Think on your feet: Seamless and Command-adaptive Transition between Human-like Locomotions

Huaxing Huang * [1], Wenhao Cui * [1] Tonghe Zhang * [2],
Shengtao Li [1], Jinchao Han [1], Bangyu Qin [1], Tianchu Zhang [1],
Liang Zheng [1], Ziyang Tang [1], Chenxu Hu [1]
Shipu Zhang [1], Zheyuan Jiang [†,1]

Abstract— **While training humanoid robots to mimic specific locomotion skills is achievable, enabling effective *meta-learning* in response to continuously changing commands presents a greater challenge. These robots must accurately track motion instructions, transition seamlessly between diverse actions, and master intermediate movements not present in their training data. In this work, we propose a novel approach that integrates human-like motion transfer with precise velocity tracking by a series of improvements to classical imitation learning. To enhance generalization, we employ the Wasserstein divergence criterion (WGAN-div). Furthermore, a Hybrid Internal Model provides structured estimates of hidden states and velocity to enhance mobile stability and environment adaptability, while a curiosity bonus fosters exploration. Our comprehensive method promises highly human-like locomotion that adapts to varying velocity requirements, direct generalization to unseen motions and multitasking, as well as zero-shot transfer to the simulator and the real world across different terrains. These advancements are validated through simulations across various robot models and extensive real-world experiments.**

I. INTRODUCTION

Humanoid robots possess great potential to mimic human behaviors, making them ideal for adapting to human-centric environments like factories and homes. However, unlike quadrupedal robots, humanoid robots struggle to learn locomotion skills effectively due to challenges such as a higher center of gravity, increased degrees of freedom, and larger body size.

Reinforcement learning (RL) has proven effective in teaching humanoids basic locomotion skills with minimal prior knowledge. However, exploring under weak reward signals can lead to unnatural gaits, resulting in high energy costs and mechanical wear that impede real-world deployment.

Imitation learning (IL) methods, such as Adversarial Motion Prior (AMP) [10], are promising for generating fluid, human-like motions by mimicking human demonstrations. Nonetheless, their application is limited by their supervised learning nature, leading to restricted generalization and a heavy reliance on high-quality expert data, which is expensive for robotics research [1].

These limitations hinder classical RL and IL algorithms from enabling humanoids to adopt fluid and versatile motion strategies like humans. This raises the question:

"Can we achieve seamless transitions between diverse human-like motions–such as walking, running, and spinning–that adapt to continuously changing velocity commands?"

In this work, we affirmatively address this question by proposing a novel humanoid locomotion learning algorithm. We integrate the human-mimicking capabilities of IL methods within an advanced RL framework while ensuring strong adherence to input velocity commands. Robots trained using our method adapt to diverse motion instructions while consistently exhibiting human-like movements, even in intermediate stages not covered in the expert data.

Our contributions include:

- **Novel architecture**: We developed a locomotion learning framework combining a Hybrid Internal Model with a WGAN-div module, enabling accurate human motion mimicry and command tracking. To our knowledge, this is the first work to achieve precise velocity tracking alongside authentic anthropomorphic movement in real-world scenarios.
- **Enhanced generalization**: Incorporating a curiosity HashNet and Wasserstein divergence loss functions allows humanoids to master unseen intermediate movements during transitions, significantly advancing imitation-learning techniques.
- **Extensive empirical evidence**: Our experiments in both simulation and real-world settings demonstrate strong adaptability to various movement commands while maintaining natural, human-like motions.

II. Related Work

A. Reinforcement Learning for Legged Locomotion

Reinforcement learning has been extensively studied in robotic locomotion [4], [8].For instance, [12] describes a setup that rapidly generates policies for robotic tasks using parallelism on a single GPU. Additionally, [9]

*Denotes equal contribution.
[1]Noetix Robotics. [2]Tsinghua University.
[†] Correspondence to: `merlin.jiang@noetixrobotics.com`

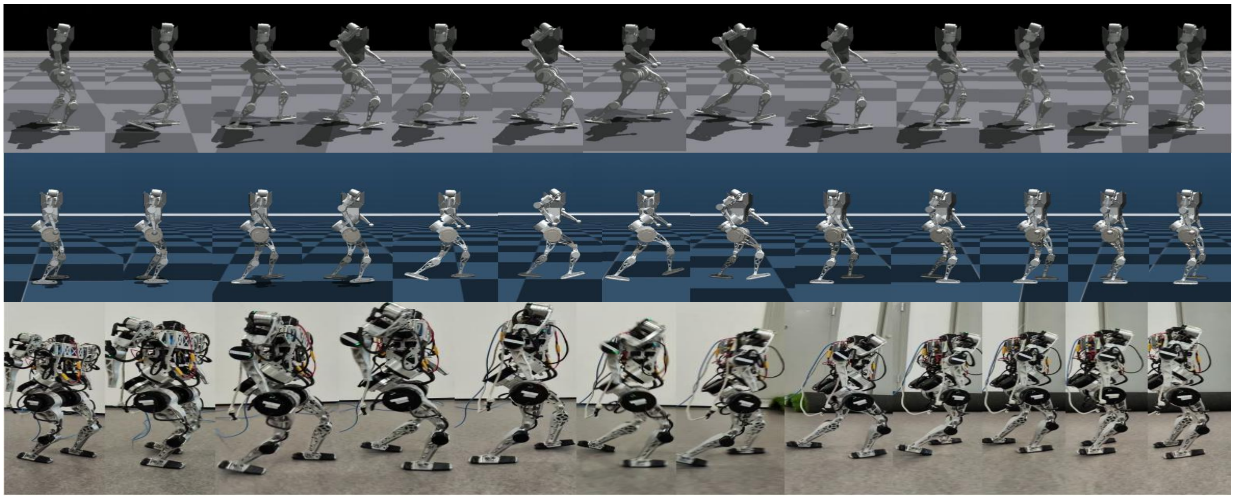
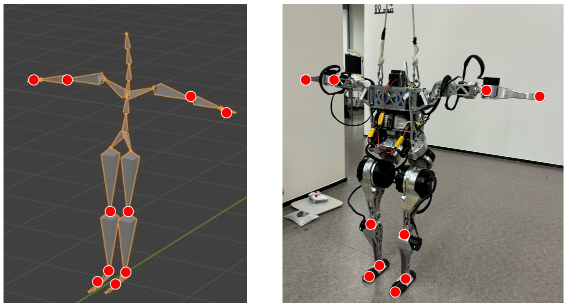


Fig. 1: Comprehensive demonstration of Noetix robot N1's locomotion skills learnt from the proposed method. The robot exhibits seamless and continuous transfer between highly human-like motion sets, accelerating from walking to running then coming to a full stop. Top to down: performance in simulator Isaac gym, Mujoco, and the real world. Parameters of the PD controller and the driving frequency are consistent in the three groups.



Real Human Motion Data

Our Robot Noetix N1

Fig. 2: Illustration of motion re-targeting from expert data (Left) to our humanoid robot "Noetix N1"(Right). N1 weights 23 kg and is of height ass 0.95 m, with 18 DoFs in total (four on each arm and five on each leg).

proposes an estimator that encodes environmental parameters through proprioceptive state histories.

While quadruped locomotion excels in navigating complex terrains, humanoid locomotion faces unique challenges due to higher degrees of freedom and the need for dynamic balance [11]. [2] developed an RL-based humanoid locomotion framework using the Legged Gym platform. [6] created a framework for training robust controllers for walking, running, and jumping skills.

B. Motion Imitation

In robotic locomotion, adapting complex motion patterns enhances robotic capabilities. Imitation Learning (IL) effectively tracks joint trajectories and extracts gait features but can struggle with discontinuities between locomotion patterns. Generative Adversarial Imitation Learning (GAIL) [3] addresses these continuity challenges. Innovations like Adversarial Motion Priors (AMP) enhance the generation of realistic motions from unstructured datasets, avoiding manual motion design constraints. These advancements support agile movements in quadrupedal and humanoid robots using refined IL techniques [13], [15], [16].

Building on these methodologies, our research proposes a framework that enhances robotic adaptability and performance in real-world environments.

III. Problem Setup

We model humanoid locomotion control as optimizing a partially observable Markov decision process $\mathcal{P} = (\mathcal{S}, \mathcal{O}, \mathcal{A}, p, r, \gamma)$, where state, observation, and action are denoted as $\mathbf{s} \in \mathcal{S}$, $\mathbf{o} \in \mathcal{O}$, and $\mathbf{a} \in \mathcal{A}$. The state transition probability is defined as $p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$. The policy π selects actions based on historical observations $\mathbf{a}_t \sim \pi(\cdot|\mathbf{o}_t^H)$. The reward function is $r_t = r(\mathbf{s}_t, \mathbf{a}_t)$ with a discount factor $\gamma \in [0, 1)$, while the objective is to maximize cumulative discounted rewards: $J(\pi) = \mathbb{E}_{\tau \sim p(\cdot|\pi)} \left[\sum_{t=0}^{+\infty} \gamma^t r(\mathbf{s}_t, \mathbf{a}_t) \right]$. Table I summarizes the construction and physical meaning of the spaces.

TABLE I: Components of Partial Observations o_t and Hidden States s_t

Entry	Dimension	Noise level	Category
Command	3	0	Observation
Base Angular Velocity	3	0.3	Observation
Base Rotation XY	2	0.09	Observation
DoF Position	18	0.075	Observation
DoF Velocity	18	2.25	Observation
DoF Action	18	0	Observation
DoF Position	18	-	Hidden State
DoF Velocity	18	-	Hidden State
Base Linear Velocity	3	-	Hidden State
Base Angular Velocity	3	-	Hidden State
Base Height	1	-	Hidden State

IV. Methodology

A. Motion Re-targeting and Correction

To achieve realistic robot movement, we utilize human motion data collected through Motion Capture (Mo-

Cap) to supervise the motion retargeting process for the Noetix N1 robot. We identify key points like toes, knees, and elbows, scale the source motion to match the robot's size (cf. Fig 2) , and apply Inverse Kinematics (IK) to compute the joint positions. Finally, we validate the motion files in simulators to ensure symmetrical movement and a straight trajectory.

B. Locomotion Learning with Hybrid Internal Model

We introduce an anthropomorphic locomotion learning framework combining velocity and implicit state estimation with human gesture supervision. Following the tradition of RL, we define a series of reward signals specified in table II) to measure how well the robot accomplishes basic locomotion tasks. To minimize simulation-to-real gap, we apply domain randominzation to the robot's hardware status and the environment's mechanical properties according to the configuration described in Table III. To take advantage of the anthropomorphism of Imitation Learning while enhancing the adaptability and stability of locomotions, we make significant improvements to classical RL framework, which is illustrated in Fig 3. In what follows, we highlight several key designs:

TABLE II: Task Reward

Reward	Equation r_i	Scale w_i
Feet slip	$\boldsymbol{\omega} \cdot I_{c(t)}$	-0.05
Feet contact forces	$\sum_i \left(\lVert \mathbf{F}_i^{\text{contact}} \rVert - F_{\max} \right)$	-0.01
Lin. velocity tracking	$\exp\left\{ -4(\mathbf{v}_{xy}^{\text{cmd}} - \mathbf{v}_{xy})^2 \right\}$	2.4
Ang. velocity tracking	$\exp\left\{ -4(\boldsymbol{\omega}_{\text{yaw}}^{\text{cmd}} - \boldsymbol{\omega}_{\text{yaw}})^2 \right\}$	1.1
Orientation	$\lvert \mathbf{g} \rvert^2$	1.0
Root accelerations	$\exp\left\{ -\left(\ddot{\mathbf{v}}_{\text{root}} \right)^3 \right\}$	0.2
Energy Cost	$\lvert \tau \rvert \lvert \dot{\mathbf{q}} \rvert$	-1e-3
Smoothness	$(\mathbf{a}_t - 2\mathbf{a}_{t-1} + \mathbf{a}_{t-2})^2$	-0.01

TABLE III: Domain Randomization

Parameter	Range	Unit
Base mass	[-5, 5]	kg
Center of mass shift	[-0.02, 0.02]	m
Friction coefficient	[0.1, 2]	-
K_p factor	[0.8, 1.2]	N·m/rad
K_d factor	[0.8, 1.2]	N·m·s/rad
Push Force	[-0.6, 0.6]	m/s
Push Torque	[-0.6, 0.6]	rad/s
Motor strength	[0.8, 1.2]	N·m
System delay	[0, 60]	ms

- A GAN-based discriminator. It provides additional supervision on the style of the robot's movements.
- A Hybrid Internal Model [7] with velocity estimate. It constructs velocity estimate with latent representation of the hidden states from historical observations.

The discriminator serves as an Imitation Learning module that provides a style reward based on the similarity between generated and expert motions, correcting unnatural joint positions. The Hybrid Internal Model (HIM) enhances locomotion imitation flexibility by generating velocity estimates and hidden states, offering rich insights into the relationship between input commands and the robot's status. This method enables humanoids trained with imitation learning and HIM to adaptively learn human-like movement in response to varying commands and diverse environments, thus improving motion stability in real-world scenarios.

The velocity estimator undergoes optimization via direct supervision, while latent state estimates are learned through contrastive learning, encouraging the agent to differentiate command status before making decisions. The policy and value networks are trained using Proximal Policy Optimization (PPO) with an added style reward. Since reference motion files include a wide variety of actions such as walking and running, the agent must learn to infer unseen intermediate states for seamless motion transfer, which contrasts with classical Imitation Learning that focuses on mimicking a limited action set. To enhance similarity with expert demonstrations while maintaining generalization, we introduce two modifications to our training pipeline, detailed in Sections IV-D and IV-C.

C. Prevent Mode Collapse by Wasserstein divergence

Our Imitation Learning (IL) module is based on Generative Adversarial Networks (GAN), which facilitates the generation of action sequences resembling motion capture data. This is accomplished by training a discriminator network $D(\cdot)$ against a generator $G(\cdot)$ to minimize the difference between real and synthetic data according to a specific criterion. To enable diverse motion learning, it is crucial that the IL module produces varied robot motions, which is essential for learning versatile movements through reinforcement learning (RL). However, traditional GANs and many of their variants face "mode collapse" where generators only capture singular peaks in the expert distribution, leading to monotonous outputs (illustrated in Fig 4). This limitation may restrict our humanoids to limited joint positions, resulting in stiff motions and reduced tracking ability. To enhance output diversity in the IL module, we employ a Wasserstein GAN with Wasserstein divergence (WGAN-div), which offers a more continuous loss function, thus ensuring that even when actions differ significantly from expert demonstrations, the gradients remain smooth. This characteristic allows for stable training throughout a broader state-action space, ultimately leading to more diverse action distributions.

D. Encourage Exploration via a Curiosity Bonus

Training robots to acquire diverse skills requires exploration of various joint angles, which is challenging to succinctly encapsulate with specific reward functions. To address this, we incorporate a curiosity reward r^c into the value function: $V = \mathbb{E} \sum_{h=0}^{+\infty} \gamma^t (r_h + r_h^c)$ This

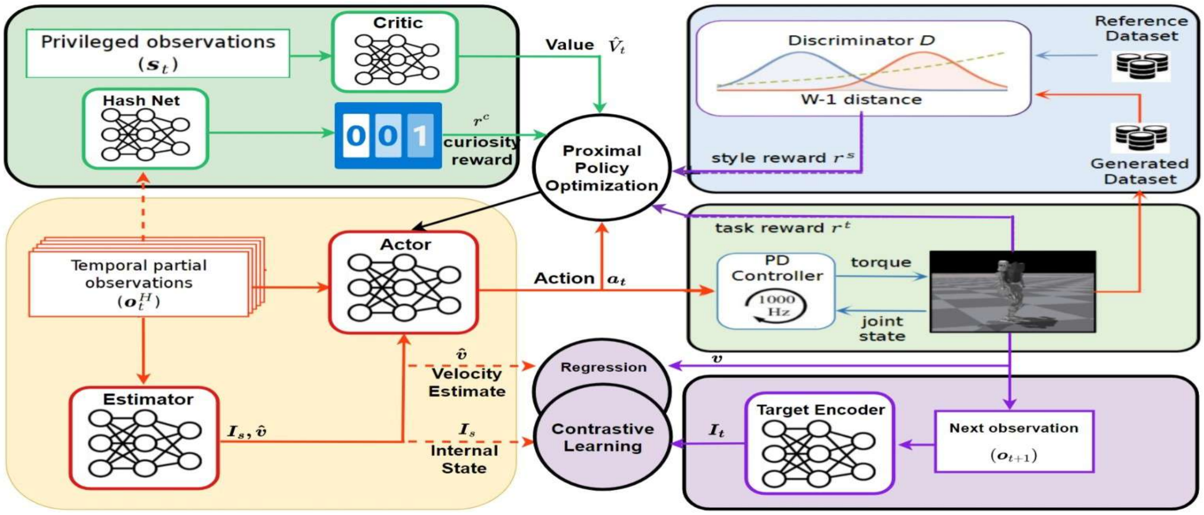


Fig. 3: Illustration of the proposed human-like locomotion learning framework. The estimator extracts information from past observations, producing a velocity estimate with internal state representation. The velocity estimate ensures mobile stability, while contrastive learning promotes future observation prediction. Imitation learning secures human-like gaits, and a curiosity bonus fosters exploration. The bottom-left block is employed in real-world deployment.

curiosity reward, rooted in bandit theory, is inversely related to $N(s_t = s, a_t = a)$, the count of historical visits to the state-action pair. Since reinforcement learning aims to maximize the value function, adding the r^c term encourages the policy to explore less-visited areas, promoting uniform exploration.

Curiosity Hash-net [14] extends this bonus design from bandit theory to continuous RL. We train a neural network $\phi(\cdot) : \mathcal{S} \rightarrow -1, 1^k$ to extract a low-dimensional binary representation of states, which we refer to as a Hashing value. The curiosity reward is then defined as $r_t^c(s_t) = \frac{1}{\sqrt{N(\phi(s_t))}}$ This adaptation retains the essence of the classical design while being suitable for continuous control environments.

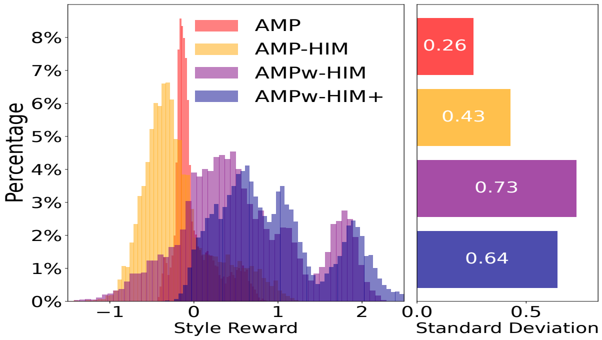


Fig. 4: Statistics of the style reward r^s obtained during training indicate the diversity of joint positions in the output actions. The RL module HIM aids in capturing a broader reward distribution (Left) with a larger standard deviation (Right). Wasserstein divergence (AMPw) captures multiple peaks, a feature not provided by the vanilla AMP design. Also note that while curiosity bonus reduces deviation, it shifts the distribution towards a higher-reward region.

V. Experiments

In this section, we present extensive ablation experiments to quantitatively support the effectiveness of the methods proposed in Sections IV-B, IV-C and IV-D. We center on showing how the HIM module, WGAN-div, and curiosity rewards help shape traditional IL algorithm (AMP) into a command-adaptive robot learning pipeline with strong generalization ability.

Section V-A is devoted to simulation experiments, while Section V-B demonstrates real-world performance. In the following sections, we denote by **AMP** the baseline agent trained without the RL module. It demonstrates the strong capability of IL in precisely mimicking movements, as well as the blatant limitation of learning to adapt to varying commands. **AMP-HIM** is the group trained with IL and HIM module simultaneously, while **AMPw-HIM** replaces the loss function of AMP-HIM from mean-square error to Wasserstein-divergence. Adding curiosity rewards to AMPw-HIM, we arrive at our full-stack approach, denoted as **AMPw-HIM+**. [1]

A. Simulation Experiments

We evaluate our proposed method in simulations based on the following four aspects, which are crucial yet distinct dimensions for assessing the vividness and effectiveness of a humanoid's movements.

1) **Basic Locomotion**: Evaluates how well the agents accomplish stable and agile locomotions that fit their hardware structure. It is reflected by task rewards such as body collision and feet slippery.
2) **Anthropomorphism**: Measures how similar the robot's gaits are to human data, which adds the vividness and naturalness to basic locomotion.
3) **Velocity Tracking**: Assesses how accurately the gait follows the required speeds, which represents mobile adaptability to changing commands.

[1] For completeness, we also include **AMPw**, whose agents are trained only with Imitation Learning with WGAN-div.

4) **Generalization**: Tests whether the algorithm is able to generalize to different robot structures, various simulated and real environments and perform unseen motions to complete multiple tasks.

1) Basic Locomotion: We train four groups of agents with the same reward and domain randomization levels and record the average task rewards in Fig 5. Ablation tests shown in Fig 6 further highlight the advantages of WGAN-div and curiosity rewards in adapting human motion patterns to the robot's form. Compared to **AMP,HIM**, WGAN produces a greater variety of joint angles for the agent to explore, while curiosity rewards help the RL policy escape local minima, optimizing within the newly explored state-action space. Thus, **AMPw,HIM** and **AMPw,HIM+** achieve significant increases in mean returns.

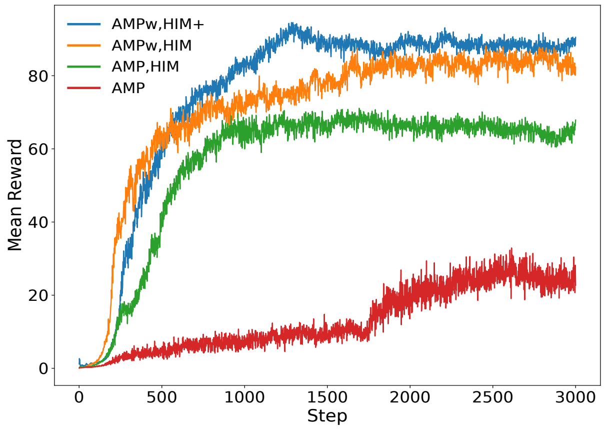


Fig. 5: Training curves of four algorithms. Vanilla AMP (red curve) fails to obtain satisfactory task rewards due to its limited generalization. RL-aided methods consistently outperforms AMP baseline.

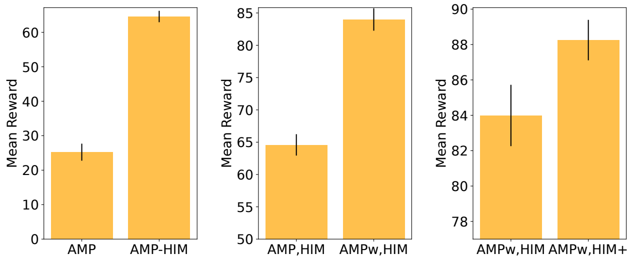


Fig. 6: Ablation study on the mean reward. Orange bars are the mean returns calculated in the last 500 iters when optimizer converges, error bars indicate their standard deviation. HIM module, WGAN-div criterion, and curiosity rewards significantly outperforms the counterparts with these structures removed.

2) Anthropomorphism: To measure the human-likeness of the generated motions, first we apply different commands to drive the agents to walk, jog, spin and walk backwards, capturing motion frames in 20 parallel experiments for each group. Then we compare the generated motions with expert data corresponding to the four movement types, calculating their similarity in joint positions using the Dynamic Time Warping (DTW) toolbox [5] in IsaacGym. The results are presented in Table IV. The baseline model, AMP, achieved the lowest DTW distances across all four movement categories, while the incorporation of HIM hurts human-like performance. This is not surprising, as RL solutions make adjustments to human gestures for better adaptability to the robot entity. However, we note that the introduction of WGAN and curiosity mitigates the negative impacts of HIM alone, correcting its joint positions to preserve human's flavor captured by the IL module. This correction effect is also evidenced by Figure 9 and 1.

	DTW Distance			
Method	Walking	Jogging	Walking Backwards	Turning Around
AMP	1160.53 ± 21.68	1488.90 ± 30.16	736.70 ± 4.02	990.19 ± 8.53
AMP,HIM	1417.92 ± 20.13	1638.21 ± 21.17	755.56 ± 2.07	1075.36 ± 4.06
AMPw,HIM	**1402.52 ± 23.55**	1640.28 ± 18.52	756.66 ± 1.81	1080.10 ± 4.68
AMPw,HIM+	1404.08 ± 20.32	**1637.08 ± 15.75**	**742.49 ± 1.65**	**1054.76 ± 3.64**

TABLE IV: Comparison of anthropomorphism between different strategies. A lower DTW score indicates stronger similarity with human expert data. Bold figures indicate the best performance in HIM-aided groups, which shows that WGAN-div and curiosity helps preserve human flavor in various locomotion tasks.

3) Velocity Tracking: Apart from comparing the ability to mimic separate human motions, we also test how the humanoids adapt to varying commanded velocities in IsaacGym, whose results are shown in Fig 7. The speed region of the reference motion files is concentrated within $[-0.4m/s, 0.4m/s]$. We apply velocity commands that sweep within $[-0.5m/s, 0.75m/s]$, making abrupt leaps during acceleration and deceleration. **AMP**-based methods output actions that stuck the robot within a small velocity region, while RL-aided groups such as **AMP,HIM** exhibit a significant increase in velocity tracking ability. To show that HIM alone is able to boost performance for various IL methods, we add an additional experiment with group **AMPw**, who follows commands even poorer than the AMP baseline, standing in sharp contrast with **AMPw,HIM**. Ablation studies

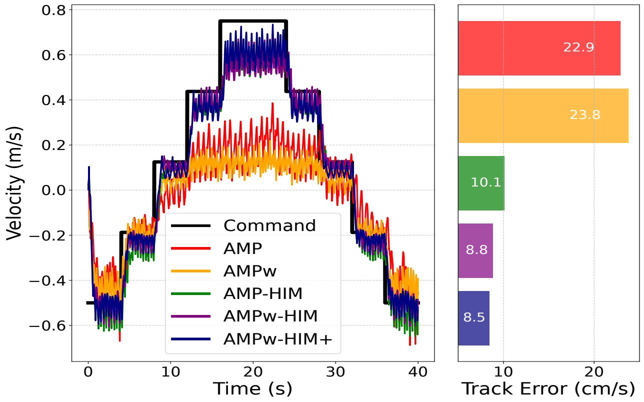


Fig. 7: Comparison of different algorithm's velocity tracking ability. Applied with abruptly changing velocities commands, pure IL methods are unable to follow unseen velocity requirements (Left), while HIM-aided policies do, yielding significantly smaller tracking error (Right).

in Fig 6 strongly prove that HIM greatly and consistently enhances the mobile adaptability upon different imitation learning methods. The success of HIM comes from its velocity and latent state estimators.

4) Generalization: Euipped with HIM and curiosity rewards, our method fosters strong generalization ability in various dimensions. **a) Hardware adaption.** Our method endows different robots with versatile locomotion ability, which includes Noetix N1 in Fig 1 and Dora in Fig 9. **b) Zero-shot transfer in various environments.** After training in IsaacGym, our robot agents are able to smoothly transfer to the other simulation envioronment MoJoCo, without any additional engineering. The transfer between simulators preserves agile locomotion ability, as demonstrated in Fig 8. Moreover, we can directly employ the same policy to robots in the real-world, using exactly the same set of PD controller parameters, which is shown in Fig 1. **c) Motion generalization and command-adaptive multitasking.** We witness direct evidence of motion generalization in the experiments. We trained agents only with expert data of straight walking and running, but the policies returned by **AMPw,HIM** are able to perform various unseen motions desired by various input commands, such as walking backwards, taking sidesteps, and spinning while running (cf. Fig 9). We explain this phenomenon with the experiment on style reward distribution. As is shown in Fig 4, WGAN-div captures the expert distribution with multiple peaks, indicating the generated joint positions possess stronger variety. This further implies RL-boosted imitation learning is able to break expert motion sequences down to basic joint position combinations with much finer granularity, thus helping agents assemble basic action units to form complex action suites, resulting in novel motions. **d) Harnessing complex terrains.** The contrastive learning module of HIM distinguished various different motion files and terrain properties, making our algorithm applicable to complex terrains like bumpy surfaces and descending planes, as shown in Fig 10.

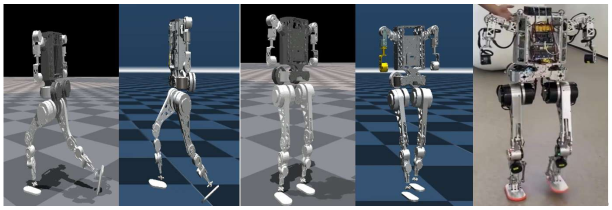
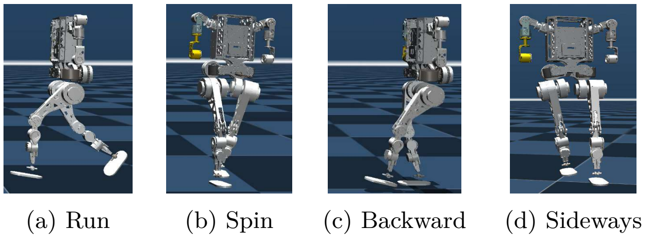


Fig. 8: Noetix Dora performance zero-shot transfer between simulator IsaacGym, MoJoCo and the real world, preserving motion versatility.

B. Real-world Evaluation

We also conduct extensive real-world experiments on robot Noetix N1. For example, in Fig. 1, the robot's gaits in the real-world highly align with other images in the two simulators, which demonstrates the strong robustness and generalization ability of our algorithm. Moreover, we show in Fig 11 that Noetix N1 is able to walk, run and stop in different environments following various command series.



(a) Run (b) Spin (c) Backward (d) Sideways

Fig. 9: Noetix Dora learns unseen body movements after zero-shot transfer to MuJoCo.

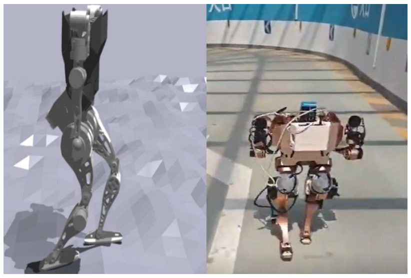


Fig. 10: Noetix N1 walks on bumpy terrains in IsaacGym (Left) and runs down to an underground parking lot (Right).

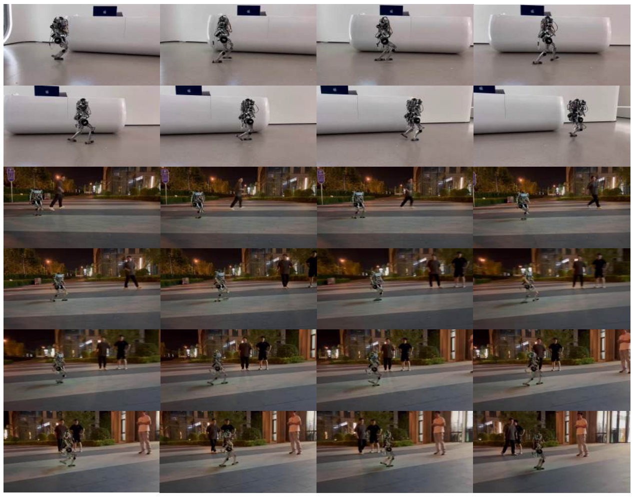


Fig. 11: Real-world experiments in different scenarios.

VI. Conclusion and Future Work

This work introduces a novel humanoid locomotion learning framework that smoothly transitions between human-like motions based on changing commands. By using a curiosity bonus and the WGAN-divergence criterion, our method enhances the AMP algorithm's generalization, while a hybrid internal model simultaneously tracks velocity and estimates unobserved states, significantly improving adaptability. Comprehensive simulations and real-world experiments validate the effectiveness of our approach.

We can extend existing findings in several directions, which include adapting our method to diverse terrains such as stairs and slopes and enhancing the versatility of robot motions by learning to hop or coordinate with hand manipulation tasks.

References

[1] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, “A survey of robot learning from demonstration,” *Robotics and Autonomous Systems*, vol. 57, no. 5, pp. 469–483, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921889008001772

[2] X. Gu, Y.-J. Wang, and J. Chen, “Humanoid-gym: Reinforcement learning for humanoid robot with zero-shot sim2real transfer,” 2024.

[3] J. Ho and S. Ermon, “Generative adversarial imitation learning,” *Advances in neural information processing systems*, vol. 29, 2016.

[4] 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.

[5] C. Li, M. Vlastelica, S. Blaes, J. Frey, F. Grimminger, and G. Martius, “Learning agile skills via adversarial imitation of rough partial demonstrations,” in *Conference on Robot Learning*. PMLR, 2023, pp. 342–352.

[6] Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, “Reinforcement learning for versatile, dynamic, and robust bipedal locomotion control,” 2024.

[7] J. Long, Z. Wang, Q. Li, J. Gao, L. Cao, and J. Pang, “Hybrid internal model: Learning agile legged locomotion with simulated robot response,” 2024. [Online]. Available: https://arxiv.org/abs/2312.11460

[8] T. Miki, J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, “Learning robust perceptive locomotion for quadrupedal robots in the wild,” *Science Robotics*, vol. 7, no. 62, p. eabk2822, 2022.

[9] I. M. A. Nahrendra, B. Yu, and H. Myung, “Dreamwaq: Learning robust quadrupedal locomotion with implicit terrain imagination via deep reinforcement learning,” 2023.

[10] 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*, vol. 40, no. 4, p. 1–20, July 2021. [Online]. Available: http://dx.doi.org/10.1145/3450626.3459670

[11] I. Radosavovic, T. Xiao, B. Zhang, T. Darrell, J. Malik, and K. Sreenath, “Real-world humanoid locomotion with reinforcement learning,” *Science Robotics*, vol. 9, no. 89, p. eadi9579, 2024. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.adi9579

[12] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, “Learning to walk in minutes using massively parallel deep reinforcement learning,” 2022.

[13] A. Tang, T. Hiraoka, N. Hiraoka, F. Shi, K. Kawaharazuka, K. Kojima, K. Okada, and M. Inaba, “Humanmimic: Learning natural locomotion and transitions for humanoid robot via wasserstein adversarial imitation,” 2024.

[14] H. Tang, R. Houthooft, D. Foote, A. Stooke, X. Chen, Y. Duan, J. Schulman, F. D. Turck, and P. Abbeel, “Exploration: A study of count-based exploration for deep reinforcement learning,” 2017. [Online]. Available: https://arxiv.org/abs/1611.04717

[15] J. Wu, G. Xin, C. Qi, and Y. Xue, “Learning robust and agile legged locomotion using adversarial motion priors,” *IEEE Robotics and Automation Letters*, 2023.

[16] Q. Zhang, P. Cui, D. Yan, J. Sun, Y. Duan, A. Zhang, and R. Xu, “Whole-body humanoid robot locomotion with human reference,” *arXiv preprint arXiv:2402.18294*, 2024.