance, using identical hyperparameters. The motion data comes from the AMASS [43] and LAFAN1 [44] datasets, retargeted using methods from prior work [2], [4].

For each motion clip, we generate augmented data by solving the aforementioned inverse kinematics problem using Mink [45], [46] and MuJoCo [10]. The offline process is highly efficient, allowing us to generate 40 minutes of augmented data for a one-minute clip in approximately one minute of wall-clock time when parallelized. This produces a dataset of tuples $(q_{\text{ref}}, W_i, K_{\text{robot}}, q_{\text{aug}})$ that defines all interaction events for training.

Training Hyperparameters. Linear stiffness commands ranged from $10\,\text{N}\,\text{m}^{-1}$ to $1000\,\text{N}\,\text{m}^{-1}$; angular stiffness from $0.1\,\text{N}\,\text{m}\,\text{rad}^{-1}$ to $10\,\text{N}\,\text{m}\,\text{rad}^{-1}$. We train using PPO with the default hyperparameters from the IsaacLab and `rsl_rl` libraries [47].

Baselines. To rigorously evaluate our method, we compare it against two carefully designed baselines that aim to isolate the different components of our framework.

1) **Stiff Baseline:** To demonstrate the value of explicit compliance, we first compare against a high-performance baseline analogous to standard motion imitation methods [1]. This **Stiff Baseline** is trained with a reward function that only incentivizes rigid tracking of the original reference motion, q_{ref}. Crucially, to ensure a fair and direct comparison, this baseline is exposed to the exact same distribution of external force perturbations during training as our compliant policy. This setup tests the emergent behavior of a state-of-the-art tracking controller when faced with physical interactions it is not explicitly rewarded to handle.
2) **`no-aug` Ablation:** To specifically isolate the contribution of our learning-from-example data generation strategy, we design an ablation called **`no-aug`**. This policy is trained with the same spring-like compliance reward, r_{spring}, as our full method, but it does not have access to the augmented dataset D_{aug}. Consequently, its motion tracking reward, reference state initializations, and termination conditions are all based on the original non-compliant reference, q_{ref}. This creates a significant learning challenge: successful compliance generates a large tracking error relative to q_{ref}, which would normally trigger an early termination and thus penalize the desired behavior. To create a meaningful and learnable task, we modify the termination condition for this ablation: an episode only terminates if the robot's state deviates significantly from q_{ref} *without* satisfying the compliant displacement objective. This necessary adjustment allows the policy to explore compliant behaviors without being immediately punished, enabling a fair evaluation of learning with reward shaping alone.



(a) IK Style 1 (b) IK Style 2 (c) Ablation: `no-aug`

Fig. 7: **Compliant Motion Augmentation provides fine-grained control over compliant style.** Three different compliant policies are shown receiving the same external force in simulation with the same commanded stiffness. By adjusting cost terms in the offline IK solver—such as adding a pelvis orientation cost in (b) compared to (a)—we can specify distinct whole-body coordination strategies. The learned policies successfully reproduce the authored styles. In contrast, the policy trained without augmented data (`no-aug`, c) adopts an unpredictable emergent posture that also performs worse.

IV. Results

A. Motion Tracking Should Be Compliant

Compliance Improves Task Generalization. Compliant imitation of a single motion can enable its generalization to different task variations. We demonstrate this through a box-picking motion. A single motion reference of a person picking a box of fixed size is used during training. During deployment, a natural approach compatible with non-compliant WBC would be to perceive the size and location of the box visually and map this to a reference motion – either through an explicit perception module or via a learned high-level policy. We are interested in the scenario where the perception module is noisy and erroneously estimates the size or location of the box. In Figure 6, we compare the force exerted on differently sized boxes by our compliant whole-body controller vs. the standard motion imitation approach; both are tracking a single original motion reference. Our method is able to maintain a lower squeezing force while successfully picking up differently sized boxes, while the standard method exhibits larger and unpredictable forces as it faces larger box sizes outside the scope of the original motion reference.

Compliance Improves Disturbance Handling. We evaluated the response of compliant policies to unseen environmental circumstances common during deployment of humanoid robots.

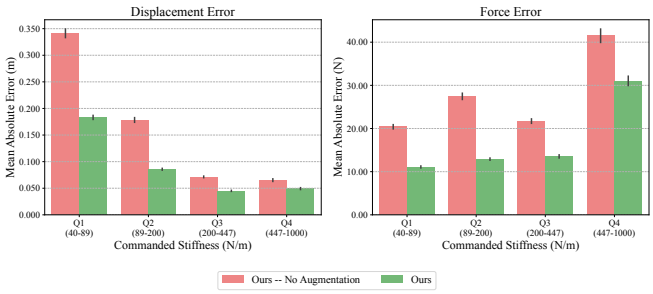


Fig. 8: **Effect of Compliant Motion Augmentation on compliance accuracy.** Policies trained with augmented compliant trajectories (green) attain lower position and force error than the `no-aug` ablation (red), with the largest gains at low stiffness where coordinated whole-body deviations are substantial.

- *T-Pose by Wall:* The robot attempts to raise its arm while standing next to a wall.
- *Walking Clip Obstacle:* The robot walks past a table and its hand clips the corner.
- *Misplaced Box:* The robot attempts to execute a bending pick while the target box is not centered, and hits its hand on the top of the box.

Figure 4 reports the maximum force the robot exerts on the environment for each task, evaluated at two different stiffness levels of our method as well as with the standard motion tracking approach. The very compliant policy exerts significantly lower forces on the environment, showing that compliant whole body control can more safely handle forceful disturbances compared to existing baselines.

Compliance Improves Safety. Figure 5 shows how modulating the commanded stiffness results in drastically different environment interactions. A small stiffness results in the robot gently pushing on the tower and deviating significantly from the T-pose reference, while a large stiffness causes the robot to strongly resist deviations to the original reference motion, consequently exerting a large force on and toppling the blocks. Figure 1 shows how the robot complies in the real world during various interactions.

B. Evaluating Stiffness Adherence

We apply external forces to the standing robot in simulation and measure the resulting displacement across a range of stiffnesses. Figure 3 shows the median effective stiffness (computed as the ratio between force and displacement) evaluated at various stiffness levels on a log-log plot. The standard motion tracking baseline, which is not conditioned on a stiffness command, yields an effective stiffness of about 500. As can be seen in the supplementary videos, the stiff policy preserves its posture when externally forced but tends to shuffle its feet, which registers as compliance in this evaluation conducted in the global reference frame. Our method displays a consistent sensitivity to the stiffness command across the entire range used in training. Figure 8 shows the displacement is often regulated below 10 cm and force error below 15 N with exceptions at the lowest stiffnesses (elevated displacement error) and highest stiffnesses (elevated force error). Figure 3 and 8 also show that training with augmented references boosts performance compared to the ablation `no-aug`, particularly at low stiffnesses where it results in a 50% reduction in displacement error.

TABLE I: **Motion tracking quality comparison.** Comparison of tracking error under no-perturbation conditions (free space) for our compliant policy and a stiff baseline on various skills. Errors are reported as mean joint position error (degrees) and keypoint Cartesian error (cm), with standard error of the mean over 36 episodes.

Skill	Ours (Compliant)		Stiff Baseline	
	Joint (°)	Keypoint (cm)	Joint (°)	Keypoint (cm)
Box Pick	5.04 ± 0.01	2.65 ± 0.01	2.04 ± 0.00	1.36 ± 0.00
Walk	6.39 ± 0.00	3.44 ± 0.00	6.09 ± 0.00	3.50 ± 0.00
Dance	11.10 ± 0.01	6.05 ± 0.01	5.16 ± 0.01	3.01 ± 0.00

C. Data Shaping Controls Behavior

A key benefit of our framework is the ability to resolve task specification ambiguities in the data augmentation stage. To illustrate this, we train compliant standing policies with two different IK-generated compliant datasets, one with a relatively higher pelvis orientation cost term that encourages the robot to squat and one with a relatively lower term that encourages the robot to bend. Figure 7 shows how the resulting policies respond to perturbations in different styles depending on the behavior designed during the IK dataset generation. It also compares the behavior of the best `no-aug` policy, which displays an emergent postural response resulting from the balance of rewards, which cannot be predicted before performing the expensive RL training.

D. Compliant Control Preserves Motion Quality

Our proposed method achieves compliance when interacting with forces and preserves competitive motion tracking accuracy in the non-perturbed case, even for long and dynamic motion clips. Under no perturbations, we compare the joint position and keypoint tracking error of our method and standard motion tracking for skills used in our demonstrations, as well as a long, challenging dance clip (`dance1_subject2` from LAFAN1 [44]) which has recently been used to demonstrate the high performance of motion tracking systems. Table I shows both our compliant policy and the standard motion tracking baseline achieve small tracking errors. This minor increase in tracking error is an expected trade-off for learning a much richer and more versatile behavioral repertoire. Figure 9 (Appendix) shows the training progression of total reward for SoftMimic vs. baseline and the convergence of compliance objectives. It takes a bit more time to train SoftMimic to convergence compared to stiff motion tracking. The policy must learn not only to track a single motion but also to embed a wide range of compliant responses.

V. Conclusion

This work introduced a formulation for learning compliant whole-body motion tracking for humanoid robots.

We demonstrate that our compliant policy outperforms the standard motion tracking baseline in generalization to unseen manipulation scenarios and in safety when handling disturbances. In qualitative experiments, we see that different user-commanded stiffness values can drastically change how the robot interacts with the environment—from gently pushing to toppling a tower of blocks—and with people during human-robot interaction. Quantitatively, we find that a compliant robot can generalize to unseen objects without applying excessive force, and when colliding with a disturbance, our policy applies nearly half the force of the standard baseline. Our policy also demonstrates good adherence to the commanded stiffness across a wide range of values and can comply across a broad workspace. Finally, under unperturbed conditions, our approach achieves tracking performance comparable to a state-of-the-art baseline across diverse motion clips, including whole-body locomotion and manipulation.

Looking forward, a key area for future work is determining how to best select stiffness for a given task. While our experiments showcase the benefits of lower stiffness for safety and generalization–demonstrating that a single fixed value can be effective for diverse scenarios such as lifting a misplaced object–we anticipate that real-world deployments will require dynamically adjusting stiffness; for example, using higher stiffness to lift a heavy box versus lower stiffness to gently hand an object to a person. Further improvements could also come from training a foundational compliant whole-body controller capable of tracking large-scale motion datasets or live teleoperation. The success of such a model depends heavily on the quality of its training data. Although our kinematic data augmentation was sufficient to realize useful behaviors, its quality could be enhanced by incorporating dynamics to yield more physically plausible motions. We could also improve robustness to multimodal behaviors and address gaps in workspace coverage caused by our rejection sampling approach. Alternatively, future research could bypass explicit data augmentation by leveraging advances in pure reinforcement learning to discover compliant skills directly from interaction. Finally, motivated by the many benefits of whole-body compliance, another promising direction is to define compliant behavior for wrenches on any link of the body, rather than only the wrists, and for multiple links simultaneously, in pursuit of fine-grained stiffness control across the robot's entire surface.

REFERENCES

[1] X. B. Peng, P. Abbeel, S. Levine, and M. Van de Panne, "Deepmimic: Example-guided deep reinforcement learning of physics-based character skills," *ACM Transactions On Graphics (TOG)*, vol. 37, no. 4, pp. 1–14, 2018.

[2] W. Xie, J. Han, J. Zheng, H. Li, X. Liu, J. Shi, W. Zhang, C. Bai, and X. Li, "Kungfubot: Physics-based humanoid whole-body control for learning highly-dynamic skills," *arXiv preprint arXiv:2506.12851*, 2025.

[3] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, "Expressive whole-body control for humanoid robots," *arXiv preprint arXiv:2402.16796*, 2024.

[4] T. He, Z. Luo, W. Xiao, C. Zhang, K. Kitani, C. Liu, and G. Shi, "Learning human-to-humanoid real-time whole-body teleoperation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 8944–8951.

[5] Z. Chen, M. Ji, X. Cheng, X. Peng, X. B. Peng, and X. Wang, "Gmt: General motion tracking for humanoid whole-body control," *arXiv preprint arXiv:2506.14770*, 2025.

[6] Q. Liao, T. E. Truong, X. Huang, G. Tevet, K. Sreenath, and C. K. Liu, "Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion," *arXiv e-prints*, pp. arXiv–2508, 2025.

[7] L. Sentis and O. Khatib, "Task-oriented control of humanoid robots through prioritization," in *IEEE International Conference on Humanoid Robots*, 2004, pp. 1–16.

[8] G. B. Margolis, G. Yang, K. Paigwar, T. Chen, and P. Agrawal, "Rapid locomotion via reinforcement learning," *Robotics: Science and Systems*, 2022.

[9] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, and G. State, "Isaac gym: High performance gpu-based physics simulation for robot learning," 2021.

[10] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 5026–5033.

[11] K. Zakka, B. Tabanpour, Q. Liao, M. Haiderbhai, S. Holt, J. Y. Luo, A. Allshire, E. Frey, K. Sreenath, L. A. Kahrs *et al.*, "Mujoco playground," *arXiv preprint arXiv:2502.08844*, 2025.

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

[13] J. Hwangbo, J. Lee, A. Dosovitskiy, D. Bellicoso, V. Tsounis, V. Koltun, and M. Hutter, "Learning agile and dynamic motor skills for legged robots," *Science Robotics*, vol. 4, no. 26, p. eaau5872, 2019.

[14] G. B. Margolis and P. Agrawal, "Walk these ways: Tuning robot control for generalization with multiplicity of behavior," in *Conference on Robot Learning*. PMLR, 2023, pp. 22–31.

[15] M. Hutter, C. Gehring, D. Jud, A. Lauber, C. D. Bellicoso, V. Tsounis, J. Hwangbo, K. Bodie, P. Fankhauser, M. Bloesch *et al.*, "Anymal-a highly mobile and dynamic quadrupedal robot," in *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2016, pp. 38–44.

[16] P. M. Wensing, A. Wang, S. Seok, D. Otten, J. Lang, and S. Kim, "Proprioceptive actuator design in the mit cheetah: Impact mitigation and high-bandwidth physical interaction for dynamic legged robots," *Ieee transactions on robotics*, vol. 33, no. 3, pp. 509–522, 2017.

[17] B. Katz, J. Di Carlo, and S. Kim, "Mini cheetah: A platform for pushing the limits of dynamic quadruped control," in *2019 international conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 6295–6301.

[18] Unitree Robotics, G1, 2025, https://www.unitree.com/g1, [Online; accessed Sep. 2025].

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

[20] Y. Ze, Z. Chen, J. P. AraÃšjo, Z.-a. Cao, X. B. Peng, J. Wu, and C. K. Liu, "Twist: Teleoperated whole-body imitation system," *arXiv preprint arXiv:2505.02833*, 2025.

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

[22] Q. Ben, F. Jia, J. Zeng, J. Dong, D. Lin, and J. Pang, "Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit," *arXiv preprint arXiv:2502.13013*, 2025.

[23] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," *arXiv preprint arXiv:2401.02117*, 2024.

[24] H.-S. Fang, B. Romero, Y. Xie, A. Hu, B.-R. Huang, J. Alvarez, M. Kim, G. Margolis, K. Anbarasu, M. Tomizuka, E. Adelson, and P. Agrawal, "Dexop: A device for robotic transfer of dexterous human manipulation," *arXiv preprint arXiv:2509.04441*, 2025.

[25] J. Eßer, G. B. Margolis, O. Urbann, S. Kerner, and P. Agrawal, "Action space design in reinforcement learning for robot motor skills," in *8th Annual Conference on Robot Learning*, 2024.

[26] Z. Fu, X. Cheng, and D. Pathak, "Deep whole-body control: learning

a unified policy for manipulation and locomotion," in *Conference on Robot Learning*. PMLR, 2023, pp. 138–149.
[27] Y. Zhang, Y. Yuan, P. Gurunath, T. He, S. Omidshafiei, A.-a. Aghamohammadi, M. Vazquez-Chanlatte, L. Pedersen, and G. Shi, "Falcon: Learning force-adaptive humanoid loco-manipulation," *arXiv preprint arXiv:2505.06776*, 2025.
[28] M. H. Raibert and J. J. Craig, "Hybrid position/force control of manipulators," 1981.
[29] N. Hogan, "Impedance control: An approach to manipulation: Part ii—implementation," 1985.
[30] O. Khatib, "A unified approach for motion and force control of robot manipulators: The operational space formulation," *IEEE Journal on Robotics and Automation*, vol. 3, no. 1, pp. 43–53, 2003.
[31] L. Sentis and O. Khatib, "Synthesis of whole-body behaviors through hierarchical control of behavioral primitives," *International Journal of Humanoid Robotics*, vol. 2, no. 04, pp. 505–518, 2005.
[32] ——, "A whole-body control framework for humanoids operating in human environments," in *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.* IEEE, 2006, pp. 2641–2648.
[33] L. Sentis, J. Park, and O. Khatib, "Compliant control of multicontact and center-of-mass behaviors in humanoid robots," *IEEE Transactions on robotics*, vol. 26, no. 3, pp. 483–501, 2010.
[34] 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.
[35] C. Ott, O. Eiberger, W. Friedl, B. Bauml, U. Hillenbrand, C. Borst, A. Albu-Schaffer, B. Brunner, H. Hirschmuller, S. Kielhofer *et al.*, "A humanoid two-arm system for dexterous manipulation," in *2006 6th IEEE-RAS international conference on humanoid robots*. IEEE, 2006, pp. 276–283.
[36] J. Englsberger, A. Werner, C. Ott, B. Henze, M. A. Roa, G. Garofalo, R. Burger, A. Beyer, O. Eiberger, K. Schmid *et al.*, "Overview of the torque-controlled humanoid robot toro," in *2014 IEEE-RAS international conference on humanoid robots*. IEEE, 2014, pp. 916–923.
[37] S. Lee, P. S. Chang, and J. Lee, "Deep compliant control," in *ACM SIGGRAPH 2022 conference proceedings*, 2022, pp. 1–9.
[38] T. Portela, G. B. Margolis, Y. Ji, and P. Agrawal, "Learning force control for legged manipulation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 15 366–15 372.
[39] A. Hartmann, D. Kang, F. Zargarbashi, M. Zamora, and S. Coros, "Deep compliant control for legged robots," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 11 421–11 427.
[40] 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.
[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.
[42] C. Ott, R. Mukherjee, and Y. Nakamura, "Unified impedance and admittance control," in *2010 IEEE international conference on robotics and automation*. IEEE, 2010, pp. 554–561.
[43] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5442–5451.
[44] F. G. Harvey, M. Yurick, D. Nowrouzezahrai, and C. Pal, "Robust motion in-betweening," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 60–1, 2020.
[45] K. Zakka, "Mink: Python inverse kinematics based on mujoco," *URL https://github. com/kevinzakka/mink*, vol. 10, 2024.
[46] J. Carpentier, G. Saurel, G. Buondonno, J. Mirabel, F. Lamiraux, O. Stasse, and N. Mansard, "The pinocchio c++ library – a fast and flexible implementation of rigid body dynamics algorithms and their analytical derivatives," in *IEEE International Symposium on System Integrations (SII)*, 2019.
[47] C. Schwarke, M. Mittal, N. Rudin, D. Hoeller, and M. Hutter, "Rsl-rl: A learning library for robotics research," *arXiv preprint arXiv:2509.10771*, 2025.

APPENDIX

A. RL Hyperparameters

We train our policies using Proximal Policy Optimization (PPO) [19] implemented in the `rsl_rl` library [47]. Hyperparameters for the PPO algorithm are detailed in Table II. The policy network is an MLP with hidden layers [512, 512, 256, 128] and an ELU activation function. The critic network is an MLP with hidden layers [512, 512, 512, 512].

B. Training Environment Details

Observation Space: The policy observation space, summarized in Table III, provides proprioceptive feedback, information about the original (non-compliant) reference motion, and the commanded stiffness.

Domain Randomization: To promote robust sim-to-real transfer, we employ standard domain randomization during training, covering both robot dynamics and observation noise. The randomization ranges are specified in Table V.

Reward Function: The total reward is a weighted sum of terms designed to encourage motion tracking, compliant interaction, and physically stable behavior. The complete reward function is detailed in Table VI.

C. Compliant Motion Augmentation Details

Our offline data generation process uses a combination of procedural event sampling and inverse kinematics to create a rich dataset of feasible compliant behaviors. The goal is to produce stylistically consistent whole-body motions that correctly respond to generalized external forces at various stiffnesses. The process can be broken down into three main stages: Event Generation, IK Solving, and Feasibility Validation.

1) Event Generation: We generate two distinct types of interaction events to ensure the policy learns a versatile set of responses. Each event is defined by a target link, a time profile, and interaction parameters. During RL training, both event types (ramp and collision) are simulated by the same virtual forcefield equation [38]. The only difference between modes is the final motion path of the forcefield's origin relative to the reference motion.

a) Ramped Wrench Events: This event type is designed to simulate controlled interactions. The sampling process proceeds in the following order:

1) **Timing and Link:** An event start time is chosen after a randomized rest period (see Table IV). A target link (left or right hand) is selected uniformly.
2) **Stiffness Sampling:** The desired robot stiffness (K_{robot}, k_r) is sampled from a *log-uniform* distribution. This ensures balanced coverage of both very compliant and very stiff behaviors.
3) **Constrained Displacement Sampling:** Crucially, we do not sample force directly. Instead, we compute a valid range of Cartesian displacements for the target link. This range is constrained by the maximum allowed force and displacement limits, given the stiffness sampled in the previous step. A target displacement is then sampled uniformly from this valid range.
4) **Peak Wrench Calculation:** The peak force $\mathbf{F}_{\text{ext}}$ is calculated as the product of the sampled stiffness and displacement. Its direction is sampled uniformly on a unit sphere. The same logic applies to the peak torque $\boldsymbol{\tau}_{\text{ext}}$.
5) **Profile Timing:** A target interaction speed is sampled. This speed is used to calculate the event's ramp-up duration ($\|\text{displacement}\| / \|\text{speed}\|$), ensuring physically plausible motion. A hold duration is then sampled, defining the complete ramp-hold-ramp profile of the event.

b) Simulated Collision Events: To better emulate unexpected contact, this mode spawns a virtual collision plane in the path of the reference motion. The interaction force is generated organically from the penetration depth of the reference hand into this plane, governed by the sampled robot and environment stiffness values. The force is applied in the plane's reference frame (along its normal).

2) IK Solving: For each timestep of a generated event, we use a differential IK solver (Mink [45] with DAQP) to find a full-body configuration $\mathbf{q}_{\text{aug}}$ that satisfies the compliant objective while maintaining balance and motion style. The solver minimizes a weighted sum of cost terms, where each term corresponds to a task objective as detailed in Table IV.

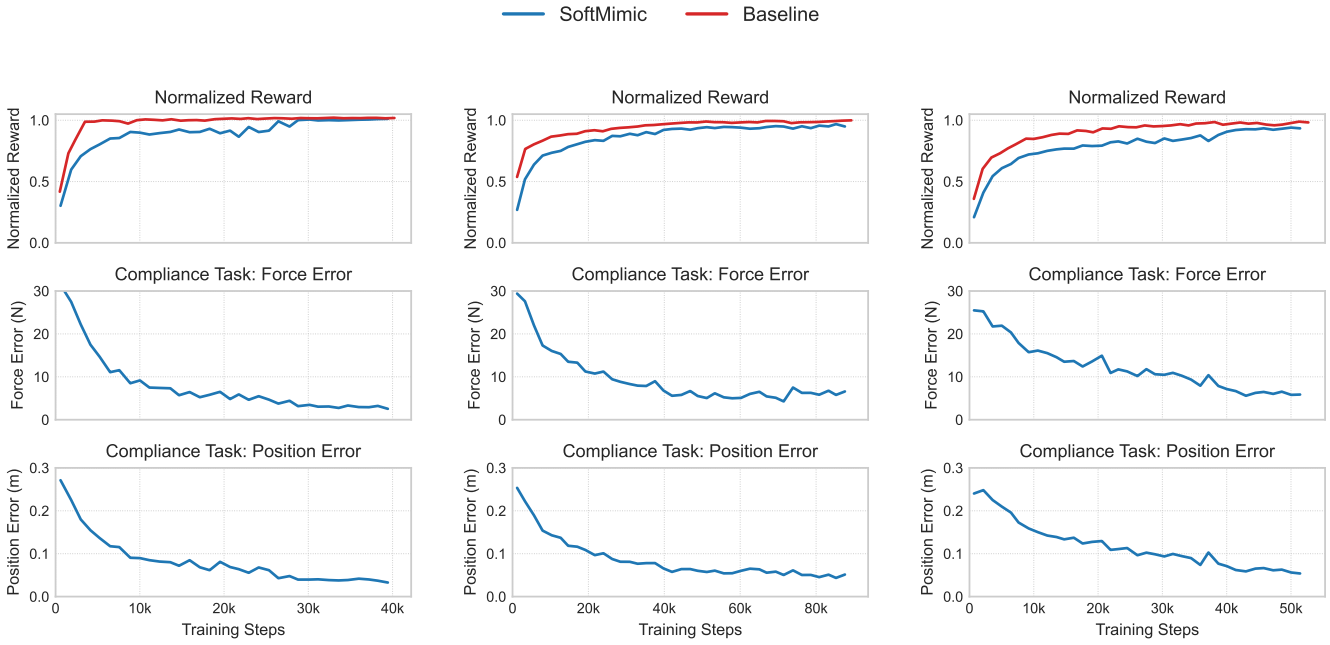
3) Feasibility Validation and Rejection Sampling: At every timestep during an event, the resulting IK solution is checked against a set of hard feasibility constraints (see Table IV). If any criterion is violated:

1) The event's magnitude is scaled down by a factor (we use 0.8). For ramped events, this means reducing the peak force; for collision events, this means shortening the event duration.
2) The entire event is re-simulated from its start time with the reduced magnitude.
3) This process repeats until the event is fully feasible or its magnitude falls below a minimum threshold (e.g., $1\,\text{N}$), at which point the event is rejected and discarded from the dataset.

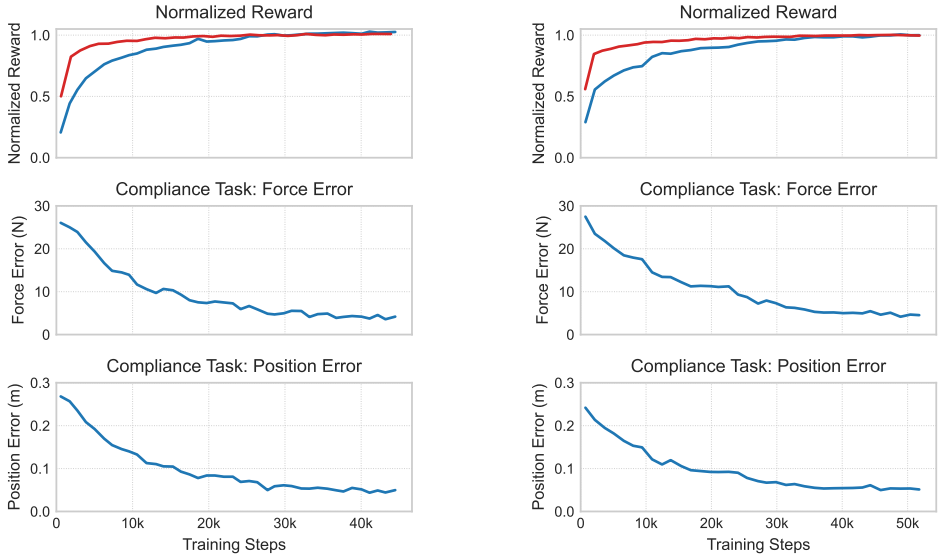
This iterative rejection sampling is critical for ensuring that the final augmented dataset D_{aug} contains only kinematically achievable and well-behaved compliant motions, simplifying the subsequent RL training problem.

D. Training Convergence Details

We render training curves for SoftMimic and the baseline in Figure 9. This illustrates the relative convergence speed of SoftMimic and the dynamics of learning force and displacement adherence over time. x-axis is number of policy update steps (each training step processing a 24-timestep rollout across 4096 parallel environments).



(a) Stand (b) Box Pick (c) T-Pose



(d) Pour (e) Walk

Fig. 9: Training performance comparison between SoftMimic and a baseline across five distinct motions. The top plot for each motion shows the normalized reward, where both policies are compared. The middle and bottom plots show force and position imitation error, respectively, for only the SoftMimic policy to highlight the training progression of its compliant behavior. Training with SoftMimic incurs modestly slower convergence while the policy learns a rich set of responses for different stiffnesses.

TABLE II: PPO hyperparameters.

Hyperparameter	Value
# Environments	4096
Timesteps per Rollout	24
Discount Factor (γ)	0.99
GAE Parameter (λ)	0.95
Learning Rate	1×10^{-3}
Schedule	Adaptive (KL target: 0.01)
Epochs per Rollout	5
Minibatches per Epoch	4
Value Loss Coefficient	1.0
Entropy Bonus	0.002
Clip Range	0.2
Max Gradient Norm	1.0
Optimizer	Adam

TABLE III: Policy observation space.

Component Group	Description
Proprioception	Joint positions (relative to default), joint velocities, base angular velocity, and projected gravity vector. A history of the last 3 timesteps is included.
Reference Motion	Reference joint positions, root height, gravity vector, base linear and angular velocity, and foot contact schedule. Includes current state, a history of 3 timesteps, and a future horizon of 20 points sampled up to 1 second ahead.
Task Commands	Logarithm of the desired translational and rotational stiffness. A history of the last 3 timesteps is included.
Action History	The previous action taken by the policy over the last 3 timesteps.

TABLE IV: **Comprehensive Data Augmentation and IK Solver Parameters.** This table details the three key components of our offline data generation pipeline. The IK solver minimizes a weighted sum of squared error norms, $\sum_i w_i \|\mathbf{e}_i\|^2$, where the error terms $\mathbf{e}_i$ are defined below. This optimization is subject to the data generation and feasibility parameters that govern the sampling of interaction events and the rejection of kinematically infeasible outcomes.

Component	Mathematical Formulation / Description	Value / Range	Units
IK Solver Objective Function Terms ($w_i \|\mathbf{e}_i\|^2$)			
Compliant Interaction	$\mathbf{e}_{\text{pos}} = (\mathbf{p}_{\text{ref}} + \mathbf{F}_{\text{ext}}/K_{\text{robot}}) - \mathbf{p}_i(\mathbf{q})$ $\mathbf{e}_{\text{rot}} = \log\left((\mathbf{R}_{\text{ref}} \exp([\boldsymbol{\tau}_{\text{ext}}/k_r]_\times))^T \mathbf{R}_i(\mathbf{q})\right)^\vee$	Weight (w_i): 5.0	-
Foot Placement	$\mathbf{e}_{\text{pos}} = \mathbf{p}_{\text{ref,foot}} - \mathbf{p}_{\text{foot}}(\mathbf{q})$ $\mathbf{e}_{\text{rot}} = \log(\mathbf{R}_{\text{ref,foot}}^T \mathbf{R}_{\text{foot}}(\mathbf{q}))^\vee$	Weight (w_i): 2.5	-
CoM Stabilization	$\mathbf{e}_{\text{CoM}} = \mathbf{c}_{\text{target}} - \mathbf{c}(\mathbf{q})$, where $\mathbf{c}_{\text{target},xy} = \mathbf{c}_{\text{ref},xy} + \frac{1}{Mg}[-\mathbf{m}_{\text{ext},y}, \mathbf{m}_{\text{ext},x}]$ ($\mathbf{m}_{\text{ext}}$ is total moment about CoP)	Weight (w_i): 0.1	-
Keypoint Posture	Tracks Cartesian poses of key links (torso, elbows, knees) against the reference motion $\mathbf{q}_{\text{ref}}$.	Weight (w_i): 0.01	-
Joint Posture	$\mathbf{e}_{\text{joint}} = \mathbf{q}_{\text{ref}} - \mathbf{q}$	Weight (w_i): 10^{-4}	-
Data Generation Hyperparameters			
Robot Stiffness (Linear)	Log-uniform sampling of commanded robot stiffness K_{robot}.	$[10, 1000]$	$\mathrm{N\,m^{-1}}$
Robot Stiffness (Angular)	Log-uniform sampling of commanded robot stiffness k_r.	$[0.1, 10]$	$\mathrm{N\,m\,rad^{-1}}$
Environment Stiffness (Linear)	Stiffness of the virtual force field or collision plane.	$[10, 1000]$	$\mathrm{N\,m^{-1}}$
Environment Stiffness (Angular)	Rotational stiffness of the virtual force field.	$[0.1, 10]$	$\mathrm{N\,m\,rad^{-1}}$
Peak Force Limit	Hard constraint on max peak force $\|\mathbf{F}_{\text{ext}}\|$.	140	N
Peak Torque Limit	Hard constraint on max peak torque $\|\boldsymbol{\tau}_{\text{ext}}\|$.	10	N m
Displacement Limit	Hard constraint on max resulting displacement.	0.7	m
Ang. Displacement Limit	Hard constraint on max resulting angular displacement.	2.0	rad
Time Between Events	Randomized rest period between interaction events.	$[0.5, 1.5]$	s
Event Hold Duration	Duration of the peak force/torque application.	$[0.5, 1.0]$	s
Target Interaction Speed	Sampled velocity used to calculate force ramp duration.	$[0.1, 1.0]$	$\mathrm{m\,s^{-1}}$
Feasibility Rejection Criteria			
Max Link Tracking Error	Maximum deviation of the solved hand pose from its compliant target pose.	Threshold: 0.05	m
Max Stance Foot Displacement	Maximum deviation of solved stance foot poses from their reference poses.	Threshold: 0.05	m
Max CoM Tracking Error	Maximum deviation of the solved CoM from its CoP-aware target in the XY-plane.	Threshold: 0.15	m

TABLE V: Domain randomization ranges.

Parameter	Range	Units
Dynamics Randomization (per episode)		
Payload Mass (added to torso)	$[-2.0, 2.0]$	kg
Link Mass Scale	$[0.7, 1.3]$	-
Base CoM Displacement (XYZ)	$[-0.02, 0.02]$	m
Joint Damping (added)	$[0, 2]$	$\mathrm{N\,m\,s\,rad^{-1}}$
Joint Armature (added)	$[0.01, 0.1]$	$\mathrm{kgm^2}$
Joint Friction (added)	$[0, 0.01]$	N m
Ground Static/Dynamic Friction	$[0.5, 2.0]$	-
Ground Restitution	$[0.0, 0.5]$	-
Observation Noise (per step)		
Joint Position Noise	$[-0.01, 0.01]$	rad
Joint Velocity Noise	$[-1.5, 1.5]$	$\mathrm{rad\,s^{-1}}$
Base Angular Velocity Noise	$[-0.2, 0.2]$	$\mathrm{rad\,s^{-1}}$
Projected Gravity Noise	$[-0.01, 0.01]$	-

TABLE VI: Reward terms and weights.

Term Description	Weight
Compliance Rewards	
Force Link Position Tracking (vs. augmented ref)	3.0
Force Link Orientation Tracking (vs. augmented ref)	3.0
Applied Force Tracking (vs. desired command)	2.0
Applied Torque Tracking (vs. desired command)	2.0
Motion Tracking Rewards	
Keypoint Position Tracking (vs. augmented ref)	2.0
Keypoint Orientation Tracking (vs. augmented ref)	2.0
Base Orientation Tracking (vs. augmented ref)	0.5
Base Linear Velocity Tracking (vs. augmented ref)	0.5
Base Angular Velocity Tracking (vs. augmented ref)	0.5
Stability and Regularization Rewards	
Alive	1.5
Joint Position Limits Penalty	-10.0
Stance Foot Stability (sliding penalty)	-0.005
Joint Velocity L2 Penalty	-2.8e-4
Action Rate L2 Penalty	-0.01
Stance Foot Joint Motion Penalty	-0.4