Preference-Conditioned Multi-Objective RL for Integrated Command Tracking and Force Compliance in Humanoid Locomotion

Tingxuan Leng[1], Yushi Wang[1], Tinglong Zheng[2], Changsheng Luo[1], and Mingguo Zhao[1]

Abstract— Humanoid locomotion requires not only accurate command tracking for navigation but also compliant responses to external forces during human interaction. Despite significant progress, existing RL approaches mainly emphasize robustness, yielding policies that resist external forces but lack compliance-particularly challenging for inherently unstable humanoids. In this work, we address this by formulating humanoid locomotion as a multi-objective optimization problem that balances *command tracking* and *external force compliance*. We introduce a preference-conditioned multi-objective RL (MORL) framework that integrates rigid command following and compliant behaviors within a single omnidirectional locomotion policy. External forces are modeled via velocity-resistance factor for consistent reward design, and training leverages an encoder-decoder structure that infers task-relevant privileged features from deployable observations. We validate our approach in both simulation and real-world experiments on a humanoid robot. Experimental results indicate that our framework not only improves adaptability and convergence over standard pipelines, but also realizes deployable preference-conditioned humanoid locomotion. Video can be found in *the link*.
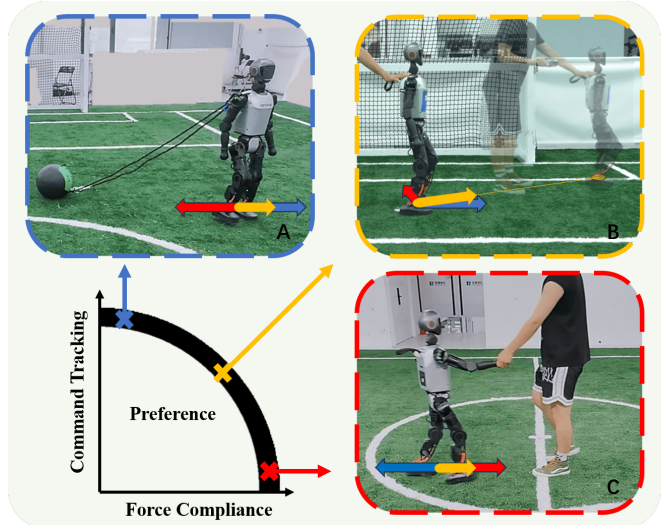
Fig. 1: **Preference-conditioned locomotion:** A single policy realizes behaviors from command tracking to human-guided compliance by adjusting the preference. Arrows indicate velocity command (blue), external force (red), and resulting humanoid velocity (yellow).

I. INTRODUCTION

Humanoid robots have become an emerging area of research in recent years, with locomotion being one of the most actively studied and developed areas. As humanoid robots are expected to operate in human-centered environments, locomotion must not only be robust but also naturally interactive: humanoids should reliably follow commands while also yielding to physical guidance from humans. This ability of *force compliance* is indispensable not only for safety and effective deployment, but also for natural and intuitive collaboration. Without it, humanoids risk exhibiting rigid or unsafe behaviors when interacting with people, such as resisting human guidance in emergencies or responding with unnatural or excessive forces.

Reinforcement learning with massively parallel simulation has enabled robust locomotion policies [1]–[5]. Recent work demonstrated walking on challenging terrains [6]–[9], blind navigation over stairs, push recovery [10]–[12], and stylized walking from human references [13], [14], showcasing RL's versatility for humanoid locomotion.

A key element of these approaches is the introduction of random force perturbations during training, which encourages the learned policy to remain stable under all conditions [1], [4], [5], but inherently biases policies toward force resistance. This training setting is essential for sim-to-real

[1]Tsinghua University, Beijing 100084, China
[2]Beijing Jiaotong University, Beijing 100044, China

transfer, and enables reliable velocity-tracking with push recovery, while simultaneously hindering force-compliance behavior.

Additionally, the pursuit of command tracking and force compliance is conflicting in nature: strong command following reduces compliance, while high compliance compromises responsiveness to current commands. Existing learning-based methods do not explicitly address this trade-off, and often compromise task-specific rewards to improve general performance [3], [15], leaving a gap between robust command tracking and interactive, compliant walking.

To bridge this gap, we regard external force not simply as perturbation but as a balancing force in continuous compliant walking, and formulate humanoid locomotion with the trade-off as a multi-objective optimization problem. A preference-conditioned MORL framework is introduced with the conflicting objectives: (1) following velocity commands, and (2) complying with sustained external forces. By conditioning the locomotion policy on a preference vector, a single integrated policy can smoothly interpolate between rigid tracking and highly compliant walking. This design allows humanoids to achieve both capabilities without requiring additional training stages or complex hierarchical architecture.

We evaluate the proposed method in simulation and on the adult-size humanoid Booster T1. Results validate that our approach is able to produce deployable omnidirectional locomotion that adapts flexibly to different given preferences.
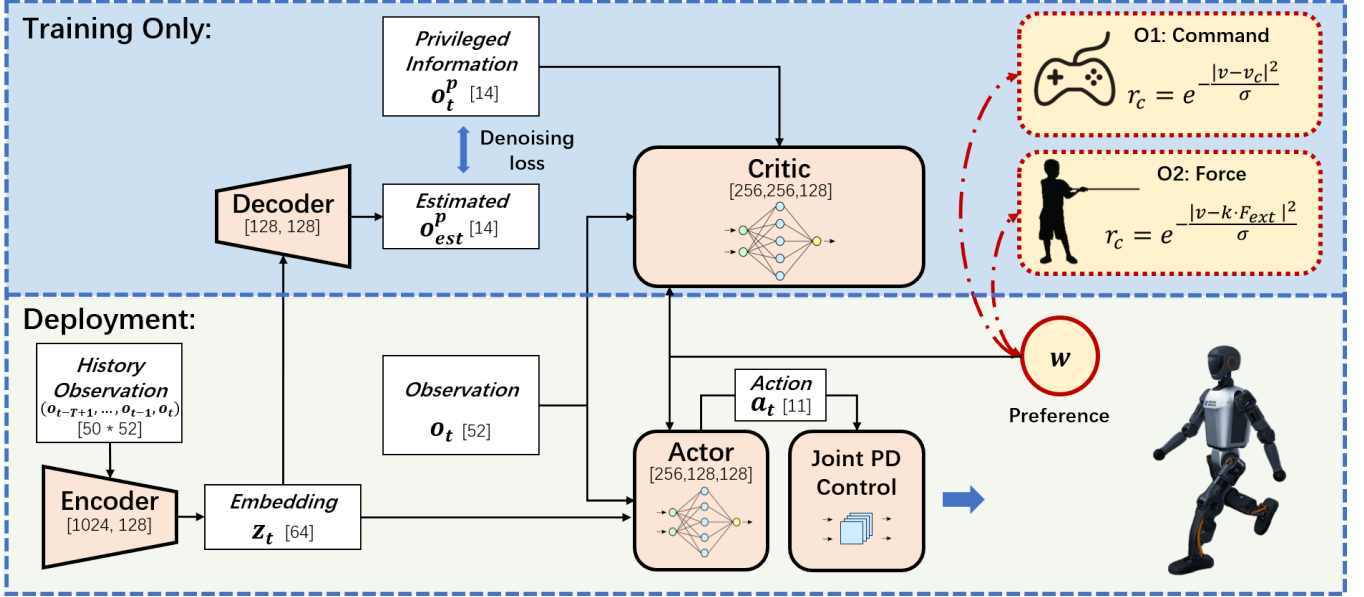
Fig. 2: **Policy training framework with privileged denoising:** An asymmetric actor–critic architecture is extended with an encoder–decoder that reconstructs privileged observations, guiding the encoder to extract force- and torque-aware latent features. At deployment, only encoder and actor remain, enabling preference-conditioned control with onboard observations, latent embedding and preference vector.

Unlike conventional policies that focus solely on disturbance resistance, our method enables humanoids to exhibit a new locomotion capability suitable for human-centered environments. The main contributions of this work are three-fold:

- We formulate humanoid locomotion as a **multi-objective problem**, where a **velocity–resistance model** provides a unified representation of commands and external forces, enabling consistent and comparable reward design.
- We propose a **preference-conditioned MORL framework** that learns a single unified policy capable of continuously balancing these objectives without hierarchical controllers or staged training.
- To our knowledge, this is among the first demonstrations of **humanoid locomotion**, validated in both **simulation and real-world deployment**, that achieves adaptive and deployable behaviors across different interaction preferences, especially demonstrating force-compliance.

II. RELATED WORK

A. RL-based Humanoid Locomotion

Reinforcement Learning has become a prominent approach for humanoid locomotion, with policies trained in large-scale parallel simulation and deployed zero-shot on hardware. Open-source frameworks [1], [4], [5] provide standardized environments and scalable pipelines, accelerating progress on robust and deployable locomotion policies for quadrupeds and humanoids.

A key paradigm in recent works is the *asymmetric actor–critic* architecture, where the critic has access to full states while the actor is constrained to partial observations. Teacher–student schemes further leverage historical observations to help the student policy infer task-relevant features. For example, RMA [7] adopts a two-stage pipeline: a teacher

policy is trained in the first stage with privileged information only available in simulation, and a student policy is distilled to rely only on deployable observations and latent embeddings. A-RMA [8] extends this to a three-stage procedure with an additional policy refinement process to adapt to a bipedal robot. More recently, [9] streamlined this paradigm into a single-stage framework, jointly training encoder and policy to enable zero-shot humanoid deployment. We build upon these advances and adopt Booster Gym [5] as the fundamental codebase and baseline environment.

B. Multi-Objective RL for Robotics

MORL has been widely applied in robotics, where objectives naturally conflict [16]. In animation and motion generation, it has been used to balance fidelity, style, and user control, as in AMOR [17] and [18], demonstrating strong performance across both robotics and animation domains. Furthermore, studies of biological and robotic gaits highlight inherent trade-offs among speed, energy, and stability [19], which further motivates the use of MORL in locomotion.

Building on this perspective, recent work has applied MORL to legged robots, coordinating objectives such as velocity tracking, torque cost, robustness, and whole-body control [20]–[22]. While these approaches show promising results, the objectives are typically *indirectly* conflicting—such as speed versus energy efficiency or fidelity versus robustness—where improvements often emerge from smooth trade-offs rather than explicit opposition.

Instead, our work focuses on a pair of *directly* conflicting objectives: velocity command tracking and force compliance. These two behaviors are inevitably opposed, as rigid command following tends to resist external forces, whereas compliant responses require following force at the expense of command accuracy. This remains an underexplored problem.

C. Locomotion with Force Adaptation

Learning-based approaches have begun to address locomotion under external forces. Advances in RL-based method has enabled legged robots to remain robust under perturbations and heavy loads [1], [7] and exert force needed for loco-manipulation tasks [23], [24]. However, force compliance is overlooked: while training policies to resist external forces, it makes compliant human guidance difficult and often produces unnatural reactions.

To encourage force-adaptive behaviors for legged robots, several works propose dedicated RL-based methods, while primarily focusing on quadruped locomotion or loco-manipulation tasks. Hartmann et al. [25] introduced compliant quadruped control through a recovery stage during training. HAC-LOCO [26] adopts a hierarchical design where a high-level residual module modulates a robust low-level policy. FACET [27] instead imitates a spring–impedance reference through a two-stage teacher–student pipeline with adaptive parameters. FALCON [28] addresses loco-manipulation with a dual-agent design and torque-limit-aware 3D end-effector force curricula, while the lower agent focusing on the supporting manipulation goals.

While these methods demonstrate effective adaptation in their respective domains, their assumptions and architectures do not directly transfer to humanoid locomotion, where balance is significantly more fragile than quadruped and omnidirectional walking itself is already a challenging task.

In contrast, our work addresses humanoid locomotion exclusively, aiming to unify velocity command tracking and external force compliance in a single policy. To our knowledge, our method is among the first to realize preference-conditioned force adaptability in humanoid omnidirectional locomotion, complementing prior quadruped and manipulation-focused compliance frameworks.

III. METHOD

In this section, we present our method for learning a humanoid locomotion policy that balances command tracking and force compliance. We first align the velocity and force from the perspective of a consistent velocity-resistance model (Section III-A). Then, we formulate the problem as a MORL task with preference conditioning (Section III-B). Building on these, we establish our training framework to obtain adaptive and deployable policy (Section III-C).

A. Velocity-Resistance Modeling of External Force

We formulate humanoid locomotion as balancing velocity tracking and force compliance in the horizontal plane, while vertical dynamics, which is commonly related with complex terrains, is beyond the scope of this work. Both linear and angular velocities are considered, corresponding respectively to external forces and torques. In practice, we envision the robot to respond in a steady and continuous manner when guided by external forces, e.g., maintaining gentle contact with a human hand during interaction.

To balance command tracking and force compliance within a single locomotion policy, velocity commands and external forces must be expressed in a comparable form. However, they inherently lie in different physical spaces. To unify them, we adopt a velocity–resistance perspective that maps sustained external forces into equivalent velocities, allowing both objectives to be represented consistently in the reward function.

In many physical systems, resistive forces scale linearly with velocity, as in viscous damping:

$$F_{\text{res}} = -B \cdot v, \tag{1}$$

where B is an effective damping coefficient. At steady state, external and resistive forces balance, yielding

$$F_{\text{ext}} + F_{\text{res}} \approx 0 \implies v \approx B^{-1} \cdot F_{\text{ext}}. \tag{2}$$

This formulation is physically suitable for humanoid locomotion: (i) human-applied forces are typically slow and low-frequency, making the steady-state approximation reasonable; (ii) it induces intuitive compliance—robots move only while being pulled, and stop naturally once the force is released; and (iii) it avoids oscillatory or unstable behaviors that can arise from high-order dynamics.

For training, we adopt a simplified mapping

$$v_{\text{ext}} = k \cdot F_{\text{ext}}, \tag{3}$$

where k is a scalar gain acting that normalizes the equivalent velocity to a feasible range. This approximation is not intended to replace full impedance control for high-bandwidth interaction, but rather provides a stable, interpretable, and physically grounded mechanism that makes external forces directly comparable to velocity commands during locomotion training with RL.

B. Multi-Objective Reinforcement Learning Formulation

We formulate humanoid locomotion as a multi-objective reinforcement learning problem.

The standard locomotion problem is modeled as a Partially Observable Markov Decision Process (POMDP) $(\mathcal{S}, \mathcal{O}, \mathcal{A}, p, r, \gamma)$, where $\mathcal{S}, \mathcal{O}, \mathcal{A}$ denote the state, observation, and action spaces, $p(s'|s, a)$ the dynamics, r the reward, and γ the discount. Policy is represented by $\pi(a|o)$. In MORL, the reward is extended to a vector

$$\mathbf{r}(s, a) = [r_1(s, a), \ldots, r_n(s, a)], \tag{4}$$

capturing multiple, possibly conflicting objectives. The goal is to approximate the Pareto front of optimal trade-offs. To this end, we adopt a preference-conditioned policy $\pi(a|o, \mathbf{w})$ with weight vector $\mathbf{w} = [w_1, \ldots, w_n]$, s.t. $\sum_i w_i = 1$. The policy maximizes the expected return:

$$J(\pi; \mathbf{w}) = \mathbb{E}_{\tau \sim p_\pi} \Big[\sum_{t=0}^{\infty} \gamma^t \, \mathbf{r}(s_t, a_t) \cdot \mathbf{w} \Big]. \tag{5}$$

Sampling diverse $\mathbf{w}$ during training enables the policy to cover a family of Pareto-optimal solutions.

For humanoid locomotion, we consider three objectives: **command tracking** (r_c), **force compliance** (r_f), and **regularization** (r_r):

$$r_c = \exp\Big(-\frac{\|v - v_c\|^2}{\sigma} \Big), \tag{6}$$

$$r_f = \exp\left(-\frac{\|v - k \cdot F_{\text{ext}}\|^2}{\sigma}\right), \qquad (7)$$

where σ is a hyperparameter. r_c measures the accuracy of command tracking, and r_f quantifies compliance to external forces in a comparable approach. The regularization term r_r is the sum of all additional rewards (e.g., base height, energy cost and stability penalties). For our task, w_c and w_f are constrained such that $w_c + w_f = 2$.

C. Policy Training with Privileged Denoising

Learning force compliance is challenging since external forces cannot be directly measured without tactile sensors. Inspired by [9], we design a one-stage privileged denoising framework, where the encoder infers force-related features from historical observations (see Figure 2)

1) Observation and Action Space: The actor receives deployable observations o (proprioception, velocity commands, gait features, etc.), while the critic additionally accesses privileged information o^p (e.g., external forces, linear velocities), available only in simulation. The action is the desired joint position $a = q_{des}$. A detailed specification of observations is shown in Table I.

TABLE I: Observation Space

Components	Dim	Critic	Actor
Preference Vector	3	✓	✓
Velocity Commands	3	✓	✓
Projected Gravity	3	✓	✓
Angular Velocity	3	✓	✓
Gait Signal	2	✓	✓
Gait Frequency	1	✓	✓
Joint Position	12	✓	✓
Joint Velocity	12	✓	✓
Last Action	12	✓	✓
Body Center of Mass	3	✓	
Body Mass	1	✓	
Linear Velocity	3	✓	
Body Height	1	✓	
Pushing Force	3	✓	
Pushing Torque	3	✓	

2) Network Architecture and Training Pipeline: We adopt an asymmetric actor–critic structure with an encoder–decoder module. The encoder maps the stacked historical observations $o_H = (o_{t-T+1}, ..., o_{t-1}, o_t)$ into a latent embedding z_t, and the decoder reconstructs privileged states o_t^p, encouraging z_t to encode task-relevant features. The actor is conditioned on (o_t, z_t, w_t) and outputs joint commands tracked by a low-level joint PD controller, while the critic has access to both observations and privileged observations (o_t, o_t^p) without requiring the embedding.

During training, the actor and critic are updated using PPO [29], with the MORL rewards defined in Section III-B. For additional implementation details of the rewards, please refer to [5]. To obtain a deployable policy, we employ a curriculum learning strategy together with domain randomization (DR). Training begins on flat terrain without velocity perturbations (sudden velocity impulses), which are gradually increased as training progresses. The environment is then switched to uneven terrain after half of the training. External forces are randomly applied during training, resampled every 30 s from $U(-50, 50)$ and later narrowed to $U(-20, 20)$ to increase sensitivity to subtle interactions. The preference weight w_c is resampled each episode from $U(0, 2)$ with $w_f = 2.0 - w_c$, while w_r is randomly sampled from $U(1, 2)$ to improve robustness.

The encoder–decoder is optimized jointly with reconstruction losses:

$$L_{denoising} = \|\hat{o}^p - o^p\|^2 \qquad (8)$$

$$L_{denoising,force} = \|\hat{F_{\text{ext}}} - F_{\text{ext}}\|^2 \qquad (9)$$

where $\hat{o}^p$ is the estimated privileged observation. The second term is introduced to emphasize force-related features.

At deployment, only the encoder and actor are used, producing a policy that adapts online according to the preference vector. For evaluation, the regularization weight is fixed to $w_r = 1$ for consistency and better robustness.

IV. SIMULATION EXPERIMENTS AND ANALYSIS

The policy is trained in Isaac Gym [30] with 4096 parallel environments. Each episode lasts 20 seconds with conditional early termination, and training is run for 20,000 epochs. All experiments are performed on the Booster T1 humanoid model at a 50 Hz control frequency, are tested after sim-to-sim to Mujoco [31] integrated in the codebase.

Direct comparison with quadruped- or manipulation-focused methods is not meaningful due to differences in morphology and task setup. Therefore, we evaluate our approach against a strong humanoid locomotion baseline built on Booster Gym [5]. The baseline policy adopts the same structure but is not specifically trained for force compliance (i.e. trained without force compliance objective). In the following sections, we first examine whether our preference-conditioned MORL policy (MORL policy in short) can learn a Pareto front that captures the trade-off between command tracking and force compliance. We then evaluate its adaptability when preference weights are switched online, compare its training convergency against a single-objective RL pipeline, and finally assess its robustness under unexpected perturbations.

A. Pareto Front across Different Preferences

To demonstrate its optimality and adaptability across multiple objectives, we evaluate the policy under command weights from 0.0 to 2.0 with interval 0.1 and corresponding $w_f = 2 - w_c$.

First, the robot operates in the opposite setting: it is commanded to walk forward with a constant velocity while subjected to three levels of backward force. The average forward velocity $\bar{v}_x$ is recorded for each trial. Similarly, angular velocity tracking is tested with constant yaw commands under three levels of external torque. For visualization, rewards for each objectives are expressed as mean squared errors (MSE):
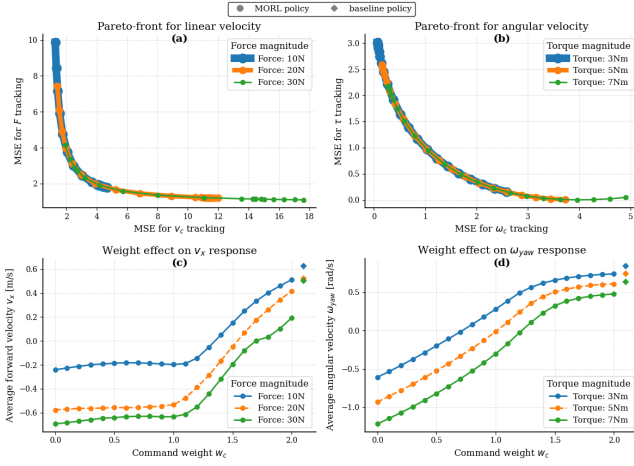
$$r_c = \mathbf{MSE}(v_x, v_{c,x}) \qquad (10)$$

Fig. 3: **Effect of reward weight w in the opposite setting:** Each trial lasts 5 s, and the baseline policy result is plotted on the rightmost side for comparison. (a)(b) Pareto-front visualization of linear and angular velocity. (c) average forward velocity $\bar{v}_x$ with $v_{c,x} = 1.0$ m/s under three levels of backward force (10 N, 20 N, 30 N). (d) average yaw angular velocity ω_{yaw} with $\omega_c = 1.0$ rad/s under three levels of applied torque (3 Nm, 5 Nm, 7 Nm).
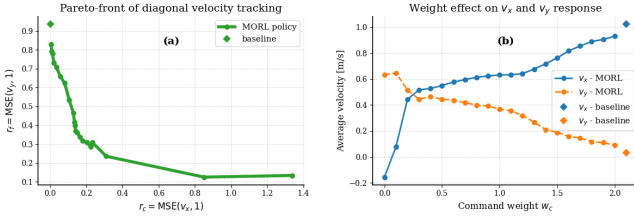

Fig. 4: **Effect of reward weight w in the orthogonal setting:** Each trial lasts 5 s, and the baseline policy result is plotted for comparison. (a) Pareto-front visualization of linear velocity. (b) velocity v_x and v_y with $v_{c,x} = 1.0$ m/s under a leftward force (30 N).

$$r_f = \mathbf{MSE}(v_x, k \cdot F_{\text{ext}}) \qquad (11)$$

The results in Figure 3 show Pareto fronts (upper panels) and velocity responses (lower panels). From the results, we observe that the majority of solutions are non-dominated, i.e., they cannot be improved in one objective without degrading another, and thus lie on the Pareto-front (PF). This indicates that our training strategy can effectively approximate the Pareto-optimal set. The PF exhibits an approximately monotonically decreasing trend, further confirming the strongly conflicting nature of this locomotion problem. Moreover, the smoothness of the curve suggests that our policy provides a continuous and feasible solution space, accommodating arbitrary user-specified preferences.

We further test an orthogonal setting where forward velocity is commanded while lateral forces are applied. As shown in Figure 4, the policy produces diagonal walking by combining both objectives, whereas the baseline fails to adapt and mainly tracking command.

During all experiments, the MORL policy succeeds without any unstable behaviors or falls, validating the reliability of our training setup. The policy demonstrates a smooth trade-off between tracking and compliance across preference weights. When $w_c = 2.0$ (and $w_f = 0.0$), the MORL policy achieves comparable tracking accuracy to the baseline. For smaller w_c, it exhibits significantly stronger compliance to
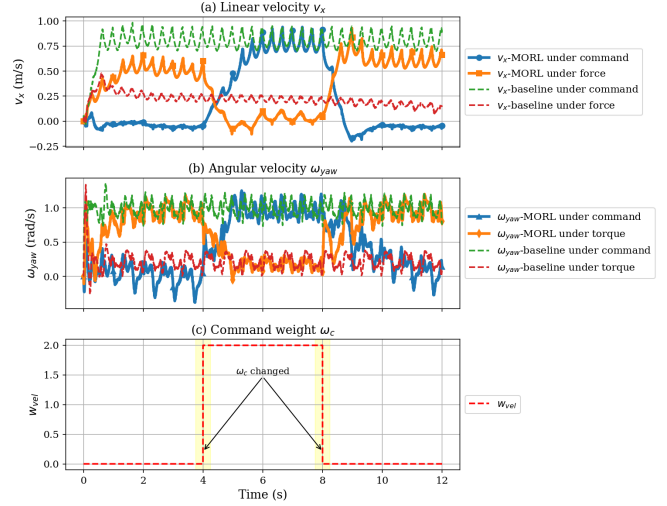

Fig. 5: **Online switching of command weight:** The trajectory lasts for 12 seconds with the command weight changed every 4 seconds. The robot is applied constant $v_{c,x} = 1.0$m/s, $F_{\text{ext}} = 20$N, $\omega_c = 1.0$rad/s, $\tau = 5$N·m separately. The policy's behavior changes corresponding to the command weight.

external forces, which the baseline lacks due to its force-unaware training. These results confirm the effectiveness of the proposed method in integrating competing objectives and highlight its versatility in adapting to diverse interaction demands.

B. Adaptation to Online Preference Switching

Furthermore, the MORL policy demonstrates the ability to adapt online when preference weights are switched during walking. The humanoid is tested under two separate conditions: either commanded with a constant forward velocity $v_{c,x}$ or subjected to a constant external force F_{ext}, with the forward velocity reward weight w_c switched twice during each trial. Similarly, angular velocity and torque tracking are evaluated under analogous separate conditions.

As shown in Figure 5, the policy transitions smoothly between behaviors without instability or sudden performance degradation, achieving performance comparable to the baseline on single-objective tracking. In contrast, the baseline policy supports only a single motion mode. These results highlight that the MORL framework not only provides a continuous spectrum of trade-offs but also enables seamless online preference adaptation, a critical capability for interactive humanoid deployment where task priorities may shift dynamically.

C. Ablation Study: MORL vs. Single-Objective RL

To further validate the effectiveness of our approach, we conduct an ablation study against a standard single-objective RL (SORL) pipeline trying to capture multiple objectives. In the single-objective setting, the trade-off is modeled as a scalar sum of rewards for conflicting objectives, with fixed weights $w_c = w_f = w_r = 1$. Both MORL and SORL use identical network architectures, optimizer settings, reward functions and amount of training steps (10,000 epochs); only the preference sampling differ. The comparison of the training curves is reported in Figure 6.
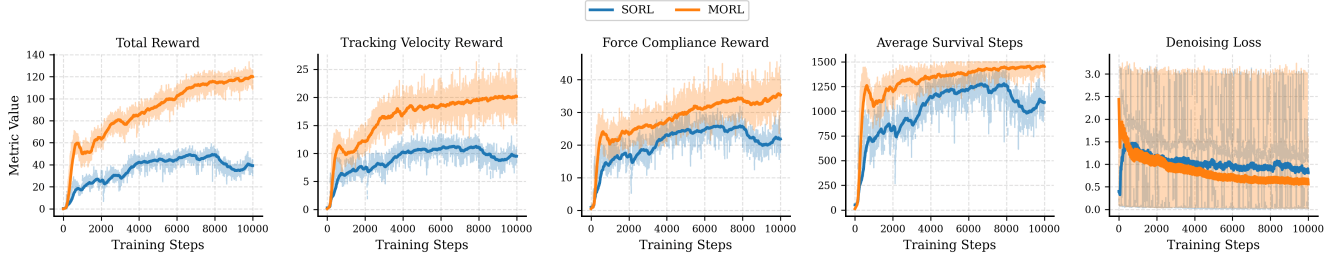
Fig. 6: Ablation study comparing MORL and SORL policies across five metrics: (1) total reward, (2) tracking velocity reward, (3) force compliance reward, (4) average survival steps, and (5) denoising loss. Curves are smoothed using Exponential Moving Average ($\alpha = 0.01$) while having the original values faded in the background.

The MORL policy achieves consistently higher rewards and improved training stability compared to SORL, with the privileged denoising process also converging faster. Counterintuitively, under extended conditions with varying preference weights and resampling process, the convergence speed and final performance of MORL are not weakened but instead enhanced, suggesting that a preference-conditioned scheme is crucial for fostering an integrated policy. This improvement likely stems from the inherent conflict between objectives. After training, the MORL agent develops adaptive locomotion skills, whereas the SORL policy fails to learn even basic walking skill and remains unstable. These results highlight the advantage of explicitly modeling conflicting objectives with MORL, leading to both more efficient training and superior emergent performance.

D. Robustness under Instant Perturbation

The test duration for this experiment is 20 seconds. The perturbation is modeled as an external force with fixed magnitude but random orientation. At every 5-second interval, a 1-second force impulse is applied to the robot's base link. We evaluate four different policy settings: (1) baseline policy, (2) MORL-c policy with $w_c, w_f = [2, 0]$, (3) MORL-m policy with $w_c, w_f = [1, 1]$, and (4) MORL-f policy with $w_c, w_f = [0, 2]$. Three levels of force magnitude are tested: $F_{ext} = 30N, 40N, 50N$. Each experiment is repeated 20 times for statistical reliability.

Two metrics are used to evaluate policy performance:

- **Success Rate:** The proportion of trials in which the humanoid doesn't fall during the testing time.
- **Peak Torque:** The average of the maximum root-mean-square (RMS) torque across all joints for each successful trial. For each trial, the RMS torque is computed over time for each joint, then the maximum RMS value among all joints is selected and are averaged over all successful trials. Only successful trials are considered to avoid spurious values from post-fall twitching.

The results in Table II indicate that while all policies perform equally well under mild perturbations (30N), MORL-based policies exhibit greater robustness at higher force magnitudes. In particular, MORL-f consistently achieves the highest success rates without increasing joint torque. Even under 50N perturbations, MORL-m and MORL-f maintain a 50% success rate, whereas the baseline fails completely.

TABLE II: Results of External Perturbation Experiments

Force	Policy	baseline	MORL-c	MORL-m	MORL-f
30N	Success Rate	100%	100%	100%	100%
	Peak Torque	15.542	14.677	15.249	14.457
	Std	±0.214	±0.146	±0.184	±0.174
40N	Success Rate	45%	90%	85%	100%
	Peak Torque	15.282	14.864	15.249	14.635
	Std	±0.165	±0.317	±0.237	±0.394
50N	Success Rate	0%	35%	50%	50%
	Peak Torque	–	14.758	15.229	14.589
	Std	–	±0.388	±0.201	±0.408

Moreover, in most cases, greater policy compliance corresponds to lower peak joint torque. These findings highlight the robustness advantage of MORL agents in terms of both stability and compliance.

V. REAL-WORLD EXPERIMENTS

We validate the proposed MORL-based locomotion policy on the Booster T1 humanoid robot through a series of hardware experiments. These experiments are designed to demonstrate two key aspects: (1) adaptability to different objective preferences, and (2) performance at specific preference settings. All experiments are documented in the supplementary video for clearer demonstration of the policies' performance.

A. Adaptability to Different Preferences

To evaluate adaptability, we conduct experiments under three preference settings identical to those in Section IV-D. The robot is guided either by joystick velocity commands or by consistent human-applied forces. The results are exclusively shown in the complement video.

As the force compliance weight w_f increases, the robot responds more readily to external forces, following human guidance with minimal resistance. Conversely, higher command-tracking weights w_c lead to faster execution of joystick inputs and stronger resistance against external forces. The baseline policy lacks such flexibility, focusing only on velocity tracking and becoming unstable under sustained human forces.

This experiment provides the most intuitive evidence that MORL enables real-time trade-offs between command tracking and force compliance.
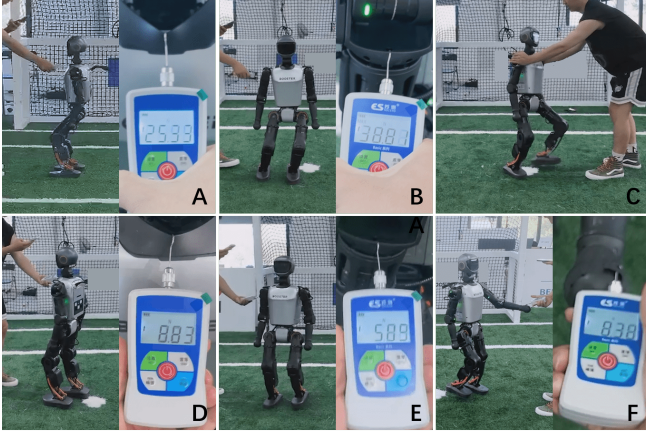
Fig. 7: **Omnidirectional force-compliance walking:** (a)(b)(c) shows the baseline policy with force in x,y and yaw direction. (d)(e)(f) shows our policy with force in x,y and yaw direction. The screenshots represent typical reading of the dynamometer in each case.

Fig. 8: **Force-compliance walking in outdoor scenarios:** (a) soccer field, (b) rough ground, (c) wooden ground, (d) smooth surface, (e) raised ground, (f) soft grass. The robot is guided by the human operator pulling its arms. The shoulder joint are kept flexible with a small pre-set $k_p = 1$. No supportive force is exerted to the robot.

B. Cross-Directional Walking

Beyond balancing between tracking and compliance, we further evaluate whether the policy can integrate multiple objectives simultaneously. Drawing from observations in earlier experiments, we choose preference weights $w_c, w_f, w_r = [1.5, 0.5, 2.0]$, which provide a practical balance between command tracking and force compliance. With this configuration, the robot is commanded to walk forward or laterally while being subjected to orthogonal external forces, as illustrated in Figure 1(b). Demonstrations of these experiments are provided in the supplementary video.

Under these settings, the robot consistently produces diagonal walking, effectively combining the commanded velocity and the compliant response to external forces. This emergent behavior highlights the MORL policy's ability not only to trade off between objectives, but also to synthesize them into coherent locomotion strategies in real time.

C. Force-Compliance Locomotion

We further assess force compliance, one of the distinguishing features of our approach. Using weights $w_c, w_f, w_r = [0.0, 2.0, 2.0]$, we pull the robot with minimal effort, measuring force with an external dynamometer Figure 7. The dynamometer is attached to the robot's neck, shoulder and hand with respect to forces in the x and y directions and torque around the z axis. Our MORL policy requires only about $10\,\mathrm{N}$ to move the robot smoothly and efficiently, while the baseline demands over $25\,\mathrm{N}$ and often exceeds the $30\,\mathrm{N}$ measurement limit. Additionally, the baseline policy generates singular actions to resist to the torque applied, making it impossible to measure with dynamometer, while our policy still remains gentle and stable. These results confirm the superior compliance of our method compared to existing locomotion frameworks.

The experiment in different outdoor environments is also conducted. In Figure 8, the robot is guided by hand-pulling and successfully traverses diverse terrains, including rough ground, a soccer field, raised surfaces, and others. The videos qualitatively showcase the practicality and robustness of our

Fig. 9: **Perturbation resistance under suspended-ball impacts**. The red arrow represents the direction of the strike. (a) Impact of a 2-kg ball, the robot maintains balance while maintaining position. (b) Impact of a 4-kg ball, the robot steps back but remains stable.

policy. Across all scenarios, the robot exhibits smooth, stable, and compliant walking behaviors, responding reliably to subtle guiding forces applied by the human operator.

D. Robustness to Instant Perturbation

To provide an additional indication of robustness on real hardware, we evaluate the system under sudden external impacts by swinging a suspended ball to strike the robot. As depicted in Figure 9, a ball is hung from the ceiling and released manually to hit the robot. The preference weight is set to be $w_c, w_f, w_r = [1.0, 1.0, 2.0]$

The experiments show that the MORL policy can withstand unexpected impacts from balls weighing up to $5\,\mathrm{kg}$ without falling (as in the video). The robot adapts by taking backward steps and absorbing the force through compliant motion. These observations align with the disturbance tolerance seen in simulation, providing additional empirical support, though this test is not the primary focus of our work.

VI. CONCLUSION

In this work, we introduced a preference-conditioned multi-objective RL framework for humanoid locomotion, which integrates command tracking and force compliance through a velocity–resistance model. The approach is simple

and effective, requiring no hierarchical control or multi-stage training, and enables a single policy to interpolate smoothly between rigid tracking and compliant behaviors. An encoder–decoder architecture inferring force-related latent features further allows the policy to be deployable on real humanoid hardware without direct force sensing.

Simulation and hardware experiments demonstrate that our method could generate adaptive and force-compliant omnidirectional walking. Beyond demonstrating feasibility, this study highlights preference-conditioned MORL as a general and versatile method for interactive humanoid locomotion, with opportunities for future extensions to richer tasks such as full loco-manipulation and higher-dimensional objectives.

ACKNOWLEDGMENT

REFERENCES

[1] 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.

[2] C. Zhang, J. Jin, J. Frey, N. Rudin, M. Mattamala, C. Cadena, and M. Hutter, "Resilient legged local navigation: Learning to traverse with compromised perception end-to-end," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 34–41.

[3] 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.

[4] X. Gu, Y.-J. Wang, and J. Chen, "Humanoid-gym: Reinforcement learning for humanoid robot with zero-shot sim2real transfer," *arXiv preprint arXiv:2404.05695*, 2024.

[5] Y. Wang, P. Chen, X. Han, F. Wu, and M. Zhao, "Booster gym: An end-to-end reinforcement learning framework for humanoid robot locomotion," *arXiv preprint arXiv:2506.15132*, 2025.

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

[7] A. Kumar, Z. Fu, D. Pathak, and J. Malik, "Rma: Rapid motor adaptation for legged robots," *arXiv preprint arXiv:2107.04034*, 2021.

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

[9] X. Gu, Y.-J. Wang, X. Zhu, C. Shi, Y. Guo, Y. Liu, and J. Chen, "Advancing humanoid locomotion: Mastering challenging terrains with denoising world model learning," *arXiv preprint arXiv:2408.14472*, 2024.

[10] D. C. Melo, M. R. Maximo, and A. M. da Cunha, "Learning push recovery behaviors for humanoid walking using deep reinforcement learning," *Journal of Intelligent & Robotic Systems*, vol. 106, no. 1, p. 8, 2022.

[11] Z. Zhuang, S. Yao, and H. Zhao, "Humanoid parkour learning," *arXiv preprint arXiv:2406.10759*, 2024.

[12] W. Xie, C. Bai, J. Shi, J. Yang, Y. Ge, W. Zhang, and X. Li, "Humanoid whole-body locomotion on narrow terrain via dynamic balance and reinforcement learning," *arXiv preprint arXiv:2502.17219*, 2025.

[13] Q. Zhang, P. Cui, D. Yan, J. Sun, Y. Duan, G. Han, W. Zhao, W. Zhang, Y. Guo, A. Zhang, *et al.*, "Whole-body humanoid robot locomotion with human reference," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 11225–11231.

[14] 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.

[15] 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.

[16] T. Shu, K. Shang, C. Gong, Y. Nan, and H. Ishibuchi, "Learning pareto set for multi-objective continuous robot control," *arXiv preprint arXiv:2406.18924*, 2024.

[17] L. N. Alegre, A. Serifi, R. Grandia, D. Müller, E. Knoop, and M. Bächer, "Amor: Adaptive character control through multi-objective reinforcement learning," in *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, 2025, pp. 1–11.

[18] A. Shrestha, P. Liu, G. Ros, K. Yuan, and A. Fern, "Generating physically realistic and directable human motions from multi-modal inputs," in *European Conference on Computer Vision*. Springer, 2024, pp. 1–17.

[19] R. Rockenfeller, R. L. Cieri, J. T. Schultz, R. Maag, and C. J. Clemente, "Coordinating limbs and spine:(pareto-) optimal locomotion in theory, in vivo, and in robots," *npj Robotics*, vol. 2, no. 1, p. 4, 2024.

[20] D. Kim, H. Kwon, J. Kim, G. Lee, and S. Oh, "Stage-wise reward shaping for acrobatic robots: A constrained multi-objective reinforcement learning approach," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 10268–10274.

[21] M. Robert, S. Brodeur, and F. Ferland, "Multi-objective algorithms for learning open-ended robotic problems," in *2024 IEEE 10th World Forum on Internet of Things (WF-IoT)*. IEEE, 2024, pp. 1–7.

[22] H.-L. Tran, L. Doan, N. H. Luong, and H. T. T. Binh, "A two-stage multi-objective evolutionary reinforcement learning framework for continuous robot control," in *Proceedings of the Genetic and Evolutionary Computation Conference*, 2023, pp. 577–585.

[23] 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. 15366–15372.

[24] 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.

[25] 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. 11421–11427.

[26] X. Zhou, X. Zhang, and Q. Zhang, "Hac-loco: Learning hierarchical active compliance control for quadruped locomotion under continuous external disturbances," *arXiv preprint arXiv:2507.02447*, 2025.

[27] 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.

[28] 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.

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

[30] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.

[31] 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.