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

HumanMimic: Learning Natural Locomotion and Transitions for Humanoid Robot via Wasserstein Adversarial Imitation

Annan Tang[1], Takuma Hiraoka[1], Naoki Hiraoka[1], Fan Shi[1,2], Kento Kawaharazuka[1], Kunio Kojima[1], Kei Okada[1] and Masayuki Inaba[1]

***Abstract*— Transferring human motion skills to humanoid robots remains a significant challenge. In this study, we introduce a Wasserstein adversarial imitation learning system, allowing humanoid robots to replicate natural whole-body locomotion patterns and execute seamless transitions by mimicking human motions. First, we present a unified primitive-skeleton motion retargeting to mitigate morphological differences between arbitrary human demonstrators and humanoid robots. An adversarial critic component is integrated with Reinforcement Learning (RL) to guide the control policy to produce behaviors aligned with the data distribution of mixed reference motions. Additionally, we employ a specific Integral Probabilistic Metric (IPM), namely the Wasserstein-1 distance with a novel soft boundary constraint to stabilize the training process and prevent model collapse. Our system is evaluated on a full-sized humanoid JAXON in the simulator. The resulting control policy demonstrates a wide range of locomotion patterns, including standing, push-recovery, squat walking, human-like straight-leg walking, and dynamic running. Notably, even in the absence of transition motions in the demonstration dataset, the robot showcases an emerging ability to transit naturally between distinct locomotion patterns as desired speed changes.**

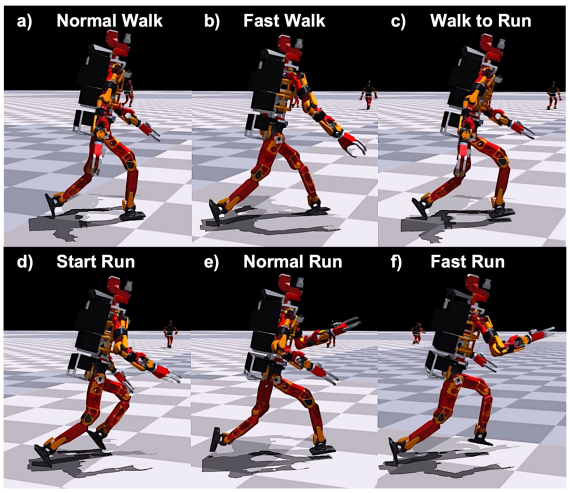

Fig. 1. Our Wasserstein adversarial imitation learning system enables a full-sized humanoid to exhibit various human-like natural locomotion behaviors and achieve seamless transitions as velocity command changes.

I. Introduction

Natural selection has shaped human ability, enabling humans to perform various locomotion behaviors and adeptly shift gait patterns in response to speed changes or external disturbances. Transferring the natural-looking locomotion and seamless transitions to humanoid robots remains a long-standing challenge, primarily due to the control complexity and intricacies of motion design.

While numerous studies based on simplified models [1] [2] [3] [4] and optimal control [5] [6] have demonstrated promising performance on structured locomotion paradigms, the intrinsically under-actuated and nonlinear characteristics of humanoids complicate the establishment of a unified model that accurately captures the dynamics across diverse gait transitions. On the other hand, deep reinforcement learning (deep RL) semi-automates the complex modeling process by maximizing the cumulative reward, leading to its growing popularity in developing advanced locomotion skills for quadrupedal robots [7] [8], bipedal robots [9] [10] and even humanoid robots [11] [12]. Nevertheless, RL-generated motions for high-DOF humanoids often exhibit undesired whole-body behaviors, including irregular arm swings, aggressive ground impacts, and unnatural gaits. Typical solutions utilize supplemental footstep planners [13], heuristic gait reward design [10] or pre-optimized gait and joint trajectory [14] to induce specific locomotion patterns. But given our limited understanding of the underlying characteristics that depict the natural behaviors of humans, these modules frequently employ basic principles like symmetry and energy minimization [15], resulting in more stereotypical robotic motions compared to humans.

For acquiring natural motions without the need for laborious reward engineering, the adversarial motion prior (AMP) [16] exploits an additional discriminator that outputs a style reward to encourage generated motions to resemble human demonstrations. In practice, discriminators trained with binary cross entropy (BCE) or least-square (LS) loss often face unstable training and model collapse, mainly due to the inadequacy of the metrics used to measure distances between non-overlapping probability distributions in high-dimensional spaces. In the closely related domain of generative adversarial networks (GANs), researchers have introduced several types of integral probability metrics (IPMs) [17] [18] [19], especially the Wasserstein distance [20] [21] [22], to address the challenges above. However, the unbounded Wasserstein distance [23] presents a significant challenge when trying to balance the style reward with other task-specific rewards like velocity tracking. Moreover,

[1]JSK Lab, Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan. tang@jsk.imi.i.u-tokyo.ac.jp
[2]AI Center, ETH Zürich, 8092 Zürich, Switzerland. fan.shi@ai.ethz.ch

the significant morphological differences between human demonstrators and humanoid robots, including joint configurations, body proportions, and bone hierarchies, pose challenges for the direct imitation of human demonstrations.

In this work, we present an adversarial imitation learning system that enables full-sized humanoids to autonomously acquire a variety of realistic locomotion behaviors through imitating human demonstrations. First, we introduce a unified primitive-skeleton motion retargeting approach to address morphological differences between arbitrary human demonstrators and humanoid robots. We exploit the power of the Wasserstein-1(W-1) distance, incorporating a novel soft boundary constraint, to ensure stable training dynamics and prevent the convergence of generated motions to a limited set of trivial modes. The learned one policy showcases a diverse array of robust and natural locomotion patterns, encompassing standing, push-recovery, squat walking, human-like straight-leg walking, dynamic running, and seamless transitions in response to changes in velocity commands, as shown in Figure 1. In short, our main contributions are: (i) Proposing an improved adversarial imitating learning system with Wasserstein critic and soft boundary constraints to address unstable training and model collapse. (ii) Detailing a unified primitive-skeleton motion retargeting technique applicable to arbitrary human skeleton sources and humanoid models. (iii) Achieving the whole-body natural locomotion and transitions for humanoids and evaluating the robustness through sim-to-sim settings in a high-fidelity simulator.

II. Related Works

RL for bipedal locomotion. Recent advancements in RL-based control strategies have significantly enhanced bipedal locomotion [24] [25]. For instance, the bipedal robot Cassie not only mastered versatile gait patterns through the use of periodic-parameterized reward functions [10] but also achieved the Guinness World Record for the fastest 100m dash using pre-optimized reference running gaits [14]. Jeon et al. [26] utilized potential-based reward shaping to ensure faster convergence and more robust humanoid locomotion. Shi et al. [27] integrated an assistive force curriculum into the learning process, allowing the acquisition of multiple agile humanoid motion skills in reference-free settings. In a more recent study, the full-sized humanoid HRP-5P [28] showcased robust walking using actuator current feedback, while Kim et al. [12] demonstrated a torque-based policy to bridge sim-to-real gaps. DeepMind [29] managed to instill agile soccer skills in a miniature humanoid via a two-stage teacher-student distillation and self-play. Additionally, attention-based transformers [11] have been employed to achieve more versatile locomotion in the humanoid Digit.

Motion imitation from real-world demonstrations. Leveraging reference motions from living creatures enables robots to acquire natural and versatile locomotion skills [30] [31] that are challenging to define manually. A predominant imitation strategy involves tracking either reference joint trajectories [32] [16] [14] or extracted gait features [33] [34]. However, these explicit tracking techniques, often limited to separate motion clips, can disrupt smooth transitions between different locomotion patterns. Drawing inspiration from generative adversarial imitation learning (GAIL) [35], Peng et al. introduced AMP [36] and successor ASE [37]. These approaches empower physics-based avatars to carry out objective tasks while simultaneously imitating the underlying motion styles from extensive unstructured datasets in an implicit manner. Variants of AMP have been further employed for learning agile quadrupedal locomotion [23] [38] [39] and terrain-adaptive skills [40] [41], exemplifying its efficacy in eliminating the need for intricate reward function design.

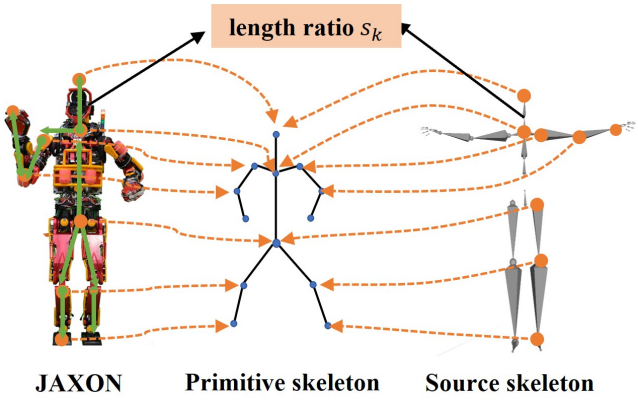

Fig. 2. Binding the humanoid JAXON and the MoCap skeleton involves merging their bones into a common primitive skeleton.

Despite the advancements in other domains, methods similar to AMP have not been extensively explored for humanoid robots. To bridge this gap, in this work, we present a Wasserstein adversarial imitation system with soft boundary constraints as an enhancement to the existing AMP techniques. We aim to provide a foundational training algorithm for future deployment on full-sized humanoid robots in real-world scenarios.

III. Motion Retargeting

To transfer reference motion to the robot, certain retargeting methods [42], [43] consider both kinematic and dynamic constraints, requiring accurate dynamic modeling or complex balance controllers. In this section, we detail a flexible motion retargeting approach based on the unified primitive skeleton, emphasizing geometry consistency. The kinematic and dynamic constraints such as feet contact state and balance will be satisfied in the reinforcement learning paradigm in the next section. Our retargeting involves four key procedures.

Unified primitive skeleton binding. Skeletal structures of humans and humanoid robots are known to correspond to homeomorphic graphs [44]. Leveraging this property, we extract what we term *primitive skeleton* that encapsulates the foundational geometric and hierarchical characteristics shared across various skeletons. In the process of primitive skeleton binding, we first construct kinematic trees for all involved skeletons. These trees are subsequently merged into

a unified primitive skeleton, retaining only a single bone between two successive key joints. Users manually select n key joints, offering an intuitive and flexible mechanism for loose binding between the source and target skeleton groups. Once binding is complete, we compute the length ratio $S = \{s_k \mid k \in \{1, \ldots, n\}\}$ for each bone within the primitive skeleton. An illustrative example of this binding between the Humanoid JAXON [45] and CMU MoCap [46] skeleton is presented in Figure 2.

Coordinate transformation. We consider the MoCap dataset represented by the predefined skeletal kinematic tree $L' = \{l'_{ij}\}$ and the corresponding motion sequence of T frames $M'_s = \{m'_t \mid t \in \{1, \ldots, T\}\}$, with frame t as $m'_t = ({}^w P'_r, {}^w R'_r, {}^0 R'_1, \ldots, {}^{j-1} R'_j)$. Here, l'_{ij} is the bone length between directly connected joints i and j, ${}^w P'_r$ and ${}^w R'_r$ represent the root's position and orientation w.r.t the world coordinate, and ${}^{j-1} R'_j$ indicates the local orientation of the source skeleton's joint j w.r.t its parent joint. Applying iterative homogeneous transformations along the given kinematic tree L', denoted by ${}^w P'_j = H\left({}^w P'_r, {}^w R'_r, {}^0 R'_1, \ldots, {}^{j-1} R'_j, L'\right)$, we derive the global position for each joint in the source skeletons. The relative position vector between adjacent key joints is computed as $\vec{r_k}' = {}^w P'_k - {}^w P'_{k-1}$. We scale it by $\vec{r}_k = s_k \cdot \vec{r_k}'$ to get the relative position vector $\vec{r}_k$ in target robot skeleton. Finally, we sum up the relative position vector along the kinematic chains and apply a transformation to get the key joint Cartesian positions w.r.t the root of robot skeletons as ${}^r P_k = H\left(\sum_{i=1}^{k} \vec{r}_i\right)$. All the end-effector poses ${}^r P_e \in \mathbb{R}^3 \times \mathbb{SO}(3)$ are incorporated into the final robot motion frame $m_t = ({}^w P_r, {}^w R_r, {}^r P_k, {}^r P_e)$, here e denotes wrists, feet and head.

Multi-Objective inverse kinematics. To map the key joint Cartesian position ${}^r P_k$ and end-effector pose ${}^r P_e$ to joint positions $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$, we formulate the whole-body inverse kinematics as a gradient-based optimization problem [47] with three goals,

$$\begin{aligned} C_1 &= \sum_k \left\| {}^r P_k - p_k(\theta) \right\|^2, \\ C_2 &= \sum_e \left\| {}^r P_e - p_e(\theta) \right\|^2, \\ C_3 &= \left\| \theta_t - \theta_{t-1} \right\|^2, \end{aligned} \tag{1}$$

where the $p_k(\theta)$ and $p_e(\theta)$ are the calculated Cartesian position and pose for joints and end-effectors during gradient descent iterations. The main goals consist of the **position goal** C_1 for all key joints and the **pose goal** C_2 for the end-effectors including hands, foot soles and head. An additional **minimal displacement goal** C_3 is introduced to maintain each joint variable close to the previous motion frames. This is crucial for the highly redundant humanoids as multiple solutions might satisfy C_1 and C_2. The overall objective function is the weighted sum of each individual goal cost,

$$C = \arg min_\theta \sum_i \kappa_i C_i(\theta). \tag{2}$$

The weights κ_i are determined as $(1, 1, 0.2)$ heuristically. The joint position and velocity limitations are incorporated into the constraints,

$$\begin{aligned} \theta_{\min} \le \theta_t \le \theta_{\max}, \\ \dot{\theta}_{\min} \le \frac{\theta_t - \theta_{t-1}}{\Delta t} \le \dot{\theta}_{\max}. \end{aligned} \tag{3}$$

Post-Processing. We compute root and joint velocities from sequential frame differences. Linear and Slerp interpolation are applied separately for position and orientation between discrete motion frames. Moreover, an exponential moving average filter is applied to smooth position and velocity spikes.

IV. Wasserstein Adversarial Imitation

Our Wasserstein adversarial imitation learning framework, as illustrated in Figure 3, incorporates actor-critic networks, a Wasserstein critic, and a motor-level Proportional Derivative (PD) controller. The actor updates the network parameters using policy gradients derived from both environment rewards and the Wasserstein critic. The Wasserstein critic undergoes adversarial training based on the Wasserstein-1 distance complemented by a soft boundary loss. When given a user-defined velocity command, our setup enables humanoids to follow the velocity, ensuring smooth transitions in locomotion.

A. Velocity-Conditioned Reinforcement Learning

We formulate the humanoid locomotion control as a velocity-goal-conditioned [48] Markov decision process, with the velocity goal $v^* \sim p(v) \in \mathcal{V}$, state $s \in \mathcal{S}$, action $a \sim \pi(\cdot|s, v^*) \in \mathcal{A}$, reward $r = r(s, a, v^*)$ and discount factor $\gamma \sim (0, 1]$. The agent updates the decision policy π through interactions with the surrounding environments to maximize the expected discounted return under the condition of the desired velocity,

$$J(\pi) = \mathbb{E}_{v^* \sim p(v), \tau \sim p(\cdot|\pi, v^*)} \left[\sum_t \gamma^t r\left(s_t, a_t, v^*\right) \right]. \tag{4}$$

The total reward terms are composed of two components:(1) velocity-tracking reward r^V, (2) style reward r^S,

$$r_t = \mu_1 r^V + \mu_2 r^S, \tag{5}$$

where μ_1 and μ_2 denote the combination weight on each term. Reward r^V encourages the robot to follow the commanded CoM velocities, it is designed as the normalized exponential errors of linear velocity v^*_{xy} and heading velocity w^*_z separately,

$$\begin{aligned} r^V =& \beta_1 \exp\left(-\frac{\left\| v^*_{xy} - v_{xy} \right\|^2}{\lambda_l |v^*_{xy}|} \right) \\ &+ \beta_2 \exp\left(-\frac{\left\| w^*_z - w_z \right\|^2}{\lambda_h |w^*_z|} \right), \end{aligned} \tag{6}$$

β_1 and β_2 are hyper-parameters to control the importance of each tracking error. The parameters λ_l and λ_h are utilized to regulate the tracking precision. Smaller values of λ_l and λ_h encourage the humanoid robots to achieve better velocity-following precision but may make it more difficult for the policy to receive rewards at the beginning of training.

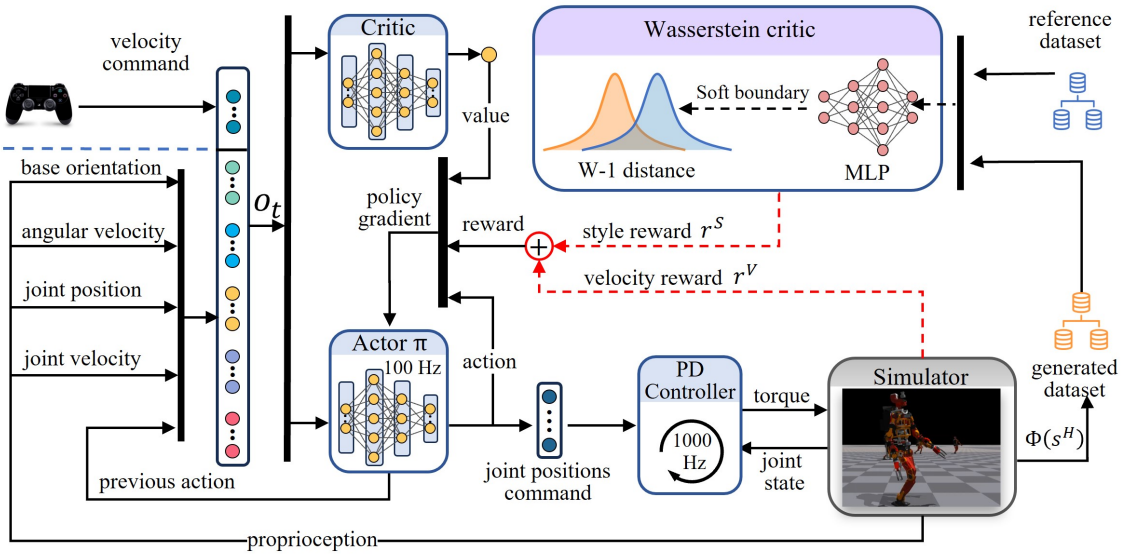

Fig. 3. **Wasserstein Adversarial Imitation Framework.** Given the robot's proprioceptive state and base velocity commands, the policy network predicts the joint position targets. A PD controller converts these targets into torques to actuate the robot. Using the reference motion dataset and policy-generated motion dataset, the Wasserstein critic updates its parameters through the soft-boundary Wasserstein-1 loss during training and predicts the style reward during roll-out. The style reward r^S is combined with the velocity reward r^V to guide policy training.

We model the actuation dynamics as a mass-damping system. A PD controller is employed to map the actions to desired torques with the target joint velocity always specified as 0.

B. Wasserstein Critic

In adversarial imitation learning, it is pivotal for the discriminator to offer an appropriate distance metric between the generated data distribution $\mathcal{Q}$ and the reference data distribution $\mathcal{P}$. In vanilla GAIL [35], the discriminator employs BCE loss, which has been shown to equate to minimizing the Jensen-Shannon Divergence [49]. When there is no overlap between two high-dimensional data distributions, it can result in gradient vanishing, severely causing unstable training and model collapse. IPMs have been proven as excellent distance measures on probabilities [17],

$$\Gamma_{\mathcal{F}}(\mathcal{P},\mathcal{Q}) := \sup_{f\in\mathcal{F}} \left| \int_{\mathcal{M}} f d\mathcal{P} - \int_{\mathcal{M}} f d\mathcal{Q} \right|, \tag{7}$$

where $\mathcal{F}$ represents a class of real-valued bounded measurable functions on Manifold $\mathcal{M}$. When $\mathcal{F} = \{f : \|f\|_L \leq 1\}$, it forms the dual representation of Wasserstein-1 distance and the typical Wasserstein loss with gradient penalty [20] becomes

$$\begin{aligned} \arg\min_{\theta} & -\mathbb{E}_{x\sim P_r}\left[D_\theta(x)\right] + \mathbb{E}_{\tilde{x}\sim P_g}\left[D_\theta(\tilde{x})\right] \\ & + \lambda \mathbb{E}_{\hat{x}\sim P_{\hat{x}}}\left[\left(\|\nabla_{\hat{x}} D_\theta(\hat{x})\|_2 - 1\right)^2\right], \end{aligned} \tag{8}$$

where $D_\theta(\cdot)$ denotes outputs from the Wasserstein critic. $x = \Phi\left(s^N\right)$ is the manually selected feature from N consecutive motions s^N that are sampled from the reference and generated motion distribution P_r and P_g. $\hat{x} = \alpha x + (1-\alpha)\tilde{x}$ are samples obtained through random interpolation between the reference samples x and generated samples $\tilde{x}$.

Soft boundary constraint. The Wasserstein critic network is used to approximate a cluster of Lipschitz-constrained functions with a linear combination architecture in the final layers. As a result, the output value is unbounded and unbiased [50] [51]. Utilizing original Wasserstein loss (8), we observed drawbacks stemming from the unbounded values. At the early training stage, there are significant differences between generated samples and real data distributions, the critic's output for generated samples converges quickly to large negative values. This renders the style reward r^S nearly zero, causing the policy to fail to learn natural motions. The unbounded value also introduces large standard deviations in style reward, which makes the training unstable. To limit the outputs from the Wasserstein critic, we modify the Wasserstein loss with a soft boundary constraint,

$$\begin{aligned} \arg\min_{\theta} & -\mathbb{E}_{x\sim P_r}\left[\tanh(\eta D_\theta(x))\right] \\ & + \mathbb{E}_{\tilde{x}\sim P_g}\left[\tanh(\eta D_\theta(\tilde{x}))\right] \\ & + \lambda \mathbb{E}_{\hat{x}\sim P_{\hat{x}}}\left[\left(\max\{0, \|\nabla_{\hat{x}} D_\theta(\hat{x})\|_2 - 1\}\right)^2\right], \end{aligned} \tag{9}$$

Where η is a hyperparameter that controls the range of boundaries. Smaller η means softer constraints that generate larger critic values. In practice, $\eta \sim (0.1, 0.5)$ is a proper range for selection. We apply a weaker gradient penalty [52] to further stabilize the training. Finally, the style reward is designed as $r^S = e^{D_\theta(\tilde{x})}$.

V. Experiment

A. Implementation Details

Actor-Critic observation space. The actor and critic networks share the same observation space. The observation space $O_{ac} \in \mathbb{R}^{102}$ consists of: (i) Base angular velocity $w_b \in \mathbb{R}^3$ expressed in base local frame. (ii) Velocity command $v^* \in \mathbb{R}^3$, including target linear velocity $v^*_{xy} \in [0, 5]$

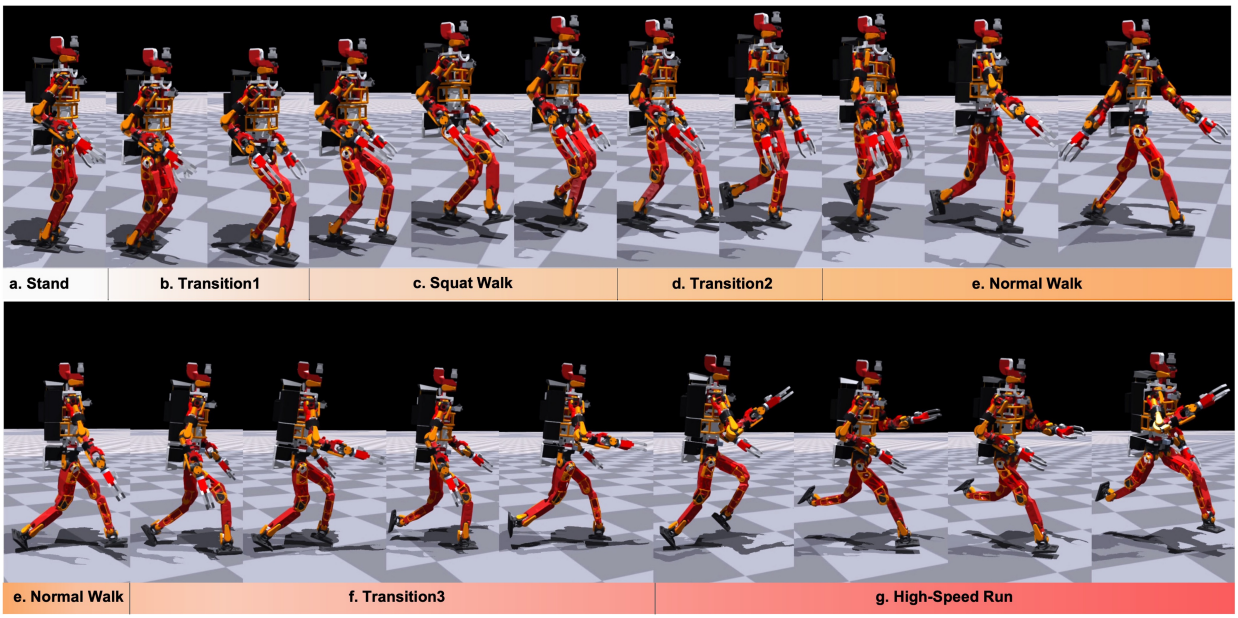

Fig. 4. Snapshots of various natural locomotion behaviors learned by the Humanoid JAXON. As the velocity command increases from 0 m/s to 5 m/s, the robot exhibits seamless transitions from standing to dynamic running.

m/s and heading velocity $w_z^* \in [-1, 1]$ rad/s. (iii) The gravity vector $z_b \in \mathbb{R}^3$ expressed in base local frame. (iv) Current joint position $\theta \in \mathbb{R}^{31}$. (v) Current joint velocity $\dot{\theta} \in \mathbb{R}^{31}$. (vi) Last-step actions $a_{t-1} \in \mathbb{R}^{31}$.

Wasserstein-Critic observation space and action space. The observation space O_{wc} of Wasserstein critic consists of state-transition pairs $\Phi\left(s^N\right) = (s_i, \ldots, s_{i+N}) \in \mathbb{R}^{78\times(N+1)}$ in N preceding time-steps. Each s_i is represented in the same style feature space where the style features are carefully hand-selected. The motion style feature $s_i \in \mathbb{R}^{78}$ is composed of: (i) Base height $p_z \in \mathbb{R}$. (ii) Base linear velocity $v_b \in \mathbb{R}^3$ expressed in base local frame. (iii) Base angular velocity $w_b \in \mathbb{R}^3$ expressed in base local frame. (iv) The gravity vector $z_b \in \mathbb{R}^3$ expressed in base local frame. (v) Joint position $\theta \in \mathbb{R}^{31}$. (vi) Joint velocity $\dot{\theta} \in \mathbb{R}^{31}$. (vii) Relative position of feet with base $r_{\text{feet}} \in \mathbb{R}^6$. The corresponding action space $\mathcal{A} \in \mathbb{R}^{31}$ of policy is chosen as 31 target joint positions within the joint angle limitation.

Reference motion dataset. The reference motion dataset includes multiple locomotion patterns. Table I depicts the Statistics details of the whole dataset used for training. Normal walk and squat walk are retargeted from CMU-MoCap dataset [46] and SFU-MoCap dataset [53]. The standstill motion is manually designed and the squat walk motion is recorded from the existing robot controller [54].

Regularization terms and domain randomization. To obtain a high-fidelity controller, we impose regularization penalties for large action jerks, significant joint torque, and acceleration. We also employ domain randomization on contact friction, restitution, joint friction, joint inertia, mass parameters, PD gains, and motor strength to avoid overfitting the environmental dynamics.

Training details. The actor, critic, and Wasserstein critic have the same MLP structures with [1024, 512, 256] hidden units and ELU activation functions. Policies update via PPO [55] with a learning rate of $l = 3.0e-5$ and around 30 hours of training in Isaac gym [56] with an NVIDIA 3090Ti.

TABLE I
STATISTICS OF REFERENCE DATA

Data Type	Duration(s)	Velocity(m/s)	Selection Probability
Stand	5.1	[0.0, 0.0]	0.15
Squat Walk	8.0	[0.2, 0.6]	0.20
Normal Walk	14.2	[0.9, 2.4]	0.35
Run	15.3	[2.9, 4.8]	0.30
Total	42.6	[0.0, 4.8]	1.00

B. Evaluation

Natural locomotion and transition. We examined the robot's ability to reproduce a range of natural locomotion behaviors from the reference dataset and to adapt to velocity commands. We set the initial desired velocity to 0 m/s and gradually increased it to 5 m/s with a constant acceleration of 0.1 m/s^2. Figure 4 presents a side view of JAXON robot's locomotion behaviors in response to changing velocities. The results indicate that our control policy not only captures diverse locomotion patterns from the reference dataset but also enables smooth transitions not present in the reference motions. The velocity-tracking curve and the z-direction feet contact force are shown in Figure 5. The velocity tracking curve in Figure 5 demonstrates the robot's capability to closely follow the desired velocity, reaching speeds of up to 5 m/s. It's important to note that this velocity tracking refers to the average velocity over a gait cycle, as opposed to instantaneous velocity. As the speed increases, the variance

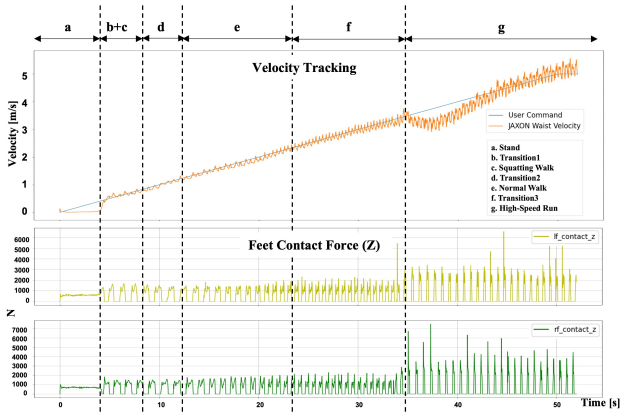

Fig. 5. Top: the velocity tracking curve, where the velocity command increases from 0.0 m/s to 5.0 m/s with a constant acceleration of 0.1 m/s^2. Middle and bottom: the contact forces in the z-direction for the left and right feet during the transition from standing to running.

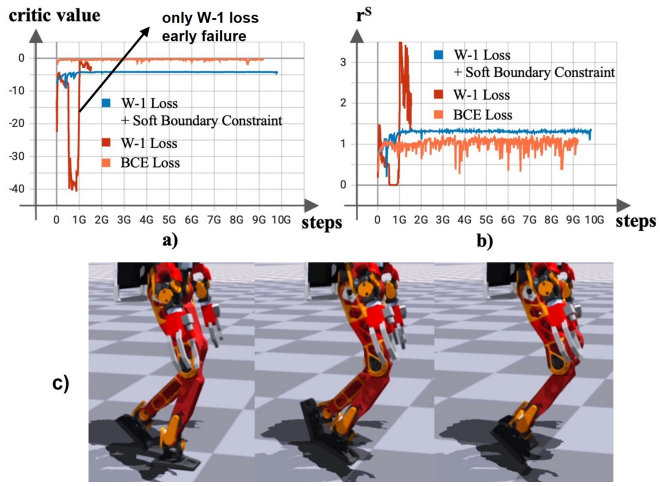

Fig. 6. a) Comparison of discriminator (critic) output values. b) Comparison of style reward values. Using only the Wasserstein-1 loss results in a wide range and significant fluctuations in both output and style reward, causing early-stage training failures. While employing the BCE loss keeps these values within a suitable range, it also leads to considerable relative fluctuations and susceptibility to model collapse and unstable training. In contrast, the Wasserstein-1 loss with soft boundary constraint ensures both the output and style reward remain within an appropriate range and exhibit minimal fluctuations, leading to a more stable training process. c) An example of model collapse with BCE loss is that the robot only learned a tiptoe walking gait close to the standing posture.

in instantaneous velocity also increases. The contact force increases dramatically with the increase in speed. During the transition from walking stage f to running stage g, there is a significant increase in stride frequency and a substantial decrease in contact time. As the robot transitions into the running gait pattern, we can observe the presence of the air phase.

Training stability and model collapse. To assess the utility of the soft-boundary-constrained Wasserstein loss, we conducted three separate training sessions, each utilizing identical hyperparameters and random seeds, but varying discriminator loss types. As depicted in Figure 6a and 6b, the contrasts in discriminator (critic) outputs and style rewards

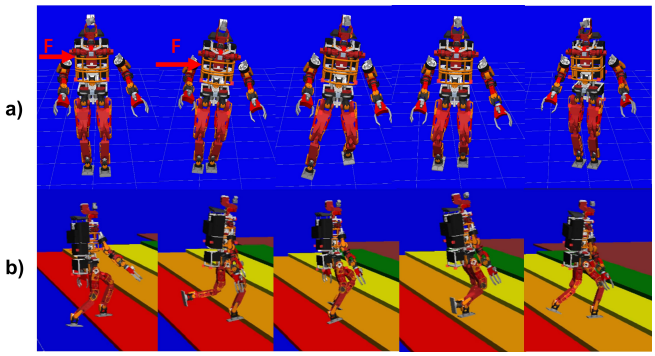

Fig. 7. Sim-to-sim robust test in high-fidelity Choreonoid simulator. a) Push-recovery: The robot takes one lateral step with its left foot to maintain balance. b) Stair-climbing: The robot navigates a set of stairs with each step height of 50mm.

are evident. With the W-1 loss, the critic's output experiences significant fluctuations spanning a broad range. This causes rapid changes in the style reward r^S during the initial training phases, culminating in a failed training attempt. Conversely, while the discriminator output using the BCE loss remains between (0,1), it still exhibits considerable relative fluctuation, resulting in volatile changes to the style reward and destabilizing the training phase. Our novel soft-boundary-constrained Wasserstein loss effectively constrains the output value within a more acceptable range and also minimizes the fluctuation in style reward, thus enhancing training stability. Beyond stability, the Wasserstein critic delivers improved assessments of distributional distances, which ultimately curtails model collapse and aberrant locomotion behaviors.

Sim-to-Sim robust test. The Choreonoid [57], integrated with real-time-control software Hrpsys, has been widely used in our previous work [58] [59] [60] and has proven to be a high-fidelity simulation environment with a small reality gap. We successfully transferred the policy from Isaac Gym to Choreonoid to facilitate future sim-to-real experiments. As depicted in Figure 7, the controller demonstrates extraordinary robustness in push-recovery and blind stair-climbing tasks.

VI. Conclusion

In this work, we present a Wasserstein adversarial imitation learning system capable of acquiring a variety of natural locomotion skills from human demonstration datasets with diverse motion behaviors. We have detailed a unified primitive-skeleton motion retargeting method, proficient in efficiently mapping motions between skeletons with significant morphological differences. Our findings highlight the system's novel ability to seamlessly transition between unique locomotion patterns as the desired speed varies, even though such transition behaviors are conspicuously absent in the reference dataset. Further experiments validate that our proposed soft-boundary-constrained Wasserstein-1 loss significantly stabilizes the training process and reduces the risk of model collapse. In the further, we aim to transfer this policy to real-world robots, to achieve versatile, natural, and dynamic locomotion for humanoids.

REFERENCES

[1] K. Miura, M. Morisawa, F. Kanehiro, S. Kajita, K. Kaneko, and K. Yokoi, "Human-like walking with toe supporting for humanoids," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2011, pp. 4428–4435.

[2] P. M. Wensing and D. E. Orin, "High-speed humanoid running through control with a 3d-slip model," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 5134–5140.

[3] T. Kamioka, H. Kaneko, M. Kuroda, C. Tanaka, S. Shirokura, M. Takeda, and T. Yoshiike, "Dynamic gait transition between walking, running and hopping for push recovery," in *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*. IEEE, 2017, pp. 1–8.

[4] T. Sugihara, K. Imanishi, T. Yamamoto, and S. Caron, "3d biped locomotion control including seamless transition between walking and running via 3d zmp manipulation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 6258–6263.

[5] K. Ishihara, T. D. Itoh, and J. Morimoto, "Full-body optimal control toward versatile and agile behaviors in a humanoid robot," *IEEE Robotics and Automation Letters*, vol. 5, no. 1, pp. 119–126, 2019.

[6] M. Chignoli, D. Kim, E. Stanger-Jones, and S. Kim, "The mit humanoid robot: Design, motion planning, and control for acrobatic behaviors," in *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2021, pp. 1–8.

[7] 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.

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

[9] Z. Xie, P. Clary, J. Dao, P. Morais, J. Hurst, and M. Panne, "Learning locomotion skills for cassie: Iterative design and sim-to-real," in *Conference on Robot Learning*. PMLR, 2020, pp. 317–329.

[10] J. Siekmann, Y. Godse, A. Fern, and J. Hurst, "Sim-to-real learning of all common bipedal gaits via periodic reward composition," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 7309–7315.

[11] I. Radosavovic, T. Xiao, B. Zhang, T. Darrell, J. Malik, and K. Sreenath, "Learning humanoid locomotion with transformers," *arXiv preprint arXiv:2303.03381*, 2023.

[12] D. Kim, G. Berseth, M. Schwartz, and J. Park, "Torque-based deep reinforcement learning for task-and-robot agnostic learning on bipedal robots using sim-to-real transfer," *IEEE Robotics and Automation Letters*, vol. 8, no. 10, pp. 6251–6258, 2023.

[13] R. P. Singh, M. Benallegue, M. Morisawa, R. Cisneros, and F. Kanehiro, "Learning bipedal walking on planned footsteps for humanoid robots," in *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*. IEEE, 2022, pp. 686–693.

[14] D. Crowley, J. Dao, H. Duan, K. Green, J. Hurst, and A. Fern, "Optimizing bipedal locomotion for the 100m dash with comparison to human running," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 12 205–12 211.

[15] W. Yu, G. Turk, and C. K. Liu, "Learning symmetric and low-energy locomotion," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–12, 2018.

[16] X. B. Peng, E. Coumans, T. Zhang, T.-W. Lee, J. Tan, and S. Levine, "Learning agile robotic locomotion skills by imitating animals," *arXiv preprint arXiv:2004.00784*, 2020.

[17] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. Lanckriet, "On integral probability metrics,\phi-divergences and binary classification," *arXiv preprint arXiv:0901.2698*, 2009.

[18] C. L. Li, W. C. Chang, Y. Cheng, Y. Yang, and B. Póczos, "Mmd gan: Towards deeper understanding of moment matching network," *Advances in neural information processing systems*, vol. 30, 2017.

[19] Y. Mroueh and T. Sercu, "Fisher gan," *Advances in neural information processing systems*, vol. 30, 2017.

[20] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *Advances in neural information processing systems*, vol. 30, 2017.

[21] R. Dadashi, L. Hussenot, M. Geist, and O. Pietquin, "Primal wasserstein imitation learning," in *ICLR 2021-Ninth International Conference on Learning Representations*, 2021.

[22] I. Durugkar, M. Tec, S. Niekum, and P. Stone, "Adversarial intrinsic motivation for reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8622–8636, 2021.

[23] 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.

[24] Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Robust and versatile bipedal jumping control through multi-task reinforcement learning," *arXiv preprint arXiv:2302.09450*, 2023.

[25] J. Siekmann, K. Green, J. Warila, A. Fern, and J. Hurst, "Blind bipedal stair traversal via sim-to-real reinforcement learning," *arXiv preprint arXiv:2105.08328*, 2021.

[26] S. H. Jeon, S. Heim, C. Khazoom, and S. Kim, "Benchmarking potential based rewards for learning humanoid locomotion," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9204–9210.

[27] F. Shi, Y. Kojio, T. Makabe, T. Anzai, K. Kojima, K. Okada, and M. Inaba, "Reference-free learning bipedal motor skills via assistive force curricula," in *The International Symposium of Robotics Research*. Springer, 2022, pp. 304–320.

[28] R. P. Singh, Z. Xie, P. Gergondet, and F. Kanehiro, "Learning bipedal walking for humanoids with current feedback," *IEEE Access*, 2023.

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

[30] S. Bohez, S. Tunyasuvunakool, P. Brakel, F. Sadeghi, L. Hasenclever, Y. Tassa, E. Parisotto, J. Humplik, T. Haarnoja, R. Hafner *et al.*, "Imitate and repurpose: Learning reusable robot movement skills from human and animal behaviors," *arXiv preprint arXiv:2203.17138*, 2022.

[31] L. Han, Q. Zhu, J. Sheng, C. Zhang, T. Li, Y. Zhang, H. Zhang, Y. Liu, C. Zhou, R. Zhao *et al.*, "Lifelike agility and play on quadrupedal robots using reinforcement learning and generative pre-trained models," *arXiv preprint arXiv:2308.15143*, 2023.

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

[33] F. Yin, A. Tang, L. Xu, Y. Cao, Y. Zheng, Z. Zhang, and X. Chen, "Run like a dog: Learning based whole-body control framework for quadruped gait style transfer," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 8508–8514.

[34] D. Kang, F. De Vincenti, N. C. Adami, and S. Coros, "Animal motions on legged robots using nonlinear model predictive control," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 11 955–11 962.

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

[36] 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.

[37] X. B. Peng, Y. Guo, L. Halper, S. Levine, and S. Fidler, "Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters," *ACM Transactions On Graphics (TOG)*, vol. 41, no. 4, pp. 1–17, 2022.

[38] A. Escontrela, X. B. Peng, W. Yu, T. Zhang, A. Iscen, K. Goldberg, and P. Abbeel, "Adversarial motion priors make good substitutes for complex reward functions," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 25–32.

[39] E. Vollenweider, M. Bjelonic, V. Klemm, N. Rudin, J. Lee, and M. Hutter, "Advanced skills through multiple adversarial motion priors in reinforcement learning," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 5120–5126.

[40] 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.

[41] Y. Wang, Z. Jiang, and J. Chen, "Amp in the wild: Learning robust, agile, natural legged locomotion skills," *arXiv preprint arXiv:2304.10888*, 2023.

[42] 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.
[43] R. Grandia, F. Farshidian, E. Knoop, C. Schumacher, M. Hutter, and M. Bächer, "Doc: Differentiable optimal control for retargeting motions onto legged robots," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–14, 2023.
[44] K. Aberman, P. Li, D. Lischinski, O. Sorkine-Hornung, D. Cohen-Or, and B. Chen, "Skeleton-aware networks for deep motion retargeting," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 62–1, 2020.
[45] K. Kojima, T. Karasawa, T. Kozuki, E. Kuroiwa, S. Yukizaki, S. Iwaishi, T. Ishikawa, R. Koyama, S. Noda, F. Sugai *et al.*, "Development of life-sized high-power humanoid robot jaxon for real-world use," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2015, pp. 838–843.
[46] CMU, "Cmu graphics lab motion capture database," http://mocap.cs.cmu.edu/.
[47] S. Starke, N. Hendrich, and J. Zhang, "Memetic evolution for generic full-body inverse kinematics in robotics and animation," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 3, pp. 406–420, 2018.
[48] E. Chane-Sane, C. Schmid, and I. Laptev, "Goal-conditioned reinforcement learning with imagined subgoals," in *International Conference on Machine Learning*. PMLR, 2021, pp. 1430–1440.
[49] L. Ke, S. Choudhury, M. Barnes, W. Sun, G. Lee, and S. Srinivasa, "Imitation learning as f-divergence minimization," in *Algorithmic Foundations of Robotics XIV: Proceedings of the Fourteenth Workshop on the Algorithmic Foundations of Robotics 14*. Springer, 2021, pp. 313–329.
[50] M. Zhang, Y. Wang, X. Ma, L. Xia, J. Yang, Z. Li, and X. Li, "Wasserstein distance guided adversarial imitation learning with reward shape exploration," in *2020 IEEE 9th Data Driven Control and Learning Systems Conference (DDCLS)*. IEEE, 2020, pp. 1165–1170.
[51] I. Kostrikov, K. K. Agrawal, D. Dwibedi, S. Levine, and J. Tompson, "Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning," *arXiv preprint arXiv:1809.02925*, 2018.
[52] H. Petzka, A. Fischer, and D. Lukovnikov, "On the regularization of wasserstein gans," in *International Conference on Learning Representations*, 2018.
[53] SFU, "Sfu motion capture database," https://mocap.cs.sfu.ca/.
[54] Y. Kojio, Y. Omori, K. Kojima, F. Sugai, Y. Kakiuchi, K. Okada, and M. Inaba, "Footstep modification including step time and angular momentum under disturbances on sparse footholds," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4907–4914, 2020.
[55] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
[56] 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.
[57] S. Nakaoka, "Choreonoid: Extensible virtual robot environment built on an integrated gui framework," in *2012 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, 2012, pp. 79–85.
[58] Y. Kakiuchi, K. Kojima, E. Kuroiwa, S. Noda, M. Murooka, I. Kumagai, R. Ueda, F. Sugai, S. Nozawa, K. Okada *et al.*, "Development of humanoid robot system for disaster response through team nedo-jsk's approach to darpa robotics challenge finals," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2015, pp. 805–810.
[59] Y. Kojio, Y. Ishiguro, F. Sugai, Y. Kakiuchi, K. Okada, M. Inaba *et al.*, "Unified balance control for biped robots including modification of footsteps with angular momentum and falling detection based on capturability," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 497–504.
[60] S. Sato, Y. Kojio, Y. Kakiuchi, K. Kojima, K. Okada, and M. Inaba, "Robust humanoid walking system considering recognized terrain and robots' balance," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 8298–8305.