2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Learning Force Control for Legged Manipulation

Tifanny Portela[12], Gabriel B. Margolis[1], Yandong Ji[13], and Pulkit Agrawal[1]

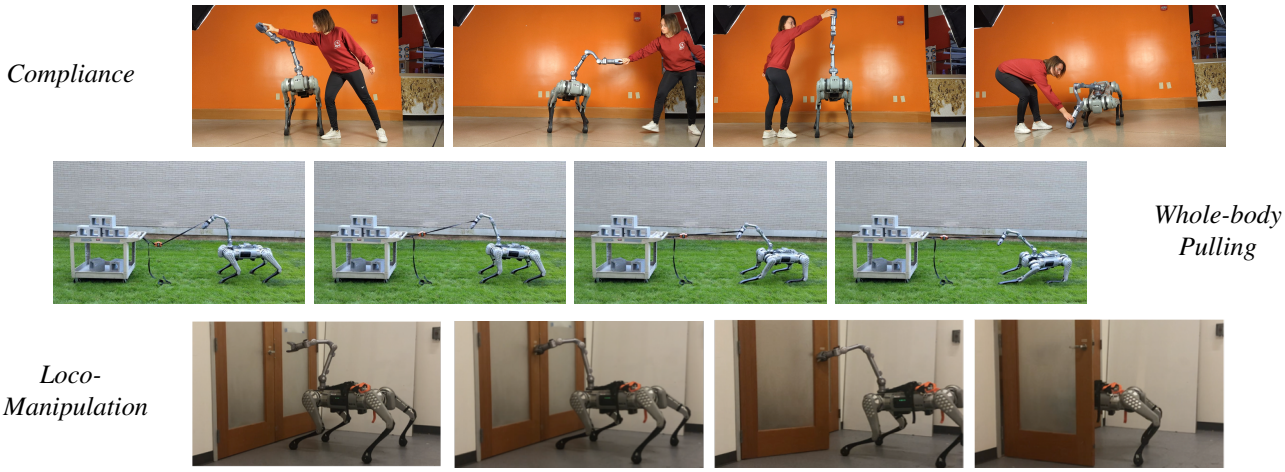


Fig. 1: We train a whole-body policy to control *contact force* as well as for end effector position control in *noncontact* scenarios. The force control mode enables compliant interaction (top) and whole-body force application (middle). The mobile end effector position and force control modes comprise a pipeline for teleoperation of locomotion, grasping, and force-controlled manipulation.

***Abstract*— Controlling the contact force during interactions is an inherent requirement for locomotion and manipulation tasks. Current reinforcement learning approaches to locomotion and manipulation rely implicitly on forceful interaction to accomplish tasks but do not explicitly regulate it. This paper proposes a reinforcement learning task specification that focuses on matching desired contact force levels. Integrating force control with the coordination of a robot's body and arm, we present an end-to-end policy for legged manipulator control. Force control enables us to realize compliant gripper and whole-body pulling movements that have not been previously demonstrated using a learned policy. It also facilitates a characterization of the force-tracking performance of learned policies in simulation and the real world, indicating their performance potential for force-critical tasks. Video is available at the project website: `https://tif-twirl-13.github.io/learning-compliance`.**

I. Introduction

Recent frameworks for training policies using reinforcement learning have achieved state-of-the-art results in legged locomotion [1]–[4], dexterous manipulation [5], [6], and drone flight [7]. The common paradigm in these works is to command a task in position or velocity space while leaving the policy to choose applied force implicitly. Compared to other forceful manipulation systems [8]–[12], typically learned policies use a coarse action space of position targets for a low-gain PD controller, a relatively low control frequency, and do not incorporate a force-torque sensor on the contact body. Regardless, reinforcement learning controllers are frequently used to generate forceful interactions such as during walking, running, parkour, fall recovery, object reorientation, and others [13]–[18]. In this work, we propose a task specification for reinforcement learning where the desired contact force is directly commanded. Leveraging this specification, we train a policy capable of applying the desired force through a manipulator mounted on a large quadruped. We characterize the force tracking and estimation capability of such a system. We also investigate the ability of the quadruped to coordinate the body and legs to increase both its reach and the force application workspace and magnitude compared to using the manipulator in isolation.

While force control is well studied for fixed-base and wheeled manipulators [19]–[23], legs can allow us to walk to an object that may be inaccessible to a wheeled robot and also allow shifting the center of mass to facilitate a larger workspace and stronger posture. Our work trains a legged manipulator to perform tasks that involve force control, ranging from a compliant mode for kinesthetic demonstration collection or safe operation around humans to high-force tasks like pulling and lifting. This is achieved by coordinating the entire body with an end-to-end policy. Learning a whole-body policy allows the robot to comply and apply force in configurations that would be outside a fixed-base manipulator's workspace. The policy is simultaneously trained for multiple tasks: mobile reaching and force control, allowing a seamless transition between modes. Our primary contributions are:

1) Propose a task specification for learning end-to-end

Research was conducted in the Improbable AI Lab at MIT. Author affiliations: [1] Improbable AI Lab. [2] EPFL. [3] University of California, San Diego.

policies for end effector force control (see Section IV).
2) Demonstrate the first deployment of learning-based whole-body force application control in legged manipulators (see Section V).
3) Characterize the effectiveness of force tracking and force estimation for learned policies in simulation and reality (Section V).

II. Related Work

A. Reinforcement Learning for Loco-Manipulation

Prior work has controlled a quadruped with mounted arm using sim-to-real reinforcement learning by formulating the task as simultaneous end effector position control and locomotion velocity control [13], [24]. Our work is highly influenced by [13], which was deployed on a real robot and displayed an increased workspace. However, it was not suitable for forceful and compliant tasks due to several factors, including the lack of force modeling during training, the competing constraint arising from explicitly commanding the gripper position relative to the body, and the smaller arm with a maximum payload rating of $0.2\,\mathrm{kg}$. Other works have learned a policy for a downstream task that requires forceful interaction, such as preventing and recovering from falls with a mounted arm [15] or dribbling a ball with the feet [14], [25]. These works demonstrate that forceful interaction is possible but optimize it in service of a single task. Without an explicit way of commanding the interaction force, we cannot repurpose these controllers for other force application tasks, and the precision of the force control in simulation and reality cannot be directly evaluated. Another line of work has used hierarchical architecture to push large objects using the robot's body, without an arm [26]. While the body and legs alone can perform useful environment interactions, a mounted arm can increase the workspace and allow the robot to control the direction and magnitude of force and torque application more precisely.

B. Forceful Control Primitives

For some manipulation tasks, particularly those involving contact interactions, it is hard to teleoperate by commanding position setpoints due to the unknown geometry of the contact surface or a desired degree of compliance in the motion. Instead, various parameterizations of control that regulate contact force have been proposed. Classic work on compliance and force control [19]–[22] established impedance control and hybrid force-velocity control. The purpose of these techniques is to define how the robot should reconcile force and position tracking errors given that these quantities cannot be independently controlled during contact. Force control methods may be implemented with closed-loop feedback from a wrist force sensor or with less precision by estimating the contact force from proprioceptive actuators and robot model [8]. Our approach does not strictly correspond to any classical control formulation because our objective function combines regularization terms common in reinforcement learning with a force tracking objective.

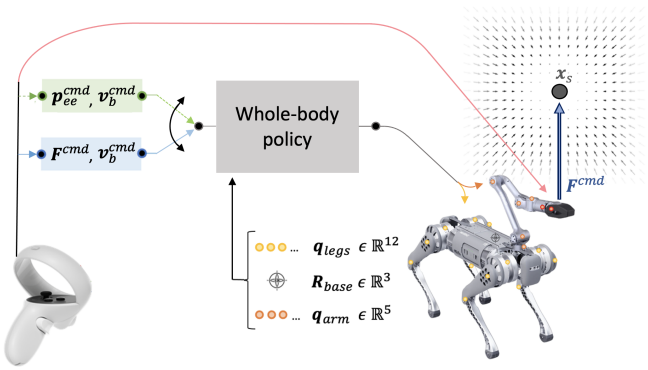


Fig. 2: **System architecture.** In addition to the position command $\mathbf{p}_{ee}^{cmd}$, we consider training a policy to track a commanded end effector force $\mathbf{F}^{cmd}$ while walking following a base velocity command $\mathbf{v}_b^{cmd}$. A human operator uses the Oculus joystick (left) to define the commands. To learn force tracking, we train the policy in a potential field where the gripper is pulled towards a randomized setpoint x_s (right).

One use case for force control is kinesthetic teaching, wherein a human physically interacts with a robot to demonstrate the movements and forces it should apply on the environment to accomplish a task [27], [28]. To receive a kinesthetic demonstration, it is necessary for the robot to comply with the operator's guidance and also estimate the amount of force exerted. This method of teaching has previously been considered challenging for legged robots and high-dimensional systems because of the cognitive challenge of reasoning about many degrees of freedom [29], which provides a motivation for developing low-level controllers that assume some control authority and ease the task.

C. Whole-body Control

Effective model-based approaches have been developed for controlling legged manipulators in a variety of tasks [9]–[12], [30]–[35]. Such works typically optimize a whole-body inverse dynamics model to stabilize a reference trajectory computed using model-predictive control or trajectory optimization. A few works have specialized in forceful interaction with a legged manipulator [9]–[12]. Early approaches [9] used trajectory optimization to generate open-loop behaviors for manipulating heavy objects and a separate online tracking controller to realize the motion. Subsequent work [10] incorporated contact point planning and control to manipulate objects using the entire body of a humanoid robot. More recently, ALMA [12] proposed a motion planning and control framework capable of coordinating dynamic and compliant locomotion and mobile manipulation based on a whole-body control task hierarchy. They demonstrated accurate contact force control through the end effector using the whole body. The ALMA system was also shown to support compliant behavior in service of a collaborative payload-carrying task.

III. Materials

Robot Quadruped: We implement our method on the Unitree B1 quadruped robot. This $55\,\mathrm{kg}$ robot stands $0.64\,\mathrm{m}$

tall. It has 12 identical electric actuators – each equipped with a joint position encoder – and an inertial measurement unit in its body to provide orientation. An onboard NVIDIA Jetson Xavier NX computer runs the control policy at 50 Hz.

Robot Arm: A Unitree Z1 arm is mounted on the B1's base. This 6 degree-of-freedom robot arm weighs 5.3 kg and has a maximum reach of 0.74 m.

Teleoperation Joystick: To control the robot and arm, we use the Oculus Meta Quest 2 headset with its two controllers (Figure 2). The motion of the right controller is mapped to position commands for the end effector. The robot's body motion, force application, and position commands at the end effector are controlled by the thumb joysticks and buttons on both controllers. The last two arm joints, responsible for controlling the gripper's roll orientation and opening angle, are directly commanded via joystick commands, as they do not influence the force or position of the end effector.

Simulator: We use Isaac Gym [36] for policy training.

IV. Method

Our system is comprised of a single learned policy that simultaneously learns locomotion and whole-body regulation of the end effector position and force. The force tracking (Section IV-D) is the main novelty of our work and enables compliant and force-controlled tasks. In order to apply forces, the robot first needs to walk to and grasp objects, so the robot alternates learning force regulation and position tracking (Section IV-E).

A. Action and Observation Space

The action space is seventeen-dimensional ($a_t \in \mathbb{R}^{17}$), controlling position targets for a proportional-derivative controller in each of the robot's joints: thigh, calf, and hip joints of the B1 robot as well as the first five joints of the arm. The position targets are computed as $\sigma_a a_t + q_{def}$, where $\sigma_a = 0.25$ is a scaling factor, a_t is the policy's output, and q_{def} denotes the default joint configuration, reflecting the robot's standard standing pose with its arm raised.

The observation, denoted as o^t, consists of the gravity vector projected in the body frame $g^t_{base} \in \mathbb{R}^3$, the feet clock timings $c^t_{feet} \in \mathbb{R}^4$ [37] , the joint positions, $q^t \in \mathbb{R}^{17}$, the joint velocities, $\dot{q}^t \in \mathbb{R}^{17}$, the actions, $a^t \in \mathbb{R}^{17}$ and the previous actions $a^{t-1} \in \mathbb{R}^{17}$:

$$o^{t+1} = [g^t_{base}, c^t_{feet}, q^t, \dot{q}^t, a^t, a^{t-1}] \in \mathbb{R}^{75} \quad (1)$$

The observation history $o^{[t-H,..,t-1,t]}$ (H=30) is concatenated to the history of task-associated commands $t^{[t-H,..,t-1,t]}_{cmd}$ as input to the policy.

B. Policy Architecture and Optimization

The policy itself includes the actor and state estimation modules, while the entire system comprises three modules: the actor network, the critic network, and the estimation module. Following a common optimization technique in sim-to-real learning [38], [39], we define a privileged state consisting of quantities that may aid learning in simulation but are not available from the real-world sensors. Our privileged state includes the robot's body velocity, the gripper position in the body frame, and the external force on the gripper. First, the state estimation module predicts an estimate $\hat{e}_t$ approximating the privileged state e_t from the observation history. This style of concurrent state estimation can improve optimization performance [39]. Then, the actor-network inputs the observation history and the state estimate $\hat{e}_t$ and outputs the action. Separately, the critic network inputs the observation history and the true privileged state e_t. The estimator, actor, and critic network are multilayer perceptions with `elu` nonlinearities and hidden layer dimensions $[256, 128]$, $[512, 256, 128]$, and $[512, 256, 128]$, respectively. The state estimator is simultaneously trained with a supervised loss while the actor and critic networks are optimized using Proximal Policy Optimization [40] with 4096 parallel environments. The maximum episode length is 20 seconds which corresponds to 1000 timesteps. To speed up training, episodes are terminated if the agent reaches a failure state. The robot is deemed to fail if the gripper is in collision or if the body height falls below 0.3 m.

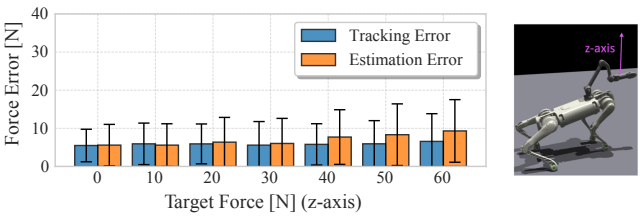


Fig. 3: **Simulated force control evaluation.** Average force tracking and estimation errors for target forces applied at 1000 setpoints sampled across the full training workspace. The bars represent the standard deviation. The accompanying illustration (right) shows how the robot applies force at the gripper along the z-axis.

C. Task Definition

The policy inputs t_{cmd} consisting of four commands:

- The desired force vector at the end effector $F^{\text{cmd}} \in \mathbb{R}^3$, is defined in Cartesian coordinates in the world frame and sampled between -70 N and 70 N for each axis.
- The desired end effector position in the body frame is expressed in spherical coordinates: $p^{\text{cmd}}_{ee} = (r, \theta, \phi)^{cmd} \in \mathbb{R}^3$, sampled from $r^{\text{cmd}} \in$ [0.3, 0.9 m], $\theta^{\text{cmd}} \in$ [-0.4π , 0.4π rad], $\phi^{\text{cmd}} \in$ [-0.6π , 0.6π rad].
- The desired linear velocity in the body frame x- and y- axes and the yaw angular velocity: $v^{\text{cmd}}_b \in \mathbb{R}^3$ are sampled from $v_x \in [-1, 1\,\text{m/s}]$, $v_y \in [-1, 1\,\text{m/s}]$, $\omega_z \in [-1, 1\,\text{rad/s}]$. To make teleoperation easier, when the magnitude of the commanded base velocity falls below 0.1, the desired contact schedule for each foot is set to zero, ensuring all feet remain grounded. This prevents the robot from stepping in place, easing teleoperation.
- The binary input C^f indicates whether the robot is in position or force tracking mode. This input is resampled randomly twice within each episode so the environments alternate between force and position tracking.

The reward is $(\mathbf{r}^b_v + \mathbf{r}^g_x + \mathbf{r}^g_f)e^{\mathbf{r}_t}$ with separate terms for the locomotion task ($\mathbf{r}^b_v$), gripper force control task ($\mathbf{r}^g_f$),

gripper position control task ($\mathbf{r}_x^g$), and safety and smoothness criteria ($\mathbf{r}_l$) (Table I). To encourage a smooth gait, the locomotion task includes a contact schedule pattern for trotting [37]. The exponential form from [39] ensures the reward will remain positive by lifting the negative penalty terms into the exponential.

During the force control task ($\mathbf{C}^f = 1$), the position commands p_{ee}^{cmd} and position tracking rewards $\mathbf{r}_x^g$ are set to zero. Conversely, during the end effector positioning task ($\mathbf{C}^f = 0$), the force commands F^{cmd} and force tracking rewards $\mathbf{r}_f^g$ are set to zero.

D. End Effector Force Task

To learn force control without learning grasping as a prerequisite, we emulate the applied force on the gripper using a soft contact model. To do so, we simulate an external force $F_e \in \mathbb{R}^3$ that pulls on the gripper as a function of its displacement from a force setpoint. The force is modulated using a proportional-derivative control scheme on the position and velocity difference between the gripper position, $\mathbf{x}_g^t \in \mathbb{R}^3$ and the setpoint $\mathbf{x}_s \in \mathbb{R}^3$:

$$F_e^t = K_p(\mathbf{x}_s - \mathbf{x}_g^t) + K_d(\dot{\mathbf{x}_s} - \dot{\mathbf{x}_g}^t) \tag{2}$$

The chosen gains, $K_p = 700$, $K_d = 6$, were tuned by inspecting the behavior of the simulator to ensure that the applied force is large enough to bring the gripper to the setpoint but small enough to avoid extreme torques or oscillations. The setpoint is defined with respect to the robot's body frame and moves with the robot during locomotion.

If we initialize the force setpoint randomly, it may be unreachable or in collision with the robot's body. To avoid this, we instead initialize the force setpoint at the gripper's initial position when alternating between the position tracking task and the force tracking task. As the robot learns to reach end effector positions across a wide workspace, it simultaneously acquires proficiency in exerting force across that workspace.

In terms of command sampling, we repeatedly sample the force target F^{cmd} and the duration of the force application t_F, which ranges between two and four seconds. The force command undergoes linear interpolation from zero to the sampled values in t_F seconds. Then, they maintain these values for a duration of 1.5 seconds, after which they are linearly interpolated back to zero in t_F seconds. To ensure that fully compliant behavior is experienced frequently, each time a force command is sampled, there is a 20% chance of being a zero force target.

E. End Effector Positioning Task

The end effector position control mode is a stepping stone towards our primary goal of forceful manipulation. This mode (i) supports walking to objects and grasping them (Section V-C), and (ii) generates diverse stable and collision-free initialization poses for the force mode (IV-D). This mode should have a wide workspace and intuitive teleoperation interface for reaching any point in the environment. This control mode reimplements elements from prior work [13] with some adaptation to suit our robot's specifications and facilitate alternation with force training.

Term	*Equation*	*Weight*
End Effector Position Control (IV-E)		
$\mathbf{r}_x^g$: gripper position	$\exp\{-\lvert p_{ee} - p_{ee}^{\text{cmd}}\rvert/0.5\}$	$5.0 * \neg C^f$
End Effector Force Control (IV-D)		
$\mathbf{r}_f^g$: gripper force	$\exp\{-\lvert F - F^{\text{cmd}}\rvert/20\}$	$5.0 * C^f$
Safety and Smoothness ($\mathbf{r}_l$)		
collision penalty	$\mathbb{1}_{\text{collision}}$	-5.0
arm joint limit	$\mathbb{1}_{q_a > 0.9 * q_a^{max} \lvert\rvert q_a < 0.9 * q_a^{min}}$	-3.0
leg joint limit	$\mathbb{1}_{q_l > 0.9 * q_l^{max} \lvert\rvert q_l < 0.9 * q_l^{min}}$	-1.0
joint velocities	$\lvert\dot{q}\rvert^2$	$-8\text{e}{-4}$
joint acceleration	$\lvert\ddot{q}\rvert^2$	$-3\text{e}{-7}$
action smoothing	$\lvert a_{t-1} - a_t\rvert$	-0.05
–	$\lvert a_{t-2} - 2a_{t-1} + a_t\rvert^2$	-0.02
arm torque limit	$\mathbb{1}_{\tau_a > 0.8 * \lvert\tau_a^{max}\rvert}$	-0.0015
Locomotion ($\mathbf{r}_v^b$)		
body velocity	$\exp\{-\lvert v_b - v_b^{\text{cmd}}\rvert/0.25\}$	1.0
swing phase	$\sum_{\text{foot}}[1 - C_i^{cmd}(t)] \exp\{-\lvert f^{\text{foot}}\rvert^2/4\}$	0.9
stance phase	$\sum_{\text{foot}}[C_i^{cmd}(t)] \exp\{-\lvert v_{xy}^{\text{foot}}\rvert^2/4\}$	4.0

TABLE I: Reward terms for learning the whole-body policy.

We adopt the command parameterization from [13], where the end effector commands are defined within a frame centered around the body and remain invariant to body height, roll, or pitch changes. Each time a new position is sampled, we linearly interpolate the position target from the previous position command to the new one for four seconds. The robot is initialized in a standing posture (shown in Fig. 5) with a small random perturbation to the joint angles.

V. Results

We evaluate the efficacy of the learned policy at end effector force tracking for compliant and forceful tasks. We also characterize the performance of end effector position tracking and show that it is sufficient to complete our system for forceful teleoperation.

A. Applying Large Forces

When lifting or pulling objects, an effective controller should realize a large applied force without undesired transients or inefficient postures that would result in the motor exceeding its safety limits. To evaluate this characteristic, the average force tracking and estimation errors for target forces applied at 1000 setpoints sampled across the full training workspace are reported in Figure 3. The tracking error represents the difference between the commanded and actual force, while the estimation error denotes the difference between the estimated and actual force. For z-axis force control, the mean absolute tracking error is around $5\,\text{N}$ for low force targets and remains below $10\,\text{N}$ across the entire training range. The estimation error is generally equal to or greater than the tracking error, suggesting that some errors may result from a lack of observability in the policy input.

To evaluate the force control performance on the real platform, we attach our robot's end effector to a dynamometer and gradually increase the downward force command

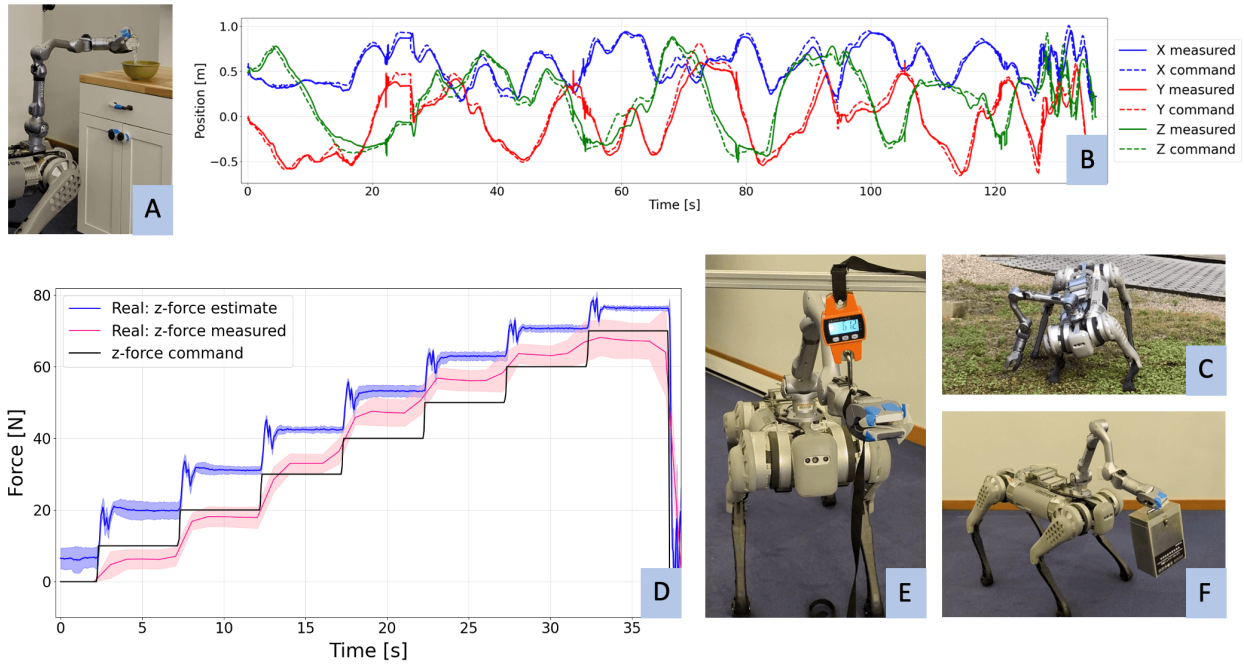


Fig. 4: **Real-world deployment.** We deploy our policy in a variety of real-world scenarios: (A) water pouring in a kitchenette; (B) tracking a long end effector trajectory with performance verified by motion capture; (C) outdoor whole-body reaching; (D, E) quantitative measurement of the force application control by pulling down on a dynamometer; (F) lifting a 4 kg box while maintaining compliance in the gripper, showcasing gravity compensation. In subfigure (D), each line represents the mean across five force setpoints, with the shaded area indicating the standard deviation.

through the entire training range (0-70 N). The dynamometer is used to record the applied force across five trials with the gripper in high, low, left, right, and middle positions (Figure 4-E). We found that the tracking error across these five trials was within the range of 5-10 N which is similar to the values observed in simulation. The estimated force tends to overshoot, suggesting a moderate sim-to-real gap. Despite significant discrepancies in force estimation, the force tracking performance remains notably proficient, suggesting that the policy somewhat disregards the estimated force feedback.

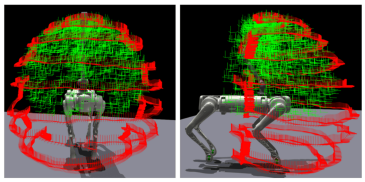
To evaluate whether our policy can coordinate the body with the legs to increase the applied force, we initialized the arm directly in front of the robot and recorded the highest pulling force we can achieve (Figure 1). The observed force of 90 N is greater than the arm's rated payload of 36 N. We also measured the force application across the reaching workspace in simulation. We found that the policy can track large forces across a large portion of the expanded workspace, although it experiences higher errors at the extreme limits of reaching.

B. Applications: Compliance and Kinesthetic Demonstrations

1) Compliant end effector state: When the controller IV-D is commanded to track zero force, this corresponds to a fully compliant mode where the posture of the body and arm coordinate to drive the gripper force application to zero in all three axes. When released, the gripper remains suspended, with the system exhibiting gravity compensation. Figure 1 and the supplementary video show an operator manipulating the system to reach various points.

Our result supports that fully compliant behavior is possible using the standard reinforcement learning architecture, which has not been previously shown. Realizing compliant behavior in learned motor policies will likely improve the safety of robots around humans and their robustness against unexpected disturbances. For example, when the quadruped walks in force control mode and bumps its gripper into a wall or obstacle, we observe that the gripper complies and moves out of the way, allowing the robot to continue walking as desired. Compliance also facilitates data collection for kinesthetic teaching, in which a human operator directly manipulates a robot's limbs to demonstrate a task without writing code or learning to use a teleoperation interface. As a proof of concept, we kinesthetically manipulated the robot to insert a drill into its charging station.

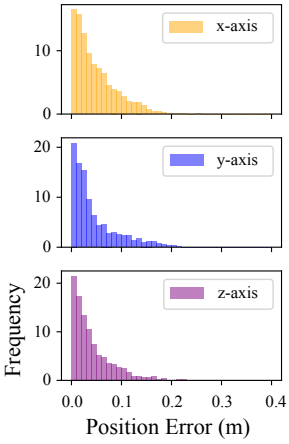
2) Compliant manipulation of heavy objects: By modulating the force application command in the z-axis, the compliant mode can be extended to the scenario where the robot is lifting an object with its arm (Figure 4-F). We tested the compliant mode with payloads of 0 kg, 2 kg, 4 kg, 6 kg. With a zero force command along the z-axis, nonzero payloads cause the gripper to sink to the ground, which is expected since it should not apply a resistive force. In this case, it takes substantial force for the human to lift the gripper with the object in grasp because the gripper will not apply any lifting force to the grasped object. Next, we increased the vertical force command to match the payload weight. For payloads



(a) Reaching workspace

Controller type	Workspace [m^3]
Arm	0.81
Whole body	1.29

(b) Workspace convex hull volumes



(c) Position error histogram

Fig. 5: **Expanded workspace.** (A-B) The whole-body policy (red) enlarges the fixed arm workspace (green) by 59%. (C) The position error distribution shows the tracking accuracy while standing in place, evaluated across 1000 end effector positions sampled from the training distribution.

up to $4\,\mathrm{kg}$, this restored gravity compensation against the payload and compliance in all axes. With the highest payload of $6\,\mathrm{kg}$, the gripper is less compliant and drifts to the center of the robot when released to the side.

C. Applications: Walking, Reaching, and Grasping

The end effector position control mode serves to establish an initial grasp or contact with the force application target. To manipulate objects that are higher or lower than the robot's body, we would like to achieve a large manipulation workspace through coordination of the body and arm.

To measure the workspace of the position reaching, we send a sequence of commands at the limit of the training distribution and record the actual gripper positions achieved. As a baseline, we measure the workspace of the arm while the base is fixed in place by randomly sampling 3000 arm configurations and recording the resulting gripper positions. To characterize the workspace in each scenario, we fit a convex hull to the reached points and compute its volume. The whole-body policy enlarges the workspace volume by $59\,\%$, demonstrating the significant benefit of whole-body coordination (Figure 5).

To grasp objects for manipulation, the end effector position tracking mode should also be accurate. We measure the distribution of error between the end effector position and the target position across a set of 1000 end effector position commands, uniformly sampled from the training distribution. The mean error is $4.6\,\mathrm{cm}$ in the x-axis, $4.8\,\mathrm{cm}$ in the y-axis, $5.5\,\mathrm{cm}$ in the z-axis.

The end effector tracking performance is evaluated on hardware by teleoperating the arm across a variety of commands within the training distribution. We record the trajectory of commands as well as the trajectory of end effector positions captured via a motion-tracking system. The path, expressed in Cartesian coordinates, can be seen in Figure 4-B. As shown, the commanded and actual end effector positions align closely. For this trajectory, the average positional errors on the x, y, and z axes are 4.42 cm, 5.37 cm, and 6.86 cm, respectively. These values being similar to the ones observed in simulation, the disparity between simulation and real-world performance in the position mode is minimal. In a series of qualitative experiments, we successfully teleoperated the end effector positioning mode for door opening (Figure 1) and water pouring (Figure 4-A).

VI. Discussion

We have demonstrated that learned whole-body manipulation policies can acquire a degree of compliance and perform force control at the end effector using only the minimal sensor configuration of joint encoders and body IMU. The force tracking performance is sufficient for some force-controlled tasks, including collecting demonstrations for kinesthetic teaching with a weight and whole-body pulling. In quantitative experiments, we demonstrate good accuracy of a learned force estimator and the force application command tracking. We also characterize the sim-to-real gap in our system. Because the policy can walk, grasp, and apply forces, our work provides a teleoperation framework suitable for teleoperation and data collection for kinesthetic teaching of forceful loco-manipulation tasks. In the future, it will be promising to explore imitation learning pipelines where the robot learns to accomplish compliant and forceful behaviors autonomously from demonstrations.

Our approach differs from some successful classical methods by realizing force control through a low-frequency neural network policy and without any force-torque sensor at the end effector. Although we showed that our policy can realize desired forces and positions, the achieved tracking accuracy of around $5\,\mathrm{cm}$ for position and $5\,\mathrm{N}$ for force commands may be insufficient for highly precise tasks. It is unknown whether this performance represents a limit of the hardware, optimization method, or choice of policy architecture. Future work could explore incorporating kinematic and dynamic models to guide the policy to a better solution.

Acknowledgement

The contributions of all authors are listed at the project website. We thank the members of the Improbable AI lab for helpful discussions and feedback. We are grateful to MIT Supercloud and the Lincoln Laboratory Supercomputing Center for providing HPC resources. This research was partly supported by Hyundai Motor Company, the DARPA Machine Common Sense Program, the MIT-IBM Watson AI Lab, and the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/). This research was also sponsored by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-21-1-0328. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes, notwithstanding any copyright notation herein.

REFERENCES

[1] 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.

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

[3] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," in *Conference on Robot Learning*. PMLR, 2022, pp. 91–100.

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

[5] T. Chen, J. Xu, and P. Agrawal, "A system for general in-hand object re-orientation," in *Conference on Robot Learning*. PMLR, 2022, pp. 297–307.

[6] T. Chen, M. Tippur, S. Wu, V. Kumar, E. Adelson, and P. Agrawal, "Visual dexterity: In-hand reorientation of novel and complex object shapes," *Science Robotics*, vol. 8, no. 84, p. eadc9244, 2023.

[7] E. Kaufmann, L. Bauersfeld, A. Loquercio, M. Müller, V. Koltun, and D. Scaramuzza, "Champion-level drone racing using deep reinforcement learning," *Nature*, vol. 620, no. 7976, pp. 982–987, 2023.

[8] S. Chiaverini, B. Siciliano, and L. Villani, "A survey of robot interaction control schemes with experimental comparison," *IEEE/ASME Transactions on mechatronics*, vol. 4, no. 3, pp. 273–285, 1999.

[9] M. P. Murphy, B. Stephens, Y. Abe, and A. A. Rizzi, "High degree-of-freedom dynamic manipulation," in *Unmanned Systems Technology XIV*, vol. 8387. SPIE, 2012, pp. 339–348.

[10] M. Murooka, S. Nozawa, Y. Kakiuchi, K. Okada, and M. Inaba, "Whole-body pushing manipulation with contact posture planning of large and heavy object for humanoid robot," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 5682–5689.

[11] B. U. Rehman, M. Focchi, J. Lee, H. Dallali, D. G. Caldwell, and C. Semini, "Towards a multi-legged mobile manipulator," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 3618–3624.

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

[13] 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.

[14] Y. Ji, G. B. Margolis, and P. Agrawal, "Dribblebot: Dynamic legged manipulation in the wild," *arXiv preprint arXiv:2304.01159*, 2023.

[15] Y. Ma, F. Farshidian, and M. Hutter, "Learning arm-assisted fall damage reduction and recovery for legged mobile manipulators," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 12 149–12 155.

[16] D. Hoeller, N. Rudin, D. Sako, and M. Hutter, "Anymal parkour: Learning agile navigation for quadrupedal robots," *arXiv preprint arXiv:2306.14874*, 2023.

[17] Z. Zhuang, Z. Fu, J. Wang, C. Atkeson, S. Schwertfeger, C. Finn, and H. Zhao, "Robot parkour learning," in *Conference on Robot Learning (CoRL)*, 2023.

[18] X. Cheng, K. Shi, A. Agarwal, and D. Pathak, "Extreme parkour with legged robots," *arXiv preprint arXiv:2309.14341*, 2023.

[19] M. T. Mason, "Compliance and force control for computer controlled manipulators," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 11, no. 6, pp. 418–432, 1981.

[20] M. H. Raibert and J. J. Craig, "Hybrid position/force control of manipulators," 1981.

[21] T. Yoshikawa, "Dynamic hybrid position/force control of robot manipulators–description of hand constraints and calculation of joint driving force," *IEEE Journal on Robotics and Automation*, vol. 3, no. 5, pp. 386–392, 1987.

[22] N. Hogan, "Impedance control: An approach to manipulation," in *1984 American control conference*. IEEE, 1984, pp. 304–313.

[23] ——, "Impedance control: An approach to manipulation: Part ii—implementation," 1985.

[24] S. Lee, S. Jeon, and J. Hwangbo, "Learning legged mobile manipulation using reinforcement learning," in *International Conference on Robot Intelligence Technology and Applications*. Springer, 2022, pp. 310–317.

[25] T. Haarnoja, B. Moran, G. Lever, S. H. Huang, D. Tirumala, M. Wulfmeier, J. Humplik, S. Tunyasuvunakool, N. Y. Siegel, R. Hafner *et al.*, "Learning agile soccer skills for a bipedal robot with deep reinforcement learning," *arXiv preprint arXiv:2304.13653*, 2023.

[26] S. Jeon, M. Jung, S. Choi, B. Kim, and J. Hwangbo, "Learning whole-body manipulation for quadrupedal robot," *arXiv preprint arXiv:2308.16820*, 2023.

[27] S. Calinon, F. Guenter, and A. Billard, "On learning, representing, and generalizing a task in a humanoid robot," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, no. 2, pp. 286–298, 2007.

[28] P. Kormushev, S. Calinon, and D. G. Caldwell, "Imitation learning of positional and force skills demonstrated via kinesthetic teaching and haptic input," *Advanced Robotics*, vol. 25, no. 5, pp. 581–603, 2011.

[29] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent advances in robot learning from demonstration," *Annual review of control, robotics, and autonomous systems*, vol. 3, pp. 297–330, 2020.

[30] H. Dai, A. Valenzuela, and R. Tedrake, "Whole-body motion planning with centroidal dynamics and full kinematics," in *2014 IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2014, pp. 295–302.

[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] M. Posa, C. Cantu, and R. Tedrake, "A direct method for trajectory optimization of rigid bodies through contact," *The International Journal of Robotics Research*, vol. 33, no. 1, pp. 69–81, 2014.

[33] J.-P. Sleiman, F. Farshidian, M. V. Minniti, and M. Hutter, "A unified mpc framework for whole-body dynamic locomotion and manipulation," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4688–4695, 2021.

[34] M. P. Polverini, A. Laurenzi, E. M. Hoffman, F. Ruscelli, and N. G. Tsagarakis, "Multi-contact heavy object pushing with a centaur-type humanoid robot: Planning and control for a real demonstrator," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 859–866, 2020.

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

[36] 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.

[37] 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.

[38] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl, "Learning by cheating," in *Conference on Robot Learning*. PMLR, 2020, pp. 66–75.

[39] G. Ji, J. Mun, H. Kim, and J. Hwangbo, "Concurrent training of a control policy and a state estimator for dynamic and robust legged locomotion," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4630–4637, 2022.

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