

StyleLoco: Generative Adversarial Distillation for Natural Humanoid Robot Locomotion

Le Ma^{1*}, Ziyu Meng^{1,2*}, Tengyu Liu¹, Yuhan Li^{1,3}, Ran Song², Wei Zhang², Siyuan Huang¹, 

¹ National Key Laboratory of General Artificial Intelligence, BIGAI

² School of Control Science and Engineering, Shandong University

³ Huazhong University of Science and Technology

*Equal contributors  huangsiyuan@bigai.ai

Abstract—Humanoid robots are anticipated to acquire a wide range of locomotion capabilities while ensuring natural movement across varying speeds and terrains. Existing methods encounter a fundamental dilemma in learning humanoid locomotion: reinforcement learning with handcrafted rewards can achieve agile locomotion but produces unnatural gaits, while Generative Adversarial Imitation Learning (GAIL) with motion capture data yields natural movements but suffers from unstable training processes and restricted agility. Integrating these approaches proves challenging due to the inherent heterogeneity between expert policies and human motion datasets. To address this, we introduce StyleLoco, a novel two-stage framework that bridges this gap through a Generative Adversarial Distillation (GAD) process. Our framework begins by training a teacher policy using reinforcement learning to achieve agile and dynamic locomotion. It then employs a multi-discriminator architecture, where distinct discriminators concurrently extract skills from both the teacher policy and motion capture data. This approach effectively combines the agility of reinforcement learning with the natural fluidity of human-like movements while mitigating the instability issues commonly associated with adversarial training. Through extensive simulation and real-world experiments, we demonstrate that StyleLoco enables humanoid robots to perform diverse locomotion tasks with the precision of expertly trained policies and the natural aesthetics of human motion, successfully transferring styles across different movement types while maintaining stable locomotion across a broad spectrum of command inputs.

I. INTRODUCTION

Natural and agile locomotion in humanoid robots represents a fundamental challenge in robotics, with far-reaching implications for human-robot interaction, disaster response, and industrial applications. While humanoid robots offer unprecedented potential for operating in human-centric environments, achieving human-like movement patterns remains difficult due to their high degrees of freedom and inherently unstable dynamics[1]. This challenge is further complicated by the fundamental trade-off between achieving precise control and maintaining natural motion qualities.

Reinforcement learning (RL) has emerged as a powerful approach for developing locomotion controllers, enabling robots to master complex movements through carefully designed reward functions. These methods often employ a two-stage learning process: first training a teacher policy that relies on privileged information (such as global positions and ground truth environmental parameters) unavailable in real-world settings, then distilling this knowledge into a student policy that operates solely on realistic sensor observations. While this approach has demonstrated impressive results in



Fig. 1. Gait pattern transitions during forward velocity (v_x) acceleration from 0.7 m/s to 1.8 m/s

terms of agility and precision, it faces two key limitations. First, the reliance on handcrafted rewards requires extensive tuning to accommodate different gaits, stride lengths, and motion parameters across varying speeds. Second, these methods often result in rigid, mechanical movements that lack the fluidity and naturalness characteristic of human motion, limiting their effectiveness in human-centric environments.

Recent advances in generative adversarial imitation learning, particularly approaches like Adversarial Motion Prior (AMP) [2], have opened new possibilities for achieving more natural robot movements by leveraging large-scale motion capture datasets such as LaFAN1 [3] and AMASS [4]. These methods employ adversarial training to ensure that robot movements closely match the statistical patterns present in human demonstrations [5]. However, their performance is fundamentally limited by the content and quality of the reference motion data. For instance, learning running behaviors becomes impossible with a dataset containing only walking motions, and acquiring diverse specialized skills often requires expensive motion capture sessions. Furthermore, these methods struggle when motion datasets lack diversity or when retargeting processes introduce artifacts, resulting in brittle behaviors that fail to generalize beyond demonstrated movements.

The limitations of both approaches highlight a critical gap in humanoid locomotion: the need to combine the precision and adaptability of RL-based controllers with the natural movement qualities captured in human demonstrations. While RL methods can learn complex skills beyond available

motion capture data, they struggle with natural movement generation. Conversely, demonstration-based methods excel at producing natural movements but are constrained by the available motion capture data. This complementary nature suggests the potential for combining both approaches, yet traditional methods struggle to bridge this gap due to the fundamental heterogeneity between expert policies trained with handcrafted rewards and the statistical patterns present in human motion datasets.

We address these challenges with *StyleLoco*, introducing a novel Generative Adversarial Distillation (GAD) framework that effectively combines knowledge from heterogeneous sources. Our approach employs a multi-discriminator architecture where separate discriminators simultaneously distill skills from both an RL-trained expert policy and motion capture demonstrations. This design allows the model to preserve the agility and precision of RL while incorporating the natural style of human movements, enabling natural skill execution even for behaviors not present in the motion capture data. Through extensive evaluations in both simulated and real-world environments, we demonstrate that *StyleLoco* enables humanoid robots to achieve superior locomotion performance compared to traditional approaches while maintaining natural, human-like movement qualities.

The key contribution of our work is three-fold.

- A novel GAD framework that enables stable policy distillation from heterogeneous sources, effectively bridging the gap between RL and demonstration-based approaches.
- A multi-discriminator architecture that successfully combines task-oriented control objectives with natural motion patterns, achieving both high performance and human-like movement qualities.
- Comprehensive validation through real-world deployment on the Unitree H1 humanoid robot, demonstrating robust and natural motion across diverse locomotion tasks and speeds.

II. RELATED WORKS

A. Humanoid Robot Locomotion

Locomotion is a critical aspect in the motion control in humanoid robots. Traditional methods typically achieve stable movement by formulating the robot's dynamics model as constrained trajectory optimization problems [6]. Model Predictive Control (MPC) [7], [8], [9] is then employed in real-time to adjust and execute this trajectory, enabling adaption to dynamic environmental changes. However, these model-based methods usually rely heavily on precise modeling of robot dynamic properties[10], [11], [12], [13], [14] and environmental conditions[15], [16], [17], [18], [12], [19], [20], [21], [22], which leads to vulnerabilities in real-world performance , especially when there is a substantial discrepancy between the applied environments and the pre-defined conditions [23]. Thus, the optimization problem for humanoid robots is slow to resolve due to the complexity of high-dimensional state and action spaces, rendering it

challenging to satisfy the demands for real-time performance and stability.

Recently, reinforcement learning (RL) has emerged as a promising paradigm for humanoid locomotion tasks. These methods design tailored reward functions to guide “try and error” feedback-based learning process. For instance, reward functions are often crafted to encourage stable walking, minimize energy consumption, or optimize trajectory tracking [24]. However, designing effective reward functions is non-trivial and often requires extensive domain expertise especially for particular locomotion gaits. Natural locomotion motions require different gaits for varying movement speeds, making the design of the reward function even more challenging. Moreover, the numerous rewards terms must strike a delicate balance between competing objectives. To alleviate these drawbacks, we incorporate diverse reference locomotion motions as style guidance to simplify the reward components and encourage the policy learn versatile gaits.

B. Imitation Learning for Humanoid locomotion

The fundamental challenges in learning high-dimensional, underactuated robotic systems include precise task specification and effective exploration. Imitation learning (IL) is a method that learns from expert demonstrations, effectively addressing challenges related to quantifying rewards. Unlike pure reinforcement learning, IL can directly leverage offline expert data to guide policy learning, significantly reducing the exploration space and obtaining dense rewards. This approach is particularly effective in real-world robotics and complex task scenarios. Typically, it involves directly following reference trajectories through motion tracking. Generative Adversarial Imitation Learning (GAIL) [25] has been applied to locomotion tasks. The traditional imitation learning method, as mentioned above, is limited in flexibility—it can only replicate reference trajectories and cannot adapt to downstream tasks. To address this limitation, AMP [2] introduces the concept of learning the style from reference motion as a constraint, guiding the policy learning process.

However, this paradigm heavily relies on expert demonstrations, and its performance can significantly degrade when the quality of demonstrations is poor or when the task changes. Since IL strategies are directly derived from the demonstrations, they are prone to overfitting to the demonstration data. As a result, when faced with novel situations, IL may lack sufficient generalization ability. Furthermore, due to the morphological differences between humanoid robots and humans, obtaining high-quality reference data proves challenging, resulting in datasets that can only encompass a limited range of instructions. This scarcity of data can compromise the stability of Generative Adversarial Imitation Learning (GAIL), leading to mode collapse. To mitigate these challenges, we supplement the expert policy as a reference motion, providing additional motion references to achieve a stable omnidirectional movement strategy.

C. Deployable Policy Distillation

In robotic locomotion control, distillation is a method that transfers knowledge from teacher policies with privileged information (e.g., full-state dynamics, simulated ground-truth forces, or ideal state estimators) to student policies for real-world deployment. This knowledge transfer enables the student to leverage the teacher’s expertise while operating under real-world constraints, such as partial observation or limited sensory inputs. There are two main approaches to distillation:

BC methods[26], [27] learn by mimicking the teacher’s actions using supervised learning on state-action pairs. BC achieves effective performance when the student operates within the teacher’s training distribution, as it directly replicates the teacher’s behavior under familiar conditions. However, its performance degrades sharply with “compounding error” [28] in out-of-distribution (OOD) scenarios (e.g., environmental perturbations, actuator noise, or unseen terrains), as BC inherently lacks the capacity to self-correct deviations from the teacher’s demonstration space. This limitation arises because BC relies solely on static datasets of teacher demonstrations, without mechanisms to adapt to novel or unexpected situations.

Another popular approach is online distillation via Dataset Aggregation (DAgger) [29], which addresses BC’s limitations by iteratively aggregating student-generated trajectories with teacher-corrected actions. Recently, DAgger and its derivative strategies have stood out as a promising distillation approach for humanoid robot [30], [31], [32], [33] to acquire deployable policies. During training, the student policy interacts with the environment, while the teacher provides corrective feedback on the student’s actions, enabling the student to refine its policy over multiple iterations. This interactive process mitigates distributional shift and improves robustness to OOD scenarios. However, DAgger still faces a fundamental challenge: the student lacks access to the teacher’s privileged information (e.g., simulated contact forces, ideal state estimators, or full-state dynamics). As a result, under partial observation or incomplete environmental feedback, the student struggles to fully replicate the teacher’s actions. [24]

III. METHOD

StyleLoco is a novel approach for learning deployable natural locomotion skills that effectively combines the precision of RL-based controllers with the naturalness of human demonstrations. At its core, **StyleLoco** employs our proposed Generative Adversarial Distillation (GAD) framework, which uses a unique double-discriminator architecture to distill knowledge from both an RL-trained teacher policy and human motion demonstrations into a deployable student policy. Through adversarial learning, our approach generates naturalistic motions beyond the constraints of available motion capture data while avoiding the artificial behaviors typically resulting from hand-crafted rewards.

StyleLoco consists of three key components: (1) a teacher policy trained with privileged information to achieve

robust omnidirectional locomotion, (2) a motion dataset containing natural human movements, and (3) our novel GAD framework that combines these sources to train a deployable student policy. The framework’s innovation lies in its ability to generate natural behaviors beyond what either source can achieve alone - overcoming both the limited coverage of motion datasets and the unnatural movements that emerge from pure RL training.

To achieve this, **StyleLoco** employs two discriminators that work in concert to adversarially shape the student policy’s behavior. One discriminator ensures the policy can replicate the robust performance of the teacher, while the other maintains consistency with natural human motion patterns. This dual-discriminator approach simultaneously serves two purposes: expanding the range of natural behaviors beyond the demonstration data, and distilling the teacher’s capabilities into a deployable policy. The resulting system produces controllers that are both highly capable and naturally moving, without being constrained to demonstrated behaviors or exhibiting artifacts from hand-crafted rewards.

A. Preliminaries

1) *Reinforcement Learning*: We formulate humanoid locomotion control as a Partially Observable Markov Decision Process (POMDP) defined by tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{O}, R, \gamma \rangle$, where \mathcal{S} represents the full state space, \mathcal{O} denotes partial observations available to the robot, \mathcal{A} is the action space, $\mathcal{T}(s'|s, a)$ describes state transitions, $R(s, a)$ defines the reward function, and $\gamma \in (0, 1]$ is the discount factor. The goal is to learn a policy $\pi(a|o)$ that maximizes expected discounted returns while operating only on partial observations $o \in \mathcal{O}$.

The locomotion task requires tracking commanded velocities $v^* = (v_x^*, v_y^*, \omega_z^*)$, where (v_x^*, v_y^*) specify desired linear velocities in local coordinate frame and ω_z^* defines the desired yaw rate. Following [34], we use the reward function:

$$r_{\text{task}}(e, \lambda) := \exp(-\lambda \cdot \|e\|^2)$$

where e represents tracking errors and λ controls their relative importance.

2) *Generative Adversarial Imitation Learning*: Generative Adversarial Imitation Learning (GAIL) learns to mimic expert behavior through adversarial training. Given a dataset of expert demonstrations $\mathcal{M} = (s_i, a_i)$ consisting of state-action pairs, GAIL trains a policy $\pi(a|s)$ that generates actions a' for given states s' . A discriminator network \mathcal{D} is employed to distinguish between state-action pairs (s, a) from the expert demonstrations and those produced by the policy π . The reward function used to train the policy is then given by:

$$r_{\text{GAIL}}(s, a) = -\log(1 - \mathcal{D}(s, a))$$

Adversarial Motion Prior (AMP) [2] extends this framework to handle settings where only state information is available in the demonstrations. Instead of operating on state-action pairs, AMP’s discriminator evaluates state transitions (s, s') , enabling imitation learning from state-only demonstrations. Additionally, AMP employs a least-squares

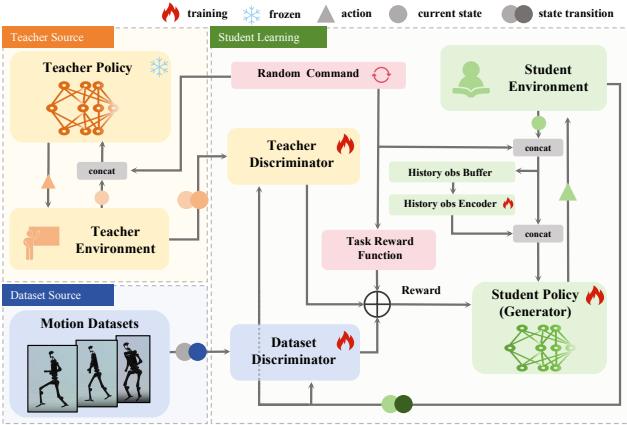


Fig. 2. Overview of the proposed Generative Adversarial Distillation (GAD) framework. Two discriminators separately evaluate the similarity of generated motions against a teacher policy and reference motion dataset, enabling the synthesis of natural and adaptive behaviors.

discriminator [35], replacing the traditional binary cross-entropy loss, which has been empirically shown to provide more stable adversarial training dynamics.

B. Generative Adversarial Distillation

The core innovation of StyleLoco is our GAD framework, which synthesizes natural and adaptive behaviors from two complementary sources: a well-trained teacher policy and a reference motion dataset. As illustrated in Fig. 2, GAD trains a student policy π_{student} alongside two AMP-style discriminators, $\mathcal{D}_{\text{teacher}}$ and $\mathcal{D}_{\text{dataset}}$. Each discriminator evaluates the student's generated state transitions against one source of reference motions: either the teacher policy or the motion dataset.

Training proceeds in an interleaving manner, alternating between updating the student policy and the discriminators. In each iteration, we first update the student policy using the combined feedback from both discriminators and then train both discriminators to better distinguish between the student's outputs and their respective reference motions.

The teacher discriminator $\mathcal{D}_{\text{teacher}}$ optimizes:

$$\begin{aligned} \arg \min_{\mathcal{D}_{\text{teacher}}} & \mathbb{E}_{(s, s') \sim \pi_{\text{teacher}}} \left[(\mathcal{D}_{\text{teacher}}(s, s') - 1)^2 \right] \\ & + \mathbb{E}_{(s, s') \sim \pi_{\text{student}}} \left[(\mathcal{D}_{\text{teacher}}(s, s') + 1)^2 \right] \\ & + \lambda \mathbb{E}_{(s, s') \sim \pi_{\text{teacher}}} \left[\|\nabla_{(s, s')} \mathcal{D}_{\text{teacher}}(s, s')\|^2 \right], \end{aligned}$$

while the reference discriminator $\mathcal{D}_{\text{dataset}}$ ensures natural motion qualities by optimizing:

$$\begin{aligned} \arg \min_{\mathcal{D}_{\text{dataset}}} & \mathbb{E}_{(s, s') \sim \mathcal{M}} \left[(\mathcal{D}_{\text{dataset}}(s, s') - 1)^2 \right] \\ & + \mathbb{E}_{(s, s') \sim \pi_{\text{student}}} \left[(\mathcal{D}_{\text{dataset}}(s, s') + 1)^2 \right] \\ & + \lambda \mathbb{E}_{(s, s') \sim \mathcal{M}} \left[\|\nabla_{(s, s')} \mathcal{D}_{\text{dataset}}(s, s')\|^2 \right], \end{aligned}$$

where λ controls the gradient penalty term that ensures stable training.

The student policy learns from a combined reward function:

$$r = r_{\text{task}} + w_{\text{teacher}} \cdot r_{\text{teacher}} + w_{\text{dataset}} \cdot r_{\text{dataset}},$$

where the discriminator rewards are computed as:

$$r_{\text{teacher}} = \max [0, 1 - 0.25(\mathcal{D}_{\text{teacher}}(s, s') - 1)^2]$$

$$r_{\text{dataset}} = \max [0, 1 - 0.25(\mathcal{D}_{\text{dataset}}(s, s') - 1)^2]$$

Both discriminators process state transitions using a consistent feature set comprising joint positions and velocities, root linear and angular velocities in the robot's local frame, base link orientation (roll and pitch), and root height. This common representation enables effective comparison across different motion sources while capturing the essential characteristics of locomotion behavior.

Deployable Policy Distillation A key aspect of our framework is enabling the student policy π_{student} to generate actions when privileged observations are unavailable in real-world deployment. While the teacher policy benefits from privileged information during training to better understand task objectives and achieve efficient convergence, the student policy must learn to generate appropriate actions using only deployable sensor observations. This asymmetric approach allows us to leverage rich state information during training while ensuring the final policy remains deployable. The specific observations available to the student policy are detailed in Table I.

C. Training Process

Curriculum Learning Teacher policy π_{teacher} training adopts a curriculum learning approach comprised of two distinct phases. The initial stability phase prioritizes maintaining balance and preventing falls, establishing fundamental stability behaviors. This is followed by the mobility phase, where the policy develops comprehensive omnidirectional locomotion capabilities. The specific reward components for each phase are detailed in Table II.

Demonstration Data Preparation The locomotion motion data in this work is sourced from the LaFAN1 dataset and meticulously retargeted to conform to the kinematic specifications of Unitree H1 robots. While this dataset offers diverse motion styles and velocity ranges, utilizing all demonstrations simultaneously introduces ambiguity in the learning process. To facilitate distinct gait style demonstrations across different velocity commands, we strategically selected motion clips with minimal or non-overlapping velocity ranges, ensuring a relatively clear behavioral boundaries between different locomotion patterns.

Asymmetric Actor-critic Architecture Student policy training utilizes an asymmetric actor-critic architecture to effectively handle partial observability in real-world conditions. The student's observation processing begins with temporal partial observations $o_t^N = [o_{t-n}, o_{t-n+1} \dots o_t]^T$. These observations are first processed through a partial states encoder \mathcal{E} to generate context latent representations, which are then combined with the current partial state observations

TABLE I
AVAILABLE OBSERVATIONS IN TRAINING

Sources	Phase	CmdVel	DoFPos	DoFVel	LastAction	Diff	BaseLinVel	BaseAngVel	RPY	Root Height	Push	Fraction	BodyMass	ContactStatus
Teacher	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Dataset				✓	✓		✓	✓	✓	✓				
Student	✓	✓	✓	✓	✓			✓	✓	✓				

Notes:

- Phase: Indicates the phase of motion, serving as a temporal marker.
- Diff: Difference between current joint angular position and reference joint angular position, calculated based on Phase.
- ContactStatus: Information regarding the stance mask and feet contact forces.

and the velocity command. The resulting combined representation passes through MLP layers to produce the final control actions.

TABLE II
REWARD DEFINITIONS USED IN TEACHER POLICY TRAINING.

Term	Definition	Weight
First Stage		
Termination	$\mathbb{I}_{\text{reset}} - \mathbb{I}_{\text{timeout}}$	-1000
Linear Velocity Tracking	$\exp\left(-\frac{\ v_{x,y}^{\text{target}} - v_{x,y}\ _2}{0.1}\right)$	10
Angular Velocity Tracking	$\exp\left(-\frac{\ \omega_z^{\text{target}} - \omega_z\ _2}{0.1}\right)$	10
Linear Velocity z	$\ v_z\ _2$	-1.0
R-P Angular Velocity	$\ \omega_{x,y}\ _2$	-0.5
Orientation	$\sum_{i \in \{x,y\}} (\text{projected gravity}_i)^2$	-1.0
Base Height	$\exp\left(-100 h_{\text{base}} - h_{\text{target}} \right)$ where $h_{\text{base}} = z_{\text{root}} - (h_{\text{feet}} - 0.08)$	0.5
Action Rate	$\ a_t - a_{t-1}\ _2$	-0.01
Energy Square	$\frac{\sum_{i=1}^{10} (\tau_i \dot{q}_i)^2}{1 + \ e_{x,y}\ _2}$	-5e-6
Stand Still	$(\sum_i q - q_{\text{default}}) \cdot \mathbb{I}_{\text{stand}}$	-1
Feet Clearance	$\sum_i \mathbb{I}\{ h_{\text{feet},i} - h_{\text{target}} < 0.01\} \cdot (1 - \text{gait phase}_i)$	2.5
Feet Contact Number	$\text{mean}(\mathbb{I}_{\{\text{contact}=\text{stance mask}\}} - \mathbb{I}_{\{\text{contact} \neq \text{stance mask}\}})$	1
Default Joint Position	$\ q_{[1:2]} - q_{[1:2]}^{\text{default}}\ _2 + \ q_{[6:7]} - q_{[6:7]}^{\text{default}}\ _2$	0.5
Action Smoothness	$\ a_{t-2} - 2a_{t-1} + a_t\ _2$	-0.001
Foot Slip	$1 - \sum_i \exp\left(-\ v_{x,y}^{\text{foot},i}\ _2\right)$	-0.05
Reference Joint Position	$\exp(-2\ q - q_{\text{ref}}\ _2) - 0.5 \min(\ q - q_{\text{ref}}\ _2, 0.5)$	10
Pelvis-Ankle y Distance	$(\ y_{\text{pelvis,pitch}} - y_{\text{ankle,L}}\ + \ y_{\text{pelvis,pitch}} - y_{\text{ankle,R}}\) \cdot \mathbb{I}\{ v_y < 0.1\}$	-5
Upper Joint Constraints	$\sum \ q_{[12:14]} - q_{[12:14]}^{\text{default}}\ + \sum \ q_{[16:18]} - q_{[16:18]}^{\text{default}}\ + \ q_{10} - q_{10}^{\text{default}}\ $	-5
Second Stage		
Joint Torque	$\ \tau\ _2$	-2e-5
Joint Acceleration	$\ q\ _2$	-1e-6
Feet Contact Forces	$\sum_i \max(\ \text{contact force}_i\ _2 - F_{\text{max}}, 0)$	-0.01
Torque When Stand-Still	$\sum_i [(\tau_t - \tau_{t-1})^2 + (\tau_t + \tau_{t-2} - 2\tau_{t-1})^2] \cdot \mathbb{I}_{\text{stand}}$	-1e-3
Body Pitch	$\ \text{pitch} - 0.01\ $	-5
Body Roll	$\ \text{roll}\ $	-10
Track Velocity Hard	$\frac{e^{-10\ v_{x,y}^{\text{target}} - v_{x,y}\ } + e^{-10\ \omega_z^{\text{target}} - \omega_z\ }}{2} - 0.2(\ v_{x,y}^{\text{error}}\ + \omega_z^{\text{error}})$	50
Ankle Air Time	$\sum_i (f_{\text{air},i} - 0.2) \cdot \mathbb{I}_{\text{first contact}, i} - \mathbb{I}_{\text{stand still}}$	100
Ankle Limits	$-\sum_{i \in \{4,9\}} \text{clip}(q_i - q_{\text{min},i}, 0) + \text{clip}(q_{\text{max},i} - q_i, 0)$	-200

Notes:

- $\mathbb{I}_A = 1$ if $A = \text{true}$ and $\mathbb{I}_A = 0$ otherwise.
- The maximum allowable feet contact force F_{max} is set to 550N

D. Implementation and Deployment Details

Both policies are implemented using the Proximal Policy Optimization (PPO) algorithm [36], with comprehensive domain randomization ensuring robust real-world transfer.

Domain Randomization Following existing researches on humanoid whole-body control, our domain randomization encompasses three aspects: physical parameter variations, systematic observation noise injection, and randomized external force perturbations. The physical parameters include variations in mass distribution, joint properties, and surface interactions. Observation noise is carefully calibrated to match real-world sensor characteristics, while external forces

simulate unexpected disturbances the robot might encounter during deployment.

Safe Deployment Safe deployment is achieved through torque limiting. This controller continuously monitors and adjusts torque outputs to remain within safe operational limits. The deployment architecture operates with the policy executing at 50Hz, while the low-level control loop maintains precise actuation at 1000Hz, ensuring responsive and stable behavior.

Real-world execution incorporates additional safety measures through continuous monitoring of joint positions, velocities, and torques. When approaching operational limits, the system smoothly modulates commands to maintain safe operation while preserving task performance. This approach enables robust deployment across varying conditions while protecting the hardware from potential damage.

IV. EXPERIMENTS

We conduct comprehensive experiments in both simulation and real-world environments to evaluate StyleLoco’s effectiveness in generating natural and adaptive locomotion behaviors. Our evaluation framework addresses four key aspects: (1) the effectiveness of GAD’s distillation capabilities, (2) the accuracy of velocity tracking during locomotion tasks, (3) the quality of motion style reproduction, and (4) real-world deployment performance.

All experiments are conducted using the Unitree H1 humanoid robot in both simulated and physical environments. For reference motions, we utilize the LaFAN1 dataset, carefully retargeted to match the H1’s kinematics. The motion data comprises global root position and orientation (quaternion), along with joint angular positions. Simulated experiments are performed in the NVIDIA Isaac Gym environment, which enables efficient parallel training and evaluation.

A. Distillation Performance

Our first set of experiments evaluates GAD’s ability to effectively distill privileged information from the teacher policy while maintaining task performance. We compare GAD against several baseline distillation approaches, measuring both task achievement and motion naturalness.

One of the main contributions of this work is the development of a Generative Adversarial Distillation method. In this context, we emphasize the ability of our single teacher discriminator (GAD-SD) to effectively distill knowledge from the teacher policy. To evaluate this capability, we compare our method against DAgger, one of the most widely used distillation methods in robot control.

First, we train an omnidirectional locomotion policy as the teacher. The command ranges used for both teacher training and the subsequent distillation experiment are listed in Table. III. We then leverage the well-trained teacher policy to guide the learning of the student policy.

TABLE III
RANGES OF LOCOMOTION TASK COMMAND

Parameter	Teacher (Unit)	Distillation student (Unit)	StyleLoco student (Unit)
Forward (v_x)	$[-1.0, 3.5] \text{ m/s}$	$[-1.0, 3.5] \text{ m/s}$	$[-1.0, 4.5] \text{ m/s}$
Lateral (v_y)	$[-0.8, 0.8] \text{ m/s}$	$[-0.8, 0.8] \text{ m/s}$	$[-1.0, 1.0] \text{ m/s}$
Angular (ω_z)	$[-1.0, 1.0] \text{ rad/s}$	$[-1.0, 1.0] \text{ rad/s}$	$[-1.5, 1.5] \text{ rad/s}$

The evaluation metrics include linear velocity tracking reward, angular velocity tracking reward, and average survival time. As shown in Table IV, while both methods successfully learn from the teacher policy, GAD-SD demonstrates superior performance, particularly in linear velocity tracking and survival time.

TABLE IV
QUANTITATIVE COMPARISON OF DISTILLATION METHODS

Method	Linear Velocity Tracking Reward(± 0.1) \uparrow	Angular Velocity Tracking Reward(± 0.1) \uparrow	Average Survival Time(± 15 steps) \uparrow
Teacher	7.403	2.824	925.9
DAgger	3.744	2.516	506.6
GAD-SD	5.679	2.653	860.3

Notes:

- Teacher: teacher policy trained with privileged information
- GAD-SD: GAD with only teacher distillation discriminator

B. Locomotion Capabilities

The second set of experiments assesses the student policy's locomotion capabilities, particularly its ability to track commanded velocities while maintaining natural motion patterns. We compare StyleLoco against state-of-the-art approaches in terms of tracking accuracy, stability, and style preservation. Table VI shows comparative results across various performance metrics.

The locomotion task evaluates the ability of student policy to track local velocity commands comprising three components: forward/backward velocity v_x , lateral velocity v_y , and rotational velocity w_z . Command values are uniformly sampled within pre-defined ranges specified in Table. III. For style imitation, we select four representative motion clips as reference targets for the style discriminator, with their corresponding velocity profiles detailed in Table. V.

TABLE V
VELOCITY PROFILES FOR MOTION CLIPS

Vel Profiles	Forward (m/s)	Lateral (m/s)	Angular (rad/s)
Slow Forward	[0.089, 1.205]	[-0.396, 0.188]	[-1.734, 0.906]
Medium Forward	[0.884, 2.067]	[-0.563, 0.306]	[-2.044, 1.963]
Fast Forward	[2.438, 4.378]	[-1.166, 0.943]	[-1.555, 3.476]
Move Backward	[-1.088, -0.350]	[-0.425, 0.365]	[-1.580, 1.981]

To comprehensively evaluate our double-discriminator framework, we compare our method against three baseline approaches:

- SD-Motion: Single-discriminator approach using only motion clips as reference.
- SD-Full: Single-discriminator approach using a combination of teacher policy online roll-out data and motion clips.
- DAgger+Style: DAgger-based teacher policy distillation combined with a separate discriminator for style learning.

The evaluation metrics are similar to those used in the distillation task experiment, with the addition of energy consumption.

As demonstrated in Table. VI, our proposed double-discriminator framework achieves superior performance in velocity tracking and survival time compared to all baseline methods. Notably, the SD-Motion approach exhibits the best energy consumption performance, suggesting that human motions are inherently energy efficient and properly incorporating motion demonstrations during training contributes to reduced energy consumption.

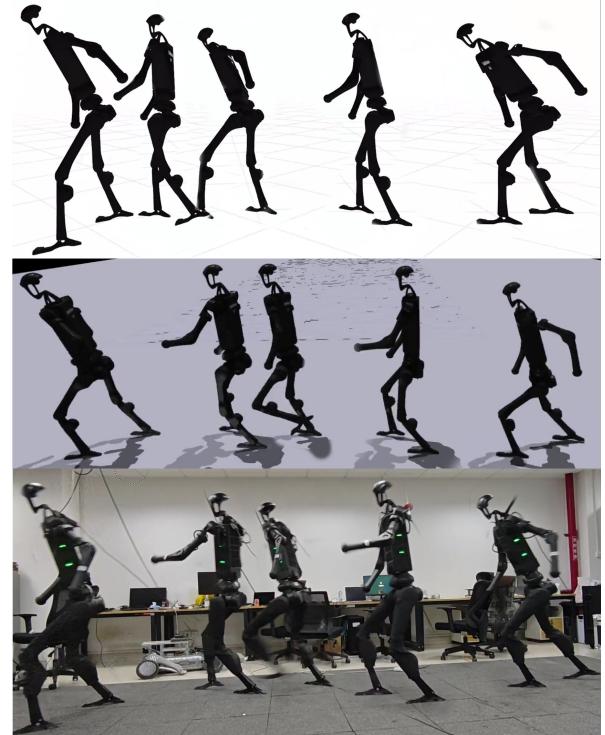


Fig. 3. From top to bottom, a stylized locomotion demonstration from LaFAN1 (Top), motions generated by student policy in simulation (Middle), motions generated by student policy deployed on real H1 robot(Bottom).

C. Evaluations on Style Imitation

To demonstrate our method's ability to combine robust locomotion skills with distinct motion styles, we evaluate a particularly challenging case: synthesizing a limping gait by combining a regular walking teacher policy with reference motions exhibiting a distinct limping pattern. Fig. 3 shows

TABLE VI
QUANTITATIVE COMPARISON OF DIFFERENT METHODS ACROSS VARIOUS METRICS

Method	Linear Velocity Tracking Reward(± 0.1) \uparrow	Angular Velocity Tracking Reward(± 0.1) \uparrow	Average Survival Time(± 15 steps) \uparrow	Energy Consumption(± 0.001) \downarrow
SD-Motion	4.229	2.249	813.2	0.065
SD-Full	4.665	2.413	824.1	0.093
DAgger+Style	5.059	2.384	826.9	0.079
GAD (Ours)	5.485	2.644	846.5	0.081

Notes:

- SD-Motion: Single discriminator with only motion demonstrations
- SD-Full: Single discriminator with both teacher roll-outs and motion demonstrations
- DAgger+Style: DAgger distillation with additional style discriminator

the comparison between the original limping motion from LaFAN1 (visualized in Rerun [37]), the synthesized motion in Isaac Gym [38], and the deployed behavior on the physical Unitree H1 robot. The results demonstrate that our method successfully maintains the characteristic limping style while preserving the fundamental locomotion capabilities of the teacher policy.

This fusion of different motion sources creates an inherent trade-off between style fidelity and command tracking accuracy, as the stylized motions often deviate significantly from the teacher’s optimal movement patterns. Our framework addresses this challenge through adjustable discriminator weights, allowing fine-tuned balance between style preservation and task performance.

D. Real Robot Deployment

The real-world deployment of our student policy on the Unitree H1 robot validates the practical effectiveness of our approach across various scenarios. As shown in Fig. 1, the robot demonstrates smooth transitions in both gait patterns and arm postures when responding to velocity command changes from low to medium speeds. The policy’s robustness is further evidenced in Fig. 4, where the robot maintains stable locomotion at high speeds up to 3 m/s. Most notably, Fig. 3 showcases our method’s unique capability to synthesize stylized gaits that combine the stability of the teacher policy with distinctive motion patterns from the reference datasets, resulting in natural and controllable locomotion behaviors.

V. CONCLUSION AND LIMITATIONS

This paper presents StyleLoco, a novel framework for humanoid locomotion that bridges the gap between robust task execution and natural motion synthesis. Through our proposed Generative Adversarial Distillation approach, we demonstrate the effective combination of privileged information from expert policies with stylistic elements from human demonstrations. Our extensive experimental results, including successful deployment on the Unitree H1 robot, validate the framework’s capability to generate stable and natural locomotion behaviors across diverse scenarios, from high-speed running at 3 m/s to stylized gaits such as limping.

The key innovation of our double-discriminator architecture enables simultaneous learning from heterogeneous

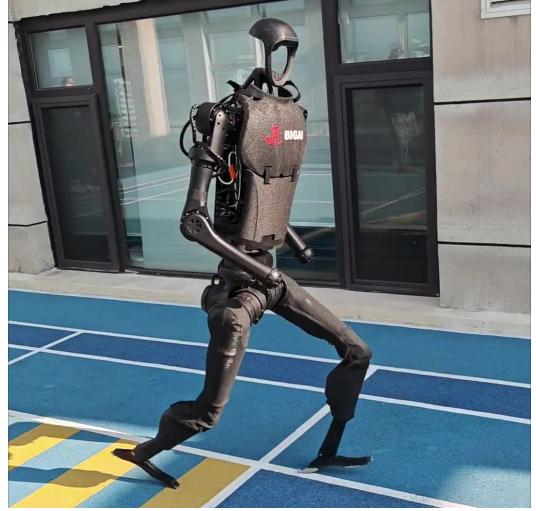


Fig. 4. H1 operating outdoors at forward velocity (v_x) of 3 m/s

sources while maintaining deployability through careful handling of privileged information. Quantitative evaluations show that StyleLoco outperforms existing approaches in both task performance and style preservation, demonstrating superior velocity tracking rewards and survival times while maintaining natural motion patterns.

Despite these achievements, several important limitations warrant future investigation. A primary challenge lies in style disambiguation when motion demonstrations share overlapping velocity ranges, potentially creating ambiguity in style selection and degrading imitation fidelity. Future research could explore automatic style clustering or context-aware selection mechanisms to address this limitation. Additionally, the current implementation relies on manual tuning of discriminator weights to balance task completion and style imitation objectives. Developing adaptive weighting schemes or automated tuning methods could enhance the framework’s practical applicability. While our method shows impressive results in locomotion tasks, its generalization to broader manipulation tasks or more complex behaviors remains to be explored, opening avenues for future research.

Despite these limitations, StyleLoco represents a step toward natural and capable humanoid robotics, offering a promising foundation for future research in combining task-oriented control with human-like motion generation.

REFERENCES

- [1] K. Darvish, L. Penco, J. Ramos, R. Cisneros, J. Pratt, E. Yoshida, S. Ivaldi, and D. Pucci, "Teleoperation of humanoid robots: A survey," *IEEE Transactions on Robotics*, vol. 39, no. 3, pp. 1706–1727, 2023.
- [2] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa, "Amp: Adversarial motion priors for stylized physics-based character control," *ACM Transactions on Graphics (ToG)*, vol. 40, no. 4, pp. 1–20, 2021.
- [3] F. G. Harvey, M. Yurick, D. Nowrouzezahrai, and C. Pal, "Robust motion in-betweening," vol. 39, no. 4, 2020.
- [4] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "AMASS: Archive of motion capture as surface shapes," in *International Conference on Computer Vision*, Oct. 2019, pp. 5442–5451.
- [5] 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.
- [6] T. Marcucci, M. Gabiccini, and A. Artomi, "A two-stage trajectory optimization strategy for articulated bodies with unscheduled contact sequences," *IEEE Robotics and Automation Letters*, vol. 2, no. 1, pp. 104–111, 2017.
- [7] G. Romualdi, S. Dafarra, G. L'Erario, I. Sorrentino, S. Traversaro, and D. Pucci, "Online non-linear centroidal mpc for humanoid robot locomotion with step adjustment," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 10412–10419.
- [8] J. Englsberger, A. Dietrich, G.-A. Mesesan, G. Garofalo, C. Ott, and A. O. Albu-Schäffer, "Mptc-modular passive tracking controller for stack of tasks based control frameworks," *16th Robotics: Science and Systems, RSS 2020*, 2020.
- [9] M. Elbaidi, G. Romualdi, G. Nava, L. Rapetti, H. A. O. Mohamed, and D. Pucci, "Online non-linear centroidal mpc for humanoid robots payload carrying with contact-stable force parametrization," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 12233–12239.
- [10] Y. Ishiguro, K. Kojima, F. Sugai, S. Nozawa, Y. Kakiuchi, K. Okada, and M. Inaba, "High speed whole body dynamic motion experiment with real time master-slave humanoid robot system," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 5835–5841.
- [11] Y. Ishiguro, T. Makabe, Y. Nagamatsu, Y. Kojo, K. Kojima, F. Sugai, Y. Kakiuchi, K. Okada, and M. Inaba, "Bilateral humanoid teleoperation system using whole-body exoskeleton cockpit tablis," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6419–6426, 2020.
- [12] J. Ramos and S. Kim, "Dynamic locomotion synchronization of bipedal robot and human operator via bilateral feedback teleoperation," *Science Robotics*, vol. 4, no. 35, p. eaav4282, 2019.
- [13] K. Ayusawa and E. Yoshida, "Motion retargeting for humanoid robots based on simultaneous morphing parameter identification and motion optimization," *IEEE Transactions on Robotics*, vol. 33, no. 6, pp. 1343–1357, 2017.
- [14] K. Hu, C. Ott, and D. Lee, "Online human walking imitation in task and joint space based on quadratic programming," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 3458–3464.
- [15] F.-J. Montecillo-Puente, M. N. Sreenivasa, and J.-P. Laumond, "On real-time whole-body human to humanoid motion transfer," in *International Conference on Informatics in Control, Automation and Robotics*, 2010. [Online]. Available: <https://api.semanticscholar.org/CorpusID:20676844>
- [16] K. Yamane, S. O. Anderson, and J. K. Hodgins, "Controlling humanoid robots with human motion data: Experimental validation," in *2010 10th IEEE-RAS International Conference on Humanoid Robots*, 2010, pp. 504–510.
- [17] A. Di Fava, K. Bouyarmane, K. Chappellet, E. Ruffaldi, and A. Kheddar, "Multi-contact motion retargeting from human to humanoid robot," in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, 2016, pp. 1081–1086.
- [18] K. Otani and K. Bouyarmane, "Adaptive whole-body manipulation in human-to-humanoid multi-contact motion retargeting," in *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, 2017, pp. 446–453.
- [19] L. Penco, B. Clement, V. Modugno, E. Mingo Hoffman, G. Nava, D. Pucci, N. G. Tsagarakis, J. B. Mouret, and S. Ivaldi, "Robust real-time whole-body motion retargeting from human to humanoid," in *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, 2018, pp. 425–432.
- [20] J. Koenemann, F. Burget, and M. Bennewitz, "Real-time imitation of human whole-body motions by humanoids," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 2806–2812.
- [21] O. E. Ramos, N. Mansard, O. Stasse, C. Benazeth, S. Hak, and L. Saab, "Dancing humanoid robots: Systematic use of osid to compute dynamically consistent movements following a motion capture pattern," *IEEE Robotics Automation Magazine*, vol. 22, no. 4, pp. 16–26, 2015.
- [22] L. Penco, K. Momose, S. McCrary, D. Anderson, N. Kitchel, D. Calvert, and R. J. Griffin, "Mixed reality teleoperation assistance for direct control of humanoids," *IEEE Robotics and Automation Letters*, vol. 9, no. 2, pp. 1937–1944, 2024.
- [23] Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Reinforcement learning for versatile, dynamic, and robust bipedal locomotion control," *The International Journal of Robotics Research*, p. 02783649241285161, 2024.
- [24] Z. Fu, A. Kumar, J. Malik, and D. Pathak, "Minimizing energy consumption leads to the emergence of gaits in legged robots," in *5th Annual Conference on Robot Learning*.
- [25] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in neural information processing systems*, vol. 29, 2016.
- [26] X. Huang, Y. Chi, R. Wang, Z. Li, X. B. Peng, S. Shao, B. Nikolic, and K. Sreenath, "Diffuseloco: Real-time legged locomotion control with diffusion from offline datasets," 2024. [Online]. Available: <https://arxiv.org/abs/2404.19264>
- [27] B. Jia and D. Manocha, "Sim-to-real robotic sketching using behavior cloning and reinforcement learning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 18272–18278.
- [28] S. Ross and D. Bagnell, "Efficient reductions for imitation learning," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, Y. W. Teh and M. Titterington, Eds., vol. 9. Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 661–668. [Online]. Available: <https://proceedings.mlr.press/v9/ross10a.html>
- [29] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 627–635.
- [30] M. Ji, X. Peng, F. Liu, J. Li, G. Yang, X. Cheng, and X. Wang, "Exbody2: Advanced expressive humanoid whole-body control," *arXiv preprint arXiv:2412.13196*, 2024.
- [31] T. He, Z. Luo, W. Xiao, C. Zhang, K. Kitani, C. Liu, and G. Shi, "Learning human-to-humanoid real-time whole-body teleoperation," *arXiv preprint arXiv:2403.04436*, 2024.
- [32] T. He, W. Xiao, T. Lin, Z. Luo, Z. Xu, Z. Jiang, C. Liu, G. Shi, X. Wang, L. Fan, and Y. Zhu, "Hover: Versatile neural whole-body controller for humanoid robots," *arXiv preprint arXiv:2410.21229*, 2024.
- [33] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. Kitani, C. Liu, and G. Shi, "Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," *arXiv preprint arXiv:2406.08858*, 2024.
- [34] 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.
- [35] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.
- [36] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
- [37] Rerun Development Team, "Rerun: A visualization sdk for multimodal data," Online, 2024, available from <https://www.rerun.io/> and <https://github.com/rerun-io/rerun>. [Online]. Available: <https://www.rerun.io>
- [38] 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. [Online]. Available: <https://arxiv.org/abs/2108.10470>