Task-Agnostic and Device-Agnostic Exoskeleton Control via Reinforcement Learning and Mixture of Experts

Zhimin Hou, Jinsen Huang, Ivan Lopez-Sanchez, and Hao Su†, *Senior Member, IEEE*

***Abstract*—Most exoskeleton control strategies have enabled them to enhance the wearer's locomotion in different motion tasks. However, two big challenges still hinder the application of exoskeletons in daily life settings: the lack of adaptation across different users and unseen devices. To address these challenges, we propose a sim-to-real reinforcement learning method to decouple the simulation and real-world physical control. Our learning-in-simulation framework can learn a device-agnostic reference prediction policy without using specific exoskeleton models. A variable impedance control was proposed to utilize the reference motion from the deployed reference prediction policy to achieve the adaptation across different users. The proposed method was validated by deploying the learned reference prediction policy on three different hip exoskeletons and three able-bodied subjects. The obtained assistive torque and power profiles demonstrate that the learned reference prediction policy can be directly deployed on unseen exoskeletons.**

***Index Terms*—Exoskeleton control, sim-to-real reinforcement learning, variable impedance control**

I. Introduction

Exoskeletons have shown promise in assisting able-bodied subjects to reduce effort while performing locomotion tasks and people with disabilities to improve mobility [1], [2], [3], [4]. While state-of-the-art exoskeletons have primarily been validated in laboratory settings for predefined tasks [5], their true potential lies in transitioning to real-world applications. Achieving this transition requires exoskeleton controllers capable of seamlessly adapting to multiple activities, diverse users, and various devices. Most exoskeleton control strategies were developed following a three-layer structure: high-level, mid-level, and low-level [3], [6]. The high-level control methods focused on detecting the ambulation modes or terrains to generate reference for the mid-level controller [7], [8]. Therefore, recent works focus on recognizing multiple ambulation modes from various sensors, such as cameras [9], IMUs [10], and EMGs. Paticularly, model-free data-driven methods were investigated for task recognition/classification [11], [8], [9] [12], [3], [13], [14]. The mid-level control layer aims to generate desired assistance according to the recognized ambulation mode and terrains [15]. Assistive torque profiles were designed as a function of estimated gait phase conditioned by the ambulation mode [9], [11]. Additionally, human-in-the-loop-optimization methods, such as Bayesian optimization, were investigated as mid-level layer control to provide personalized assistance for specific task [7]. The low-level control layer generally focuses on generating motor control commands to realize the desired assistance. Feedback controllers were typically investigated to compensate for the dynamics in actuation or robot system [16]–[18]. However, these exoskeleton control strategies depend on extensive human experiments or collect sufficient data from human users. It is challenge to achieve the generalization across tasks, users, and devices.

Sim-to-real technique has the potential to develop the exoskeleton controllers from physics-based musculoskeletal simulators [19], [20]. The human kinematic and biomechanical feedback of human musculoskeletal systems can reduce extensive human experiments or the efforts of collecting reference dataset from human body [21], [22]. However, in contrast to sim-to-real learning methods for legged robots [23] and humanoid robots [24], the exoskeleton needs to physically interact with the human musculoskeletal system. Their control is a non-linear, high-dimensional, and over-actuated control problem [25], [26], [27]. Imitation learning methods, generative adversarial imitation learning (GAIL) and adversarial motion priors (AMP), were employed to mimic the reference motion [28], [29], [30]. Fortunately, reinforcement learning (RL) based methods have demonstrated great success in reproducing human locomotion behaviors of the human musculoskeletal system [31], [32], [33]. Recently, the learned human locomotion behaviors were employed for exoskeleton control policy learning [34], [1]. However, two challenges continue to hinder the broader applications of sim-to-real exoskeleton control learning for physical exoskeletons.

The first challenge is lack of a sim-to-real learning method that learn the exoskeleton control policy across multiple ambulation modes and gait patterns. Most hierarchical control methods focus on ambulation or terrain recognition and gait phase estimation to design the necessary assistive profile [9], [11], [35]. However, these methods depend on large parameters tuning and the results are limited to the accuracy of estimation. Model-free learning-based methods were developed to leverage neural networks to directly generate the desired assistance

This work is supported in part by the National Science Foundation (NSF) Faculty Early Career Development Program (CAREER) award CMMI 1944655, National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR) Switzer Research Distinguished Fellow (SFGE22000372).

Z. Hou, J. Zhu, I. Lopez-Sanchez, Y. Yan, J. Huang, and H. Su are with Lab of Biomechatronics and Intelligent Robotics, Department of Mechanical and Aerospace Engineering, North Carolina State University, Raleigh, NC 27695, USA.

Hao Su is also with Joint NCSU/UNC Department of Biomedical Engineering, North Carolina State University, Raleigh, NC, 27695; University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.

†Corresponding author: hao.su796@ncsu.edu;

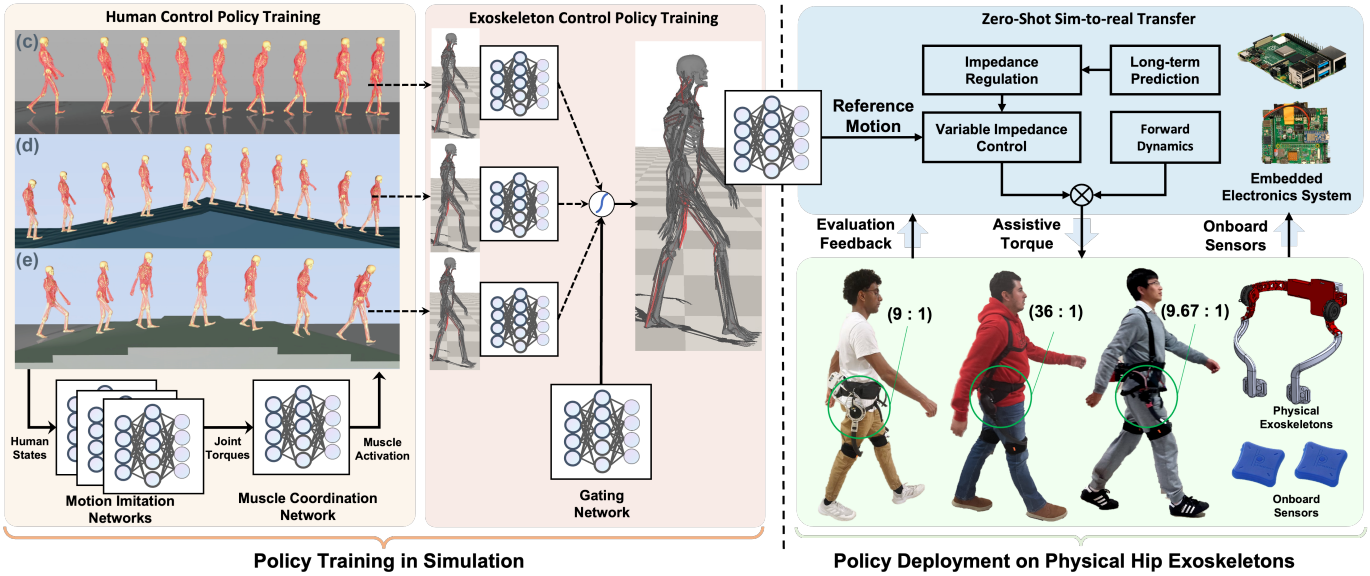

Fig. 1. Overall diagram of our device-agnostic sim-to-real RL for exoskeleton control. A reference prediction policy is trained by actor-critic networks using a full-body musculoskeletal human model in simulation. Additionally, the motion imitation network and muscle coordination network are trained to replicate the input reference activities for the human model. The offline policy training in simulation usually takes about 10 hours using a GPU. After the reference prediction policy has converged, the actor-network will be deployed on the customized electronic system to generate reference motion for variable impedance control . At each actuation step of policy deployment on different physical exoskeletons, the variable impedance control can adjust the assistive torque in real-time for each user.

TABLE I
COMPARISON WITH STATE-OF-THE-ART EXOSKELETON CONTROL METHODS

Methods	Activities/Tasks/ambulations	Adaptation	
		Users	Devices
[18]	–	–	√
[3], [35], [12]	√	–	–
[36], [17], [37]	–	√	–
[7], [38], [39], [37]	√	√	–
[1]			
Ours	√	√	√

profiles [40]. For instance, a temporal convolutional neural network (TCN) was formulated to estimate the biological hip joint torques and produce corresponding assistance profiles that align with human motion [3], [12]. However, these methods are constrained by their reliance on biologically inspired assistance profiles, which may not be suitable for maximizing user benefits in terms of reducing human muscle efforts. RL-based methods can optimize the assistance profile according to human states and reward feedback.

The second challenge is to obtain the user-adaptive and device-agnostic exoskeleton controller in addition to the generalization across multiple ambulation modes. Most sim-to-real robot learning methods focused on learning the control policy from simulated physical interaction [41]. Furthermore, random randomization typically was utilized to enhance the robustness adapting to robot or environment dynamics [42], [43]. However, these sim-to-real RL methods with domain randomization can only mitigate the sensitivity to minor variations in device parameters [1] [44]. Therefore, the inclusion of a specific exoskeleton model in simulation inherently limits the generalization of the learned controllers adapting to other robotic systems with different actuators or mechanisms. Furthermore, human biomechanics vary widely across individuals due to differences in anthropometry and gait patterns, making it challenging to reproduce human user individual behaviors. Therefore, the online tuning according to human kinematic and biomechanical feedback are still necessary to derive personalized assistive strategies [36]–[38]. Furthermore, adaptive controllers can also be explored as low-level control to compensate for robot unknown dynamics, resulting from the actuation and transmission systems [18], [7]. RL policy hierarchy and decoupling are practical ways to improve the sample efficiency [45], [46], [47].

To address the first challenge of learning an exoskeleton control policy for multiple ambulation modes, a two-stage training scheme was presented. In contrast to prior works [1] in Table **??**, we decoupled the human control policy learning and exoskeleton control policy learning. We employ a teacher-student scheme to reproduce the natural human-like gait and biomechanical feedback for multiple human locomotion behaviors learning. Afterwards, the exoskeleton control policy learning was formulated as a multi-task learning problem. Compared to existing methods in Table **??**, we present a mixture of experts method to learn a general assistive policy by interacting with several learned human control policies for different ambulation modes.

To address the second challenge of achieving device-agnostic and user-adaptive exoskeleton control, our scheme eliminates the dependence on device-specific models during simulation training. Unlike existing works in Table **??**, the policy learned

in simulation mainly focuses on obtaining the generalization across different ambulation modes, the ideal assistive torque is directly applied to the human joint. We decouple the exoskeleton control policy into a device-agnostic reference prediction policy and a lower-level interactive control to complete the physical interaction. A device-agnostic reference prediction policy is learned to generate the reference motion that can generalize across exoskeletons with varying actuators and transmission systems. Therefore, the learned device-agnostic reference prediction policy can be zero-shot transferred to physical exoskeletons. The variable impedance control was implemented by compensating to the actual dynamics of the exoskeleton. The impedance parameters were also updated according to the human individual feedback.

MoE framework can alleviate gradient conflicts by directing gradients to specialized experts, thus improving training efficiency and overall performance.

II. Problem Formulation

A. Problem Statement

The proposed sim-to-real exoskeleton control learning framework, illustrated in Fig.1, consists of two main phases: policy training in simulation and policy deployment on physical exoskeletons. In the simulation phase, learning an exoskeleton control policy relies on accurate modeling and control of the human musculoskeletal system across multiple locomotion modes, such as level ground, slopes, and stairs [48]. As shown in Fig.1, multiple human control policies are first trained using a velocity-commanded reinforcement learning (RL) [49] and curriculum learning scheme to reproduce each desired locomotion behaviors with adaptive walking speed. Afterwards, a universal exoskeleton control policy was developed by leveraging a mixture-of-experts approach, which integrates several exoskeleton control policies learned by interacting with each human musculoskeletal system actuated by locomotion specific human control policy. Furthermore, MoE could smooth the transition between two locomotion modes. Finally, the exoskeleton control policy with mixed parameters is then zero-shot transferred to a customized electronic system for its deployment on physical exoskeletons. To achieve the generalization across unseen exoskeletons and users, the dynamics of the physical exoskeleton are compensated for variable impedance controller.

1) Task Definition: This study focuses on three commonly used terrains $\chi \in \Xi$: level ground, slope, and stairs (see Fig. 1). For each locomotion, human need to walking forward along the direction $\mathbf{d}_t$ at a target speed $v \in [v_{min}, v_{max}]$. Therefore, the objective of exoskeleton control policy is to provide the desired assistance for the human model walking on specific terrain and desired walking speed.

2) Simulation Platform: The human musculoskeletal system and terrains are built based on DART [25], which is used for physics-based dynamic simulation. The full-body human musculoskeletal system consists of a rigid humanoid skeleton model coupled with computational muscle-tendon models. Most importantly, the exoskeleton model and interaction model are not required in simulation since the assistive force is ideally applied on the human joints. In this study, we take hip exoskeleton as an example.

B. Human Musculoskeletal System Modeling

1) Skeleton Modeling: The dynamics of the human musculoskeletal model are formulated as follows:

$$\boldsymbol{\mathcal{M}}^h(\boldsymbol{q}^h)\ddot{\boldsymbol{q}^h} + \boldsymbol{\mathcal{C}}^h(\boldsymbol{q}^h, \dot{\boldsymbol{q}^h})\dot{\boldsymbol{q}^h} = \boldsymbol{J}_m^T \boldsymbol{F}_m(\boldsymbol{a}) + \boldsymbol{J}_c^T \boldsymbol{F}_c + \boldsymbol{\tau}_{ext} \tag{1}$$

where $\boldsymbol{q}^h \in \mathbb{R}^{n_h}$ and $\dot{\boldsymbol{q}^h} \in \mathbb{R}^{n_h}$ are human joint angles and velocities, respectively. $\boldsymbol{\mathcal{M}}^h(\cdot)$ is the mass matrix, $\boldsymbol{\mathcal{C}}^h(\cdot,\cdot)$ is the Coriolis and gravitational forces. $\boldsymbol{F}_m \in \mathbb{R}^{n_h}$ and $\boldsymbol{F}_c \in \mathbb{R}^{n_h}$ are the muscle and constraint forces. $\boldsymbol{J}_m$ and $\boldsymbol{J}_c$ are Jacobian matrices, which map muscle forces and constraint forces to the human joint space. $\boldsymbol{a} \in \mathbb{R}^{n_m} \in [0,1]^{n_m}$ is muscle activation. $\boldsymbol{\tau}_{ext} \in \mathbb{R}^{n_h}$ is the external torque, including the applied force by the exoskeleton and the interaction force with the environment.

2) Muscle Modeling: According to the Hill-type model [25], the $i-$th muscle-tendon unit is formulated to derive the muscle force $F_m(a_i; l_i, v_i)$ from activation as $F_{max} \cdot [a_i \cdot f_l(l_i) \cdot f_v(v_i) + F_p(l_i)]$. l_i and v_i represent the normalized muscle length and the rate of muscle changes, respectively. $F_p(l_i)$ is the passive force developed by the muscle. F_{max} is maximum isometric muscle force. Therefore, the muscle force can be derived as follows:

$$\boldsymbol{F}_m(\boldsymbol{a}) = \boldsymbol{F}_m^{max} \cdot [\boldsymbol{a} \cdot \boldsymbol{F}_l + \boldsymbol{F}_m^p] = \boldsymbol{A} \cdot \boldsymbol{a} + \boldsymbol{b} \tag{2}$$

where $\boldsymbol{F}_l$ is the function affected by muscle length and the rates of muscle changes.

III. Device-agnostic Sim-to-real Exoskeleton Control Policy Learning

Given the human musculoskeletal model and reference motion dataset, the sim-to-real task-and-device agnostic exoskeleton control policy learning is learned by two phases (see Fig. 1). The human musculoskeletal system is actuated by H human control policy to replicate each ambulation modes one by one (Section III-A). Afterwards, the exoskeleton control policy learning is formulated as a multi-task RL problem, a mixture of expert scheme is proposed to learn a general policy by interacting with each human control policy (Section III-B).

A. Human Control Policy Training via Goal-conditioned RL

A parameterized human control policy $\boldsymbol{\pi}_h(\boldsymbol{a}|\boldsymbol{s}^h; \boldsymbol{\phi}_h, \boldsymbol{\chi})$ is learned to output muscle activation for replication of the reference motion under specific locomotion mode $\boldsymbol{\chi} \in \boldsymbol{\Xi}$. Inspired by the nature muscle synergies controlled by central nervous system [25], [27], [50], [51], we decoupled each human control policy as two sub policies: human motion imitation policy $\boldsymbol{\pi}_h(\boldsymbol{u}^h|\boldsymbol{s}^h; \boldsymbol{\phi}_h^s, \boldsymbol{\chi})$ and human muscle coordination policy $\boldsymbol{\pi}_m(\boldsymbol{a}|\boldsymbol{u}^h, \boldsymbol{s}_m^h; \boldsymbol{\phi}_h^m)$. Here, $\boldsymbol{u}^h$ is the actuation torque for desired motion imitation. For sample-efficient learning [25], [29], [49], the human motion imitation policy is decoupled into a reference motion prediction policy $\boldsymbol{\pi}_r(\boldsymbol{u}_r^h|\boldsymbol{s}^h; \boldsymbol{\phi}_r, \boldsymbol{\chi})$ and a joint-level PD controller $\boldsymbol{\pi}_{PD}(\boldsymbol{u}^h|\boldsymbol{u}_r^h, \boldsymbol{s}^h; \boldsymbol{\phi}_{PD})$. Therefore, the human motion imitation policy learning is degraded to

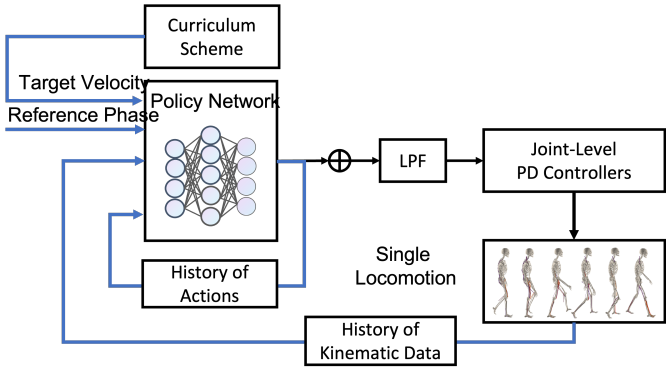

Fig. 2. Human imitation control policy framework.

learn a human reference motion prediction policy that outputs the reference position $\boldsymbol{u}_r^h$ for joint-level PD controller (see Fig. 2). While the human muscle coordination policy and the joint-level PD controller is task independent. The human muscle coordination policy can be trained via a regression-based method [1], [25].

We formulate the human musculoskeletal system control problem as a Markov Decision Process (MDP) $< \mathcal{S}_h, \mathcal{U}_h, \mathcal{R}_h, \mathcal{T}_h, \mathcal{P}_h^0, \gamma_h >$. $\boldsymbol{s}^h \in \mathcal{S}_h$ is the state of the human musculoskeletal system. $\boldsymbol{u}^h \in \mathcal{U}_h$ is the action taken by the human motion imitation policy. $\mathcal{R}_h : \mathcal{S}_h \times \mathcal{U}_h \rightarrow r^h(\boldsymbol{s}^h, \boldsymbol{u}^h) \in \mathbb{R}$ is the reward function. $\mathcal{P}_h : \mathcal{S}_h \times \mathcal{U}_h \rightarrow \mathcal{S}_h$ is the environment transition function. p_h^0 is an initial human state distribution and $\gamma_h \in (0, 1)$ is the discount factor. The parameter of the human reference motion prediction policy is updated by maximizing the expected cumulative reward, as follows:

$$\boldsymbol{\mathcal{J}}(\boldsymbol{\phi}_r) = \mathbb{E}_{\mathcal{T} \sim p_{\boldsymbol{\phi}_r}(\mathcal{T})} \left[\sum\nolimits_{t=0}^{\infty} \gamma_h^t r_t(\boldsymbol{s}_t^h, \boldsymbol{u}_t^h) \right] \tag{3}$$

where $\mathcal{T}$ is the sampled human motion trajectory during one episode. $\boldsymbol{s}_0^h \sim \mathcal{P}_h^0$, and action $\boldsymbol{u}_t^h \sim \boldsymbol{\pi}_r(\boldsymbol{u}_t^h | \boldsymbol{s}_t^h; \boldsymbol{\phi}_r)$ is sampled at each control timestep $\boldsymbol{t}$ and the muscle activation $\boldsymbol{a}$ is derived from $\boldsymbol{\pi}_m(\boldsymbol{a} | \boldsymbol{u}^h, \boldsymbol{s}^h; \boldsymbol{\phi}_m)$ to drive the human musculoskeletal system. The next state is sampled by $\boldsymbol{s}_{t+1}^h \sim \mathcal{P}_h(\boldsymbol{u}_t^h | \boldsymbol{s}_t^h; \boldsymbol{\phi}_h)$ and the reward r_t is calculated.

Similar to the velocity-goal-conditioned locomotion control [30], the state space [52], [53], [54] for reference motion prediction policy learning is defined as follows

$$\boldsymbol{s}_t^h = [\mathbf{O}_t^h, \mathbf{U}_t^h, \varsigma, \boldsymbol{c}_t] \tag{4}$$

where $\mathbf{O}_t^h = [\boldsymbol{o}_t^h, \cdots, \boldsymbol{o}_{t-T}^h]$ represents the history of the past human states and $\mathbf{U}_t^h = [\boldsymbol{u}_t^h, \cdots, \boldsymbol{u}_{t-T}^h]$ represents the history of the past output reference motions. A phase variable $\varsigma \in [0, 1]$ is employed to align with the reference motion. $\boldsymbol{c}_t$ is the velocity command.

Natural walking with changing gait by designing the following reward function consisting of three subrewards as follows:

$$\begin{aligned} r_t^h &= w^I r_t^I + w^T r_t^T + w^G r_t^G + w^C r_t^C \\ r_t^I &= w^p r_t^p + w^v r_t^v \\ r_t^G &= \exp[-\sigma_G || \dot{\mathbf{x}}_t^{com} - p_t^e ||^2] \\ r_t^C &= w^m r_t^m + w^s r_t^s \end{aligned} \tag{5}$$

where r_t^I is the sub-reward used to encourage motion imitation performance. r_t^T is a terrain-related reward to encourage the walking on selected terrain. r_t^G is the sub-reward used to encourage to fulfill the tasks-specific objective, here is to track desired walking speed. $\dot{\mathbf{x}}_t^{com}$ is the actual center-of-mass velocity of human musculoskeletal model at the time step $\boldsymbol{t}$. r_t^C is the sub reward to fulfill the external requirements, such as energy reduction r_t^m and smooth motion r_t^s. $\{w^I, w^G, w^C, w^p, w^v, w^m, w^s, w^e\}$ are the weights of each sub-reward, which should be pretested for each task.

In order to learn each human motion imitation policy $\boldsymbol{\pi}(\boldsymbol{u}_r^h | \boldsymbol{s}^h; \boldsymbol{\psi}_r)$ for specific locomotion mode using Proximal Policy Optimisation (PPO) [55], chosen for its efficiency and suitability for parallel computation. Curriculum learning [49], [41] is employed to train the policy for various velocity tracking. Therefore, we adopt the following update rule to automatically adjust the curriculum during each locomotion mode training.

$$f_t^v = \tag{6}$$

where k is episode index.

Termination criteria:

As illustrated in Fig. 2, multilayer perception (MLP)-based encoder is employed for the actor and critic networks. The implementation details of using PPO to train human control policy is concluded in Algorithm 1.

Algorithm 1 Human Control Policy Training Using PPO

Initialization: parameters of actor network and critic network $\boldsymbol{\phi}_{\boldsymbol{a}}^0$
for Each Iteration $i \in 1, 2, \cdots, I$ **do**
  for Each agent $j \in 1, 2, \cdots N$ **do**
    Collect trajectories $\{\boldsymbol{\mathcal{T}_j}\}$ from current reference prediction policy $\boldsymbol{\pi}(\cdot|\cdot; \boldsymbol{\phi}_a^i)$ for T timesteps
    Compute estimated advantages $\boldsymbol{\Omega}_j = \{\widehat{\boldsymbol{A}}_t\}_{t=1}^T$
    $\mathcal{D}_i \leftarrow \mathcal{D}_i + \{\boldsymbol{\mathcal{T}_j}, \boldsymbol{\Omega}_j\}$
  end for
  Update parameters $\boldsymbol{\phi}_{\boldsymbol{a}}^*$ for L epochs using minibatch data with size M sampled from $\mathcal{D}_i$
  $\boldsymbol{\phi}_a^{i+1} \leftarrow \boldsymbol{\phi}_a^*$
end for

B. Reference Prediction Policy Training via Mixture of Experts

The objective of learning exoskeleton control policy is to provide necessary assistance for multiple locomotion modes ($\boldsymbol{\chi} \in \boldsymbol{\Xi}$) defined in Section II-A, which is formulated as a multi-task RL problem [56], [57], [58]. A parameterized exoskeleton policy $\boldsymbol{\pi}(\boldsymbol{u}^e | \boldsymbol{s}^e; \boldsymbol{\psi}_e)$ is learned from the interactive behaviours by ideally applying assistive force on the desired human joints, such as the hip joints in Fig. 1. $\boldsymbol{u}^e$ is the motor control commands for the exoskeleton. $\boldsymbol{s}^e$ is the full state of the exoskeleton agent, which should include the human states and the robot states. However, the exoskeleton control policy will be deployed on physical exoskeletons (see human-robot interaction system depicted in Fig. 1). For each locomotion mode, only partial observation of human state $\boldsymbol{s}_t^e$ can be measured from on-board sensors, such as IMUs. Each locomotion mode is formulated as a MDP $< \mathcal{S}_e, \mathcal{U}_e, \mathcal{R}_i, \mathcal{P}_i, p_e^0, \gamma_e >$. All tasks are sharing the state space and action space, distinguished by

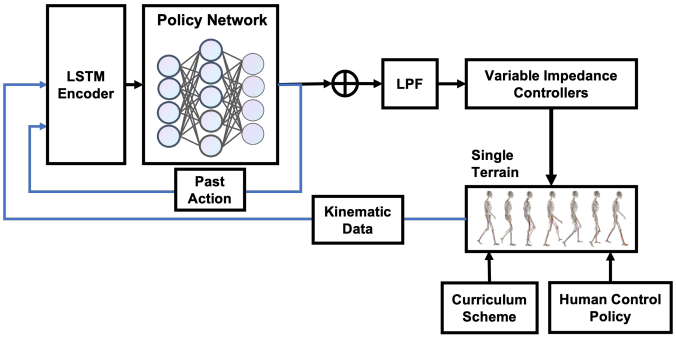

Fig. 3. Exoskeleton control policy for single locomotion mode.

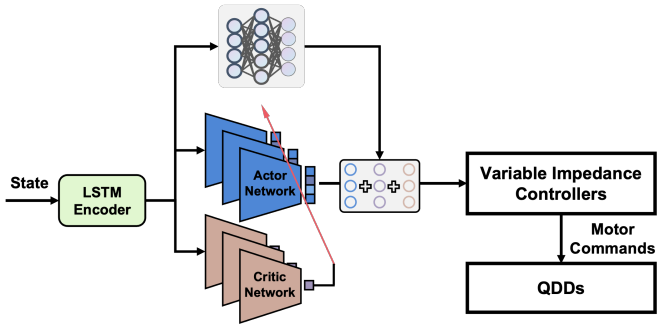

Fig. 4. Overview of exoskeleton control policy learning based on MoE framework.

different transition function $\mathcal{P}_i$ and reward function $\mathcal{R}_i$ [56]. The object is to maximize the average expected cumulative reward across all tasks, which are uniformly sampled during training. The objective function is defined as follows:

$$\mathcal{J}(\psi_e) = \mathbb{E}\left[\sum_{\mathbf{h}\in\mathcal{H}}\sum_{t=0}^{\infty}\gamma_e^t r_e^i(\boldsymbol{s}_t^e, \boldsymbol{u}_t^e, \boldsymbol{s}_{t+1}^e)\right] \tag{7}$$

where $r_e^i(\boldsymbol{s}_t^e, \boldsymbol{u}_t^e, \boldsymbol{s}_{t+1}^e)$ represents the received reward of $i-th$ locomotion mode.

We decouple the exoskeleton control policy into learning a device-agonistic reference prediction policy in simulation and a variable impedance controller that can be adapted to unseen physical exoskeletons and users (see Fig. 1), as follows:

$$\boldsymbol{\pi}_e(\boldsymbol{u}^e|\boldsymbol{s}^e;\boldsymbol{\psi}_e) = \boldsymbol{\pi}_r(\boldsymbol{u}_r^e|\boldsymbol{s}^e;\boldsymbol{\psi}_r) \circ \boldsymbol{\pi}_{VIC}(\boldsymbol{u}^e|\boldsymbol{s}^e, \boldsymbol{u}_r^e;\boldsymbol{\psi}_u) \tag{8}$$

where $\boldsymbol{\pi}_r(\boldsymbol{u}_r^e|\boldsymbol{s}^e;\boldsymbol{\psi}_r)$ is the reference prediction policy learned from the simulation to obtain the reference motion intention $\boldsymbol{u}_r^e$ for the variable impedance control (see Fig. 3) [59]. Given the device-agnostic reference motion $\boldsymbol{u}_r^e$, we implement the variable impedance control $\boldsymbol{\pi}(\boldsymbol{u}^e|\boldsymbol{s}^e, \boldsymbol{u}_r^e;\boldsymbol{\psi}_u)$ to obtain the motor control commands $\boldsymbol{u}^e$ for desired assistance. Given impedance controller with the fixed parameter $\bar{\boldsymbol{\psi}}_u$, the exoskeleton control policy $\boldsymbol{\pi}_e(\boldsymbol{u}^e|\boldsymbol{s}^e;\boldsymbol{\psi}_e)$ will be degraded to optimize the reference prediction policy $\boldsymbol{\pi}_r(\boldsymbol{u}_r^e|\boldsymbol{s}^e;\boldsymbol{\psi}_r)$.

The policy parameter $\boldsymbol{\psi}_r$ will be updated by maximizing the objective function by mixture of experts (MoE) [60], [61]. Therefore, H experts are integrated using a routing function

Algorithm 2 Reference Prediction Policy Training Using PPO

Initialization: parameters of actor network and critic network $\boldsymbol{\phi}_a^0$
for Each Iteration $i \in 1, 2, \cdots, I$ **do**
  for Each agent $j \in 1, 2, \cdots N$ **do**
    Collect trajectories $\{\boldsymbol{\mathcal{T}_j}\}$ from current reference prediction policy $\boldsymbol{\pi}(\cdot|\cdot;\boldsymbol{\phi}_a^i)$ for T timesteps
    Compute estimated advantages $\boldsymbol{\Omega}_j = \{\widehat{\boldsymbol{A}}_t\}_{t=1}^T$
    $\mathcal{D}_i \leftarrow \mathcal{D}_i + \{\boldsymbol{\mathcal{T}_j}, \boldsymbol{\Omega}_j\}$
  end for
  Update parameters $\boldsymbol{\phi}_a^*$ for L epochs using minibatch data with size M sampled from $\mathcal{D}_i$
  $\boldsymbol{\phi}_a^{i+1} \leftarrow \boldsymbol{\phi}_a^*$
end for

to learn the reference prediction policy as follows:

$$\begin{aligned}\boldsymbol{\pi}_r(\boldsymbol{u}_r^e|\boldsymbol{s}^e;\boldsymbol{\psi}_r) &= \sum_{i=0}^{H}\boldsymbol{\pi}(g_i|\boldsymbol{s}^e;\boldsymbol{\theta})\boldsymbol{\pi}_r(\boldsymbol{u}_r^e|\boldsymbol{s}^e;\boldsymbol{\psi}_r^i)\\ s.t.&\sum_{i=0}^{H}\boldsymbol{\pi}(g_i|\boldsymbol{s}^e;\boldsymbol{\theta}) = 1\end{aligned} \tag{9}$$

where $\boldsymbol{\pi}_r(\boldsymbol{u}_r^e|\boldsymbol{s}^e;\boldsymbol{\psi}_r^i)$ is the $i-th$ reference prediction policy. $\boldsymbol{\pi}(g_i|\boldsymbol{s}^e;\boldsymbol{\theta})$ is a gating policy to output the weight of each expert reference prediction policy. The goal of the gating policy is to maximize the conditional expectation of the return, as follow:

$$\boldsymbol{Q}_G^{\boldsymbol{\pi}}(\boldsymbol{s},) = \mathbb{E}\left[\boldsymbol{\pi}_r(\boldsymbol{u}_r^e|\boldsymbol{s}^e;\boldsymbol{\psi}_r^i)\boldsymbol{Q}^{\pi}(\boldsymbol{s}^e, \boldsymbol{u}_r^e)\right] \tag{10}$$

where $\boldsymbol{Q}^{\pi}(\boldsymbol{s}^e, \boldsymbol{u}_r^e)$ is the action-value function. When we adopt the deterministic action $\boldsymbol{u}_r^e$, $\boldsymbol{Q}_G^{\boldsymbol{\pi}}(\boldsymbol{s},) = \boldsymbol{Q}^{\pi}(\boldsymbol{s}^e, \boldsymbol{u}_r^e)$.

Therefore, we adopt a softmax gating policy, and the parameter $\boldsymbol{\theta}$ can be optimized to maximize the mutual information.

$$\boldsymbol{\pi}(g_i|\boldsymbol{s}^e) = \frac{\exp(\boldsymbol{Q}^{\pi}(\boldsymbol{s}^e, \boldsymbol{\pi}_r(\boldsymbol{u}_r^e|\boldsymbol{s}^e;\boldsymbol{\psi}_r^i)))}{\sum_{i=0}^{G}\exp(\boldsymbol{Q}^{\pi}(\boldsymbol{s}^e, \boldsymbol{\pi}_r(\boldsymbol{u}_r^e|\boldsymbol{s}^e;\boldsymbol{\psi}_r^i)))} \tag{11}$$

The observation space for training reference prediction policy is defined as follows:

$$\boldsymbol{s}_t^e = [\boldsymbol{q}_t^h, \dot{\boldsymbol{q}}_t^h] \tag{12}$$

where $\boldsymbol{q}_t^h$ and $\dot{\boldsymbol{q}}_t^h$ are the actual angle and velocity of both hip joints at each control step $\boldsymbol{t}$, which are measured from onboard sensors.

The reward of each observation-action pair comprises several subreward as follows:

$$r_e^i(\boldsymbol{s}_t^e, \boldsymbol{u}_t^e, \boldsymbol{s}_{t+1}^e) = w^h r_h^i + w^m r_m + w^c r_c \tag{13}$$

where r_h^i is the sub-reward that used to encourage the human performance under $i-th$ locomotion mode, defined in (5). r_m is the sub-reward to encourage the reduction of muscle work. r_c is the sub-reward to encourage the smoothness of the generated assistive torque. $\{w^h, w^m, w^c\}$ are the weights of each sub-reward, which should be pretested. Furthermore, only the muscle activation $\boldsymbol{m}_l \in \mathbb{R}^3 0$ of potentially assisted

muscles was employed to evaluate the reference prediction policy , as follows:

$$r_m = \exp\left(-\sigma_m ||\boldsymbol{m}_l||^2\right) \tag{14}$$

where σ_m is a sensitive factor.

To mitigate the dependence of the historic information, a LSTM encoder [44] is formulated to process the input partial human observation. The implementation of training reference prediction policy based on soft actor-critic (SAC) [62] is concluded in Algorithm 2. Furthermore, the termination criteria of training reference prediction policy depends on the termination criteria of training human control policy.

IV. Real-time Variable Impedance Control for Policy Development

The reference prediction policy learned can be deployed on the physical exoskeletons for same target joints (see Fig. 1), such as hip exoskeletons. The actor network with optimized parameters ψ_r^* of the reference prediction policy can be zero-shot transferred and run on the customized electronic system (see Fig. 1). The objective of variable impedance control is to tune the impedance parameters for different human users and exoskeletons. The dynamics of exoskeleton and physical interactions with human limb are formulated (Section IV-B and IV-C), which are estimated and compensated in variable impedance control . An intention-driven impedance modulation scheme is designed to generate the adaptive assistive torque for each individual based on the dynamics compensation (Section IV-D).

perform zero-shot sim-to-real transfer.

A. Reference Prediction Policy Deployment on Physical Exoskeletons

Without loss of generality, the hip exoskeleton is taken as an example. The input to the actor network is the observation of human state $\boldsymbol{o}_t^e$, including the human joint position $\boldsymbol{q}_t$ and velocity $\dot{\boldsymbol{q}}_t$, that can be measured by IMUs. The sensing and control frequency f_p of the hardware platform may be limited and different from it in the simulation f_s. When the reference prediction policy has convergent, the episode return has achieved become stable. Therefore, during the policy deployment, at each actuation step, the actor network can predict reference position $\boldsymbol{q}_r$ and velocity $\dot{\boldsymbol{q}}_r$. The variable impedance control is implemented to achieve adaptation on different exoskeletons and users by compensating for the dynamics of devices and human limbs. While the deployed physical exoskeleton may have different motors, gears, and mechanical structures as depicted in Fig. 1, which were not considered in the simulation training. The dynamics of physical exoskeletons and human limbs are estimated and compensated. Sim-to-real [63],

B. Dynamic Model of Physical Exoskeleton

The dynamics of exoskeletons can be modeled by using the Euler-Lagrange method, expressed as:

$$\boldsymbol{\mathcal{M}}^e(\boldsymbol{q}^e)\ddot{\boldsymbol{q}}^e + \boldsymbol{\mathcal{C}}^e(\boldsymbol{q}^e, \dot{\boldsymbol{q}}^e)\dot{\boldsymbol{q}}^e + \boldsymbol{\mathcal{G}}^e(\boldsymbol{q}^e) = \boldsymbol{u}^e + \boldsymbol{\tau}_h \tag{15}$$

where $\boldsymbol{q}^e, \dot{\boldsymbol{q}}^e, \ddot{\boldsymbol{q}}^e \in \mathbb{R}^{n\times 1}$ are the generalized coordinates of the exoskeletons. $\boldsymbol{\mathcal{M}}^e(\boldsymbol{q}^e) \in \mathbb{R}^{n\times n}$ is the inertia matrix; $\boldsymbol{\mathcal{C}}^e(\boldsymbol{q}^e, \dot{\boldsymbol{q}}^e) \in \mathbb{R}^{n\times n}$ is the Coriolis and centrifugal matrix; $\boldsymbol{\mathcal{G}}^e(\boldsymbol{q}^e) \in \mathbb{R}^{5\times 1}$ is the gravity vector. $\boldsymbol{u}^e \in \mathbb{R}^{n\times 1}$ is the vector of motor control commands composed of the applied force on the robot joints. $\boldsymbol{\tau}_h \in \mathbb{R}^{n\times 1}$ is the human exerted force applied on the robot joints.

C. Dynamic Model of Human Limbs

The dynamic model of individual human limbs needs to be considered in human motion intention estimation during physical interaction. By referring to [17], [64], the dynamics of human limb is described as

$$-\boldsymbol{M}_h\ddot{\boldsymbol{q}}_h - \boldsymbol{D}_h(\dot{\boldsymbol{q}}_h - \dot{\boldsymbol{q}}_h^\star) - \boldsymbol{K}_h(\boldsymbol{q}_h - \boldsymbol{q}_h^\star) = \boldsymbol{\tau}_h \tag{16}$$

where $\boldsymbol{M}_h \in \mathbb{R}^{n\times n}$, $\boldsymbol{D}_h \in \mathbb{R}^{n\times n}$, and $\boldsymbol{K}_h \in \mathbb{R}^{n\times n}$ are the inertia, damping, and spring matrices of the human limb, respectively, which varies between human subjects and hard to measure. $\boldsymbol{q}_h \in \mathbb{R}^n$, $\dot{\boldsymbol{q}}_h \in \mathbb{R}^n$, and $\ddot{\boldsymbol{q}}_h \in \mathbb{R}^n$ are the actual human joint position, velocity, and acceleration, respectively. $\{\boldsymbol{q}_h^\star \in \mathbb{R}^n, \dot{\boldsymbol{q}}_h^\star \in \mathbb{R}^n, \ddot{\boldsymbol{q}}_h^\star \in \mathbb{R}^n\}$ are the desired human joint position, human joint velocity, and human joint acceleration, respectively.

Most human movements are relatively slow resulting in the damper and spring terms are dominant. According to [17], the desired position of human limb representing the human motion intention can be obtained from (16), as

$$\boldsymbol{q}_h^\star = \boldsymbol{q}_h - \boldsymbol{K}_h^{-1}[\boldsymbol{\tau}_h + \boldsymbol{D}_h(\dot{\boldsymbol{q}}_h - \dot{\boldsymbol{q}}_h^\star)] \tag{17}$$

Due to the intrinsic uncertainties in human motion intention and sensory noise, the joint angles of the human limb are subject to random variations. The autoregression model has demonstrated that it is helpful to model limb positions as random variables [17]. The future position of human limb can be predicted as the long-term effect of human motion intention. The human motion intention is estimated using the Gaussian mixture model (GMM), as follows

$$\mathcal{P}(\boldsymbol{q}_h^\star) = \sum_{c=1}^{C} \varpi_c \mathcal{N}(\boldsymbol{q}_h | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \tag{18}$$

where the number of Gaussian components C and its parameters $\{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c, \varpi_c\}$ need to be optimized using online EM algorithm [65]. At $t{+}1$ actuation time step, given the estimated parameters and current position $\boldsymbol{q}_h(\boldsymbol{t})$, the input vector is defined as

$$\begin{aligned} \boldsymbol{q}_h^L &= [\boldsymbol{q}_h(\boldsymbol{t}-L), \boldsymbol{q}_h(\boldsymbol{t}-L+1), \cdots, \boldsymbol{q}_h(\boldsymbol{t})] \\ \boldsymbol{q}_h^\star(\boldsymbol{t}+1) &= \mathbb{E}[\boldsymbol{q}_h^\star(\boldsymbol{t}+1) | \boldsymbol{q}_h^L] \end{aligned} \tag{19}$$

D. Impedance Adjustment Scheme

The reference prediction policy learned from simulation predicted a short-term reference motion for the target joints based on the user's history kinematics. Simultaneously, a long-term prediction of the human joint motion is obtained using a Gaussian Mixture Model/Regression (GMM/GMR) trained to capture variations in human motion intention. An online adaptation scheme utilizes the difference between the predicted

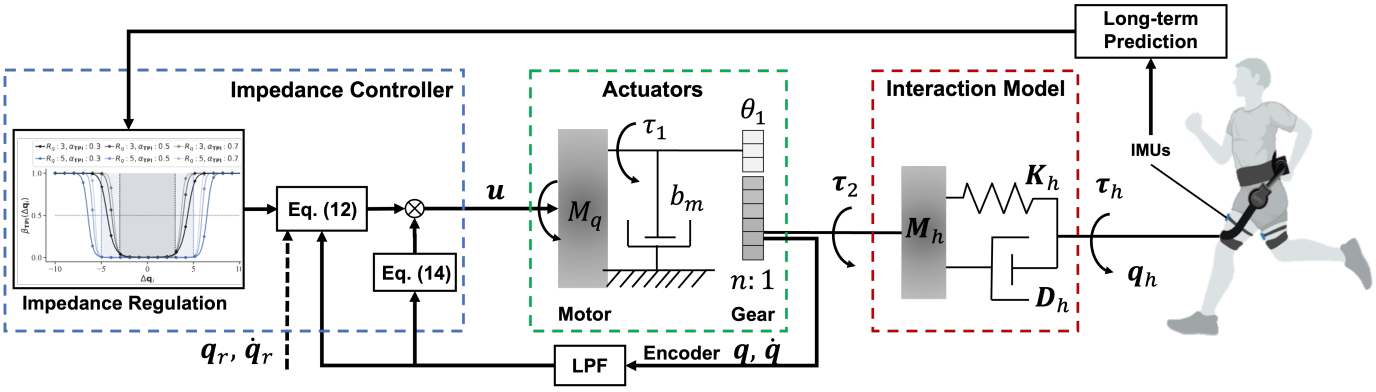

Fig. 5. Weight factors for adapting the impedance along the position difference under different threshold R_q and sensitive factor $\alpha_{\mathbf{TPI}}$. The threshold and sensitive factor should be pretested for each human user.

Algorithm 3 Reference Prediction Policy Deployment on Physical Exoskeletons

Initialization: optimized policy parameter $\boldsymbol{\phi_a^*}$
Initialization: GMMs parameters $\{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c, \varpi_c\}_{c=1}^{C}$
Initialization: pretested impedance parameter $\boldsymbol{\phi_u^0}$
for Each actuation step $\boldsymbol{t}$ **do**
  Observe state $\boldsymbol{o}_t$
  Predict reference motion $\boldsymbol{q}_r \sim \boldsymbol{\pi}(\cdot|\boldsymbol{o}_t; \boldsymbol{\phi_a^*})$
  Predict long-term motion intention $\boldsymbol{q}_h^\star$ based on (19)
  Update impedance parameter $\boldsymbol{\phi_u^k} \leftarrow \boldsymbol{\phi_u^0}$
  Obtain motor control command $\boldsymbol{u}_t$ from (21) to (25)
end for

short-term and long-term reference to tune the stiffness and damping parameters for adjusting the assistive torque profile according to each user's biomechanics actual feedback.

The objective of impedance control is to maintain a desired impedance model, as

$$\boldsymbol{M}_d(\ddot{\boldsymbol{q}} - \ddot{\boldsymbol{q}}_r) + \boldsymbol{C}_d(\dot{\boldsymbol{q}} - \dot{\boldsymbol{q}}_r) + \boldsymbol{K}_d(\boldsymbol{q} - \boldsymbol{q}_r) = \boldsymbol{\tau}_h \tag{20}$$

where $\boldsymbol{M}_d \in \mathbb{R}^{n\times n}$, $\boldsymbol{C}_d \in \mathbb{R}^{n\times n}$, and $\boldsymbol{K}_d \in \mathbb{R}^{n\times n}$ are the desired inertia, damping, and stiffness matrices. $\boldsymbol{\tau}_h$ is the human exerted force. As shown in Fig. 1, the reference motion $\{\boldsymbol{q}_r, \dot{\boldsymbol{q}}_r\}$ are derived from the reference prediction policy $\boldsymbol{\pi}(\boldsymbol{a}|\boldsymbol{s}; \boldsymbol{\phi}_a)$, which are considered as the short-term motion intention. $\boldsymbol{\psi}_u = \{\boldsymbol{M}_d, \boldsymbol{C}_d, \boldsymbol{K}_d\}$ are the impedance parameters that should be adjusted for each deployed exoskeleton and user. Firstly, the maximal impedance parameter $\boldsymbol{\psi}_u^{\mathbf{max}}$ for specific device is determined manually based on the property of actuators. Secondly, $\boldsymbol{\psi}_u^t$ is adapted impedance at $\boldsymbol{t}$ actuation step according to the difference between the actual position and the long-term predicted motion intention using (18).

Variable impedance control is employed to derive the motor control commands to provide the desired assistive torque, as

$$\boldsymbol{u} = \boldsymbol{u}_{fd} + \boldsymbol{u}_{ff} \tag{21}$$

where $\boldsymbol{u}_{fd}$ is the feedback control term to adapt the assistance for each individual and $\boldsymbol{u}_{ff}$ is the feedforward control term to compensate for the dynamics of the exoskeletons and human limbs.

When the desired inertia equals the robot's inertia, the human exerted force $\boldsymbol{\tau}_h$ is not required to be measured [17]. The variable impedance control term $\boldsymbol{u}_{fd}$ is designed as

$$\begin{aligned} \boldsymbol{u}_{fd} &= -\boldsymbol{C}_d^{\boldsymbol{t}}(\dot{\boldsymbol{q}} - \dot{\boldsymbol{q}}_r) - \boldsymbol{K}_d^{\boldsymbol{t}}(\boldsymbol{q} - \boldsymbol{q}_r) \\ \boldsymbol{C}_d^{\boldsymbol{t}} &= \boldsymbol{\omega}(\boldsymbol{q}, \boldsymbol{q}_h^\star) \circ \boldsymbol{C}_d^{\mathbf{min}} + \boldsymbol{C}_d^{\mathbf{max}} \\ \boldsymbol{K}_d^{\boldsymbol{t}} &= \boldsymbol{\omega}(\boldsymbol{q}, \boldsymbol{q}_h^\star) \circ \boldsymbol{K}_d^{\mathbf{min}} + \boldsymbol{K}_d^{\mathbf{max}} \end{aligned} \tag{22}$$

where $\boldsymbol{\omega}(\boldsymbol{q}, \boldsymbol{q}_h^\star) \in [0, 1]$ is a diagonal weighting factor matrix. For exoskeleton, the corresponding weight factor is defined based on the difference between the actual joint position $\boldsymbol{q}$ and the long-term estimated joint position $\boldsymbol{q}_h^\star$. $[\boldsymbol{K}_d^{\mathbf{min}}, \boldsymbol{K}_d^{\mathbf{max}}]$ and $[\boldsymbol{C}_d^{\mathbf{min}}, \boldsymbol{C}_d^{\mathbf{max}}]$ are the stiffness range and damping range for each user, which are defined according to $\boldsymbol{\psi}_u^{\mathbf{max}}$ for specific device. The weight factor of i-th ($i \in \{1, \cdots, n\}$) dimension of robot joint can be derived as

$$\begin{aligned} \beta_{\mathbf{IMP}}^i &= \frac{1}{1 + \exp(\alpha_{\mathbf{TPI}}\varepsilon(\Delta\boldsymbol{q}_i) - R_{\boldsymbol{q}})} \\ \varepsilon(\Delta\boldsymbol{q}_i) &= ||\Delta\boldsymbol{q}_i||^2 - R_{\boldsymbol{q}}^2 \end{aligned} \tag{23}$$

where $\Delta\boldsymbol{q}_i = \boldsymbol{q}_i - \boldsymbol{q}_{hi}^\star$ is the the tracking error of $i-$th dimension and $\alpha_{\mathbf{TPI}} \in (0, 1)$ is a sensitive factor. R_q is the predefined threshold. As illustrated in Fig. 5, .

A linear function regression is employed to estimate the dynamics and the feedforward term is designed as

$$\boldsymbol{u}_{ff} = \boldsymbol{Y}_q(\boldsymbol{q}, \dot{\boldsymbol{q}}, \dot{\boldsymbol{q}}_z, \ddot{\boldsymbol{q}}_z)\boldsymbol{\psi}_q \tag{24}$$

where $\boldsymbol{Y}_q$ is linear regression matrices and $\boldsymbol{\psi}_q$ is estimated according to the impedance error as follows

$$\boldsymbol{\psi}_q = \widehat{\boldsymbol{\psi}}_q - \boldsymbol{L}_q\boldsymbol{Y}_q^T(\cdot, \cdot, \cdot, \cdot)\Delta\boldsymbol{q} \tag{25}$$

V. Experiments

A. Experimental Setups

1) Simulation Setups for Reference Prediction Policy Training: The human musculoskeletal model is built based on MASS and the dynamics [] is simulated by DART []. The simulation is run on the PC with one GPU (GEFORCE RTX 4080). PPO and SAC were employed to train the reference prediction policy .

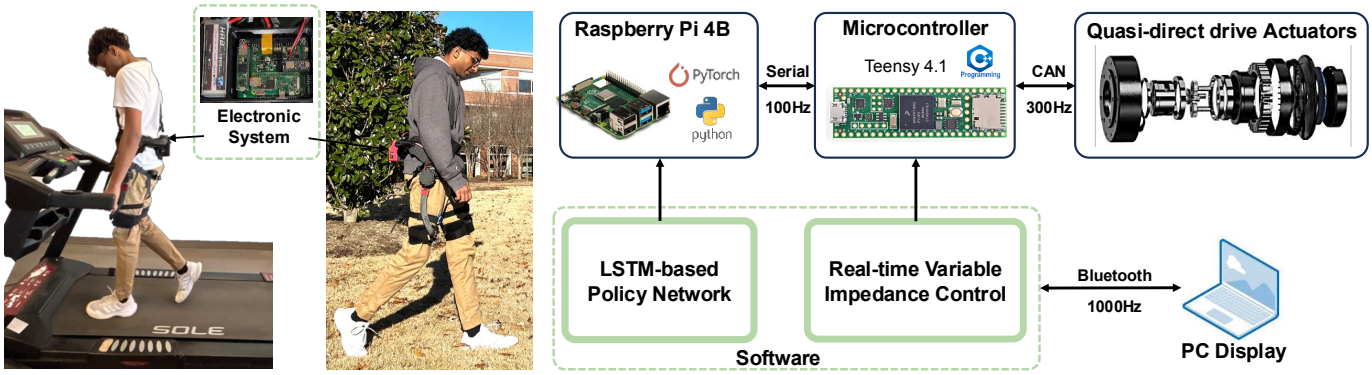

Fig. 6. Experimental setups for policy deployment. (a) Experiments demonstrating policy deployment on hip exoskeletons for treadmill and outdoor walking. (b) Customized electronic system for policy deployment on quasi-direct driven hip exoskeletons. The forward propagation of trained reference prediction policy with optimized parameters is run on Raspberry Pi 4B to generate the reference position and velocity. The variable impedance control is run on Teensy 4.1 to send the motor command to actuators via CAN at a frequency of 300Hz. The obtained reference data from Raspberry Pi 4B are sent to the Teensy 4.1 via serial port at a frequency of 100Hz. Additionally, an interface is developed by Qt5 to visualize the results on PC.

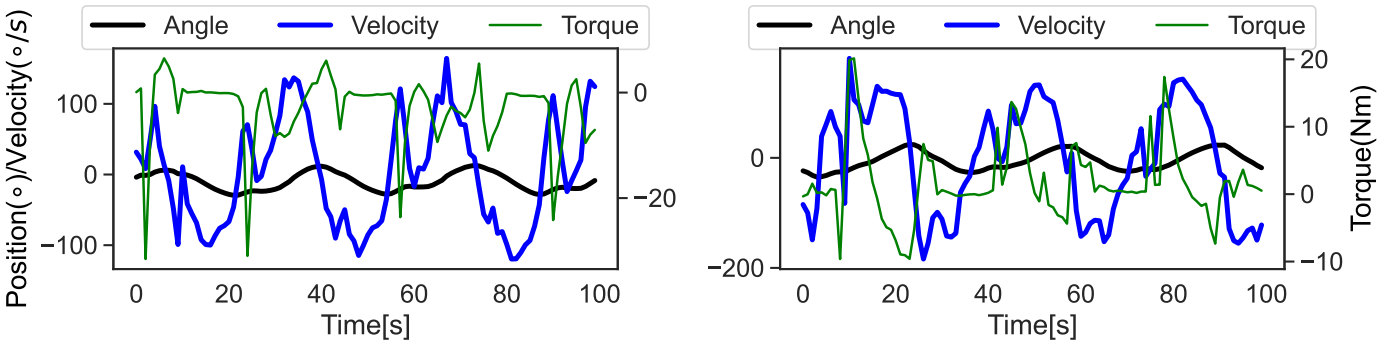

Fig. 7. Alignment of simulation and physical exoskeleton. Hip joint position, velocity and biological torque of learned human agent in simulation and physical robot are compared. The simulation is running at 30Hz and the real-world interactive control is running at 100Hz.

TABLE II
HYPER-PARAMETERS OF PPO

Parameter	Value
Number of Actors	$N = 16$
Number of Iterations	$I = 5000$
Number of Epochs	$L = 10$
Horizon of Each Episode	$T = 2048$
Minibatch Size	$M = 128$
Discount	$\gamma = 0.99$
Clip rate	$\epsilon = 0.2$
Learning Rate	$\alpha_a = 0.0001, \alpha_c = 0.001$
Simulation frequency	600Hz
Control frequency	30Hz
Reward weights	$w^m =$, $w^{as} =$, $w^p =$

TABLE III
DEVICE-RELATED PARAMETERS FOR POLICY DEPLOYMENT

Exoskeleton Name	Values
Hip #1	$\boldsymbol{C}_d^{\min} = \text{diag}(1,1), \boldsymbol{C}_d^{\max} = \text{diag}(1,1)$ $\boldsymbol{K}_d^{\min} = \text{diag}(1,1), \boldsymbol{K}_d^{\max} = \text{diag}(1,1)$ $\boldsymbol{L}_q = \text{diag}(0.001, 0.001),\ f_p = 10\text{Hz}$
Hip #2	$\boldsymbol{C}_d^{\min} = \text{diag}(1,1), \boldsymbol{C}_d^{\max} = \text{diag}(1,1)$ $\boldsymbol{K}_d^{\min} = \text{diag}(1,1), \boldsymbol{K}_d^{\max} = \text{diag}(1,1)$ $\boldsymbol{L}_q = \text{diag}(0.001, 0.001),\ f_p = 10\text{Hz}$
Hip #3	$\boldsymbol{C}_d^{\min} = \text{diag}(1,1), \boldsymbol{C}_d^{\max} = \text{diag}(1,1)$ $\boldsymbol{K}_d^{\min} = \text{diag}(1,1), \boldsymbol{K}_d^{\max} = \text{diag}(1,1)$ $\boldsymbol{L}_q = \text{diag}(0.001, 0.001),\ f_p = 10\text{Hz}$

2) Customized Electronic System for Policy Deployment: The customized electronic system developed for policy deployment is shown in Fig. 6. The trained reference prediction policy will be validated on three hip exoskeletons of walking on treadmill and walking outdoors as depicted in Fig. 6(a). The reference prediction policy is developed on Raspberry Pi 4B. The human kinematics is measured using IMU (LPMS-B2, ALUBI) and the robot kinematics is measured from motor encoders via CAN. The collected state information is processed via low-pass filter and sent to Raspberry Pi 4B at 100Hz. The variable impedance control is run on Teensy 4.1 at 100Hz and the motor command is sent to the QDD actuators at 300Hz. The interface with human users is displayed on PC via Bluetooth at 100Hz.

TABLE IV
USE-RELATED PARAMETERS FOR POLICY DEPLOYMENT

Subject Name	Values
Subject #1	$C = 10,\ \alpha_{\text{TPI}} = 0.01,\ R_q = 0.5\text{deg}$
Subject #2	$C = 10,\ \alpha_{\text{TPI}} = 0.01,\ R_q = 0.5\text{deg}$
Subject #3	$C = 10,\ \alpha_{\text{TPI}} = 0.01,\ R_q = 0.5\text{deg}$

3) Details of Hip Exoskeletons for Policy Deployment: Three hip exoskeletons, as depicted in Fig. 1, were developed and customized by our lab to validate our proposed method. The hip exo $R\#1$ use T-motor (Ak80-9) with the gear ratio of $9:1$, and the hip exo $R\#2$ is using Sig Motors (6010-H) with the gear ratio of $9.67:1$, and a commercial Hip Exo $R\#3$ is developed by Kenqing with the gear ratio of $36:1$.

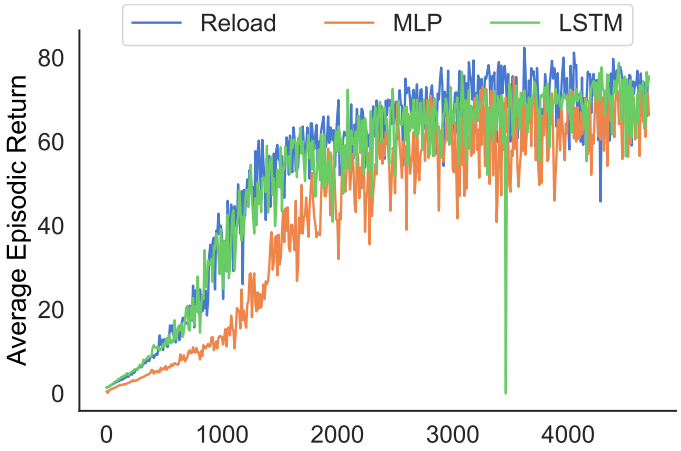

Fig. 8. Episode return of exoskeleton control agent using LSTM-based reference prediction policy vs using MLP-based reference prediction policy .

B. Experimental Protocol

1) Experimental Protocol of Training Reference Prediction Policy in Simulation: The reference prediction policy will be trained by two most popular RL methods for continuous action control. The hyper-parameters of training PPO and train SAC are listed in Table II. Furthermore, two state representation methods, MLP and LSTM, are developed to encode the state for actor and critic networks. The learned network parameters are saved and reloaded by the networks on Raspberry Pi 4B.

2) Experimental Protocol of Walking on Treadmill: Three able-bodied subjects (one female and two males) from our lab are recruited for validation of effectiveness and generalization. Fig. 6(a) illustrates the subject #1 wearing the exoskeleton #2 for treadmill walking. All three subjects are allowed to wear three exoskeletons introduced in Section V-A1, with the learned reference prediction policy for level ground walking on treadmill with three different speeds, $0.75m/s$, $1.25m/s$, and $1.75m/s$ [66]. The generated assistance force profile and power was calculated to evaluate the performance.

3) Experimental Protocol of Walking Outdoor: One representative Subject #1 is asked to walk outdoors following the predefined route including level ground and stairs under changed walking speed.

C. Experimental Results

1) Results of Average Return of Training reference prediction policy : After training 5000 iterations by PPO, the reference prediction policy convergent, the average return over training iterations is plotted in Fig. 8. The LSTM-based actor network can achieve higher episode reward than MLP-based actor network. Furthermore, the reloaded motion imitation network and muscle coordination work can speed up exoskeleton policy learning. Before deploying the trained policy on physical exoskeleton, as shown in the Fig. 9, the colormap to visualize the muscle activation of the human imitation agent assisted by the Exo control policy are plotted. Since the reference motion only covers limited walking velocities, the generalization of learned Exo control policy is validated on other velocities $1.5m/s$, $1.75m/s$, and $2.0m/s$. The assistive torque profile and power under three walking velocities during a gait cycle are plotted in Fig. 10. The generalization on another human model with different height is also tested and plotted in Fig. 10(b) and Fig. 10(c).

Unpowered exo and Non-exo

2) Results of Assistive Torque and Power Deployed on Three Hip Physical Exoskeletons: The hyper parameters of variable impedance control for each physical exoskeleton are listed in Table III. The actual assistive torque and power of three walking speed at $0.75m/s$, $1.25m/s$, and $1.75m/s$ were plotted in Fig. 11.

3) Results of Assistive Torque and Power of Three Healthy Subjects Wearing Exoskeleton $R\#1$: The hyper parameters of impedance adjustment scheme are summarized in Table IV. The baseline method is the controller with the reference output from the learned reference prediction policy and use fixed impedance parameters $\{\boldsymbol{K}, \boldsymbol{C}\}$. The results are illustrated in Fig. 13.

4) Results of Predicted Reference Position and Impedance of Walking Outdoors: A representative subject is instructed to walking outdoors with changing speed. The adapted assistive force profile and corresponding changed impedance were plotted in Fig. **??**.

VI. Conclusion and Discussion

The human motion imitation and muscle coordination network have been trained separately. TCN can also be employed to .

Compared to the sim-to-real RL with dynamic randomization,

Compared to existing interactive control.

References

[1] S. Luo, M. Jiang, S. Zhang, J. Zhu, S. Yu, I. Dominguez Silva, T. Wang, E. Rouse, B. Zhou, H. Yuk *et al.*, "Experiment-free exoskeleton assistance via learning in simulation," *Nature*, vol. 630, no. 8016, pp. 353–359, 2024.

[2] S. Yu, T.-H. Huang, X. Yang, C. Jiao, J. Yang, Y. Chen, J. Yi, and H. Su, "Quasi-direct drive actuation for a lightweight hip exoskeleton with high backdrivability and high bandwidth," *IEEE/ASME Transactions on Mechatronics*, vol. 25, no. 4, pp. 1794–1802, 2020.

[3] D. D. Molinaro, I. Kang, and A. J. Young, "Estimating human joint moments unifies exoskeleton control, reducing user effort," *Science Robotics*, vol. 9, no. 88, p. eadi8852, 2024.

[4] C. Siviy, L. M. Baker, B. T. Quinlivan, F. Porciuncula, K. Swaminathan, L. N. Awad, and C. J. Walsh, "Opportunities and challenges in the development of exoskeletons for locomotor assistance," *Nature Biomedical Engineering*, vol. 7, no. 4, pp. 456–472, 2023.

[5] G. Durandau, W. F. Rampeltshammer, H. van der Kooij, and M. Sartori, "Neuromechanical model-based adaptive control of bilateral ankle exoskeletons: Biological joint torque and electromyogram reduction across walking conditions," *IEEE Transactions on Robotics*, vol. 38, no. 3, pp. 1380–1394, 2022.

[6] R. Baud, A. R. Manzoori, A. Ijspeert, and M. Bouri, "Review of control strategies for lower-limb exoskeletons to assist gait," *Journal of neuroengineering and rehabilitation*, vol. 18, pp. 1–34, 2021.

[7] Y. Chen, S. Miao, G. Chen, J. Ye, C. Fu, B. Liang, S. Song, and X. Li, "Learning to assist different wearers in multitasks: Efficient and individualized human-in-the-loop adaptation framework for lower-limb exoskeleton," *IEEE Transactions on Robotics*, 2024.

[8] R. L. Medrano, G. C. Thomas, C. G. Keais, E. J. Rouse, and R. D. Gregg, "Real-time gait phase and task estimation for controlling a powered ankle exoskeleton on extremely uneven terrain," *IEEE Transactions on Robotics*, vol. 39, no. 3, pp. 2170–2182, 2023.

[9] Y. Qian, Y. Wang, C. Chen, J. Xiong, Y. Leng, H. Yu, and C. Fu, "Predictive locomotion mode recognition and accurate gait phase estimation for hip exoskeleton on various terrains," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6439–6446, 2022.

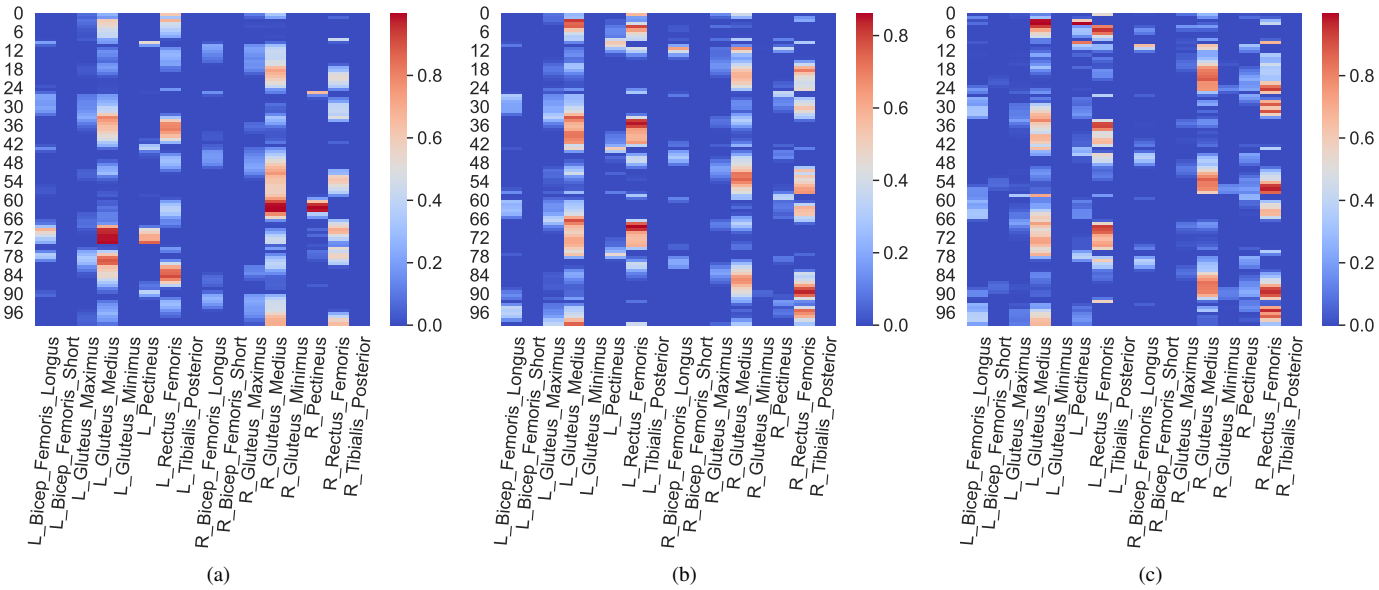

Fig. 9. Performance of muscle activation during evaluation using the learned policy of three training iterations. (a) Policy saved after #3 iterations. (b) Policy saved after #10 iterations (c) Policy saved after #30 iterations.

[10] I. Kang, D. D. Molinaro, S. Duggal, Y. Chen, P. Kunapuli, and A. J. Young, "Real-time gait phase estimation for robotic hip exoskeleton control during multimodal locomotion," *IEEE robotics and automation letters*, vol. 6, no. 2, pp. 3491–3497, 2021.

[11] X. Zhang, E. Tricomi, X. Ma, M. Gomez-Correa, A. Ciaramella, F. Missiroli, L. Mišković, H. Su, and L. Masia, "A lower limb wearable exosuit for improved sitting, standing, and walking efficiency," *IEEE Transactions on Robotics*, 2024.

[12] D. D. Molinaro, K. L. Scherpereel, E. B. Schonhaut, G. Evangelopoulos, M. K. Shepherd, and A. J. Young, "Task-agnostic exoskeleton control via biological joint moment estimation," *Nature*, vol. 635, no. 8038, pp. 337–344, 2024.

[13] P. R. Shetty, J. A. Menezes, S. Song, A. J. Young, and M. K. Shepherd, "Ankle exoskeleton control via data-driven gait estimation for walking, running, and inclines," *IEEE Robotics and Automation Letters*, 2025.

[14] Q. Zhang, J. Si, X. Tu, M. Li, M. D. Lewek, and H. Huang, "Toward task-independent optimal adaptive control of a hip exoskeleton for locomotion assistance in neurorehabilitation," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024.

[15] J. K. Mehr, E. Guo, M. Akbari, V. K. Mushahwar, and M. Tavakoli, "Deep reinforcement learning based personalized locomotion planning for lower-limb exoskeletons," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 5127–5133.

[16] X. Li, Y. Pan, G. Chen, and H. Yu, "Adaptive human–robot interaction control for robots driven by series elastic actuators," *IEEE Transactions on Robotics*, vol. 33, no. 1, pp. 169–182, 2016.

[17] Y. Huo, X. Li, X. Zhang, X. Li, and D. Sun, "Adaptive intention-driven variable impedance control for wearable robots with compliant actuators," *IEEE Transactions on Control Systems Technology*, vol. 31, no. 3, pp. 1308–1323, 2022.

[18] G. Aguirre-Ollinger and H. Yu, "Lower-limb exoskeleton with variable-structure series elastic actuators: Phase-synchronized force control for gait asymmetry correction," *IEEE Transactions on Robotics*, vol. 37, no. 3, pp. 763–779, 2020.

[19] A. Rajagopal, C. L. Dembia, M. S. DeMers, D. D. Delp, J. L. Hicks, and S. L. Delp, "Full-body musculoskeletal model for muscle-driven simulation of human gait," *IEEE transactions on biomedical engineering*, vol. 63, no. 10, pp. 2068–2079, 2016.

[20] C. L. Dembia, N. A. Bianco, A. Falisse, J. L. Hicks, and S. L. Delp, "Opensim moco: musculoskeletal optimal control," *PLOS Computational Biology*, vol. 16, no. 12, p. e1008493, 2020.

[21] V. Firouzi, A. Seyfarth, S. Song, O. von Stryk, and M. Ahmad Sharbafi, "Biomechanical models in the lower-limb exoskeletons development: A review," *Journal of NeuroEngineering and Rehabilitation*, vol. 22, no. 1, p. 12, 2025.

[22] M. Abdullah, A. A. Hulleck, R. Katmah, K. Khalaf, and M. El-Rich, "Multibody dynamics-based musculoskeletal modeling for gait analysis: a systematic review," *Journal of NeuroEngineering and Rehabilitation*, vol. 21, no. 1, p. 178, 2024.

[23] S. Ha, J. Lee, M. van de Panne, Z. Xie, W. Yu, and M. Khadiv, "Learning-based legged locomotion: State of the art and future perspectives," *The International Journal of Robotics Research*, p. 02783649241312698, 2024.

[24] 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.

[25] S. Lee, M. Park, K. Lee, and J. Lee, "Scalable muscle-actuated human simulation and control," *ACM Transactions On Graphics (TOG)*, vol. 38, no. 4, pp. 1–13, 2019.

[26] K. He, C. Zuo, C. Ma, and Y. Sui, "Dynsyn: dynamical synergistic representation for efficient learning and control in overactuated embodied systems," ser. ICML'24. JMLR.org, 2024.

[27] C. Zuo, K. He, J. Shao, and Y. Sui, "Self model for embodied intelligence: Modeling full-body human musculoskeletal system and locomotion control with hierarchical low-dimensional representation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 13 062–13 069.

[28] 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.

[29] 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.

[30] 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," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 13 107–13 114.

[31] S. Song, Ł. Kidziński, X. B. Peng, C. Ong, J. Hicks, S. Levine, C. G. Atkeson, and S. L. Delp, "Deep reinforcement learning for modeling human locomotion control in neuromechanical simulation," *Journal of neuroengineering and rehabilitation*, vol. 18, pp. 1–17, 2021.

[32] J. Weng, E. Hashemi, and A. Arami, "Natural walking with musculoskeletal models using deep reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 4156–4162, 2021.

[33] H.-J. Geiβ, F. Al-Hafez, A. Seyfarth, J. Peters, and D. Tateo, "Exciting action: Investigating efficient exploration for learning musculoskeletal humanoid locomotion," in *2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids)*. IEEE, 2024, pp. 205–212.

[34] S. Luo, G. Androwis, S. Adamovich, E. Nunez, H. Su, and X. Zhou, "Robust walking control of a lower limb rehabilitation exoskeleton coupled

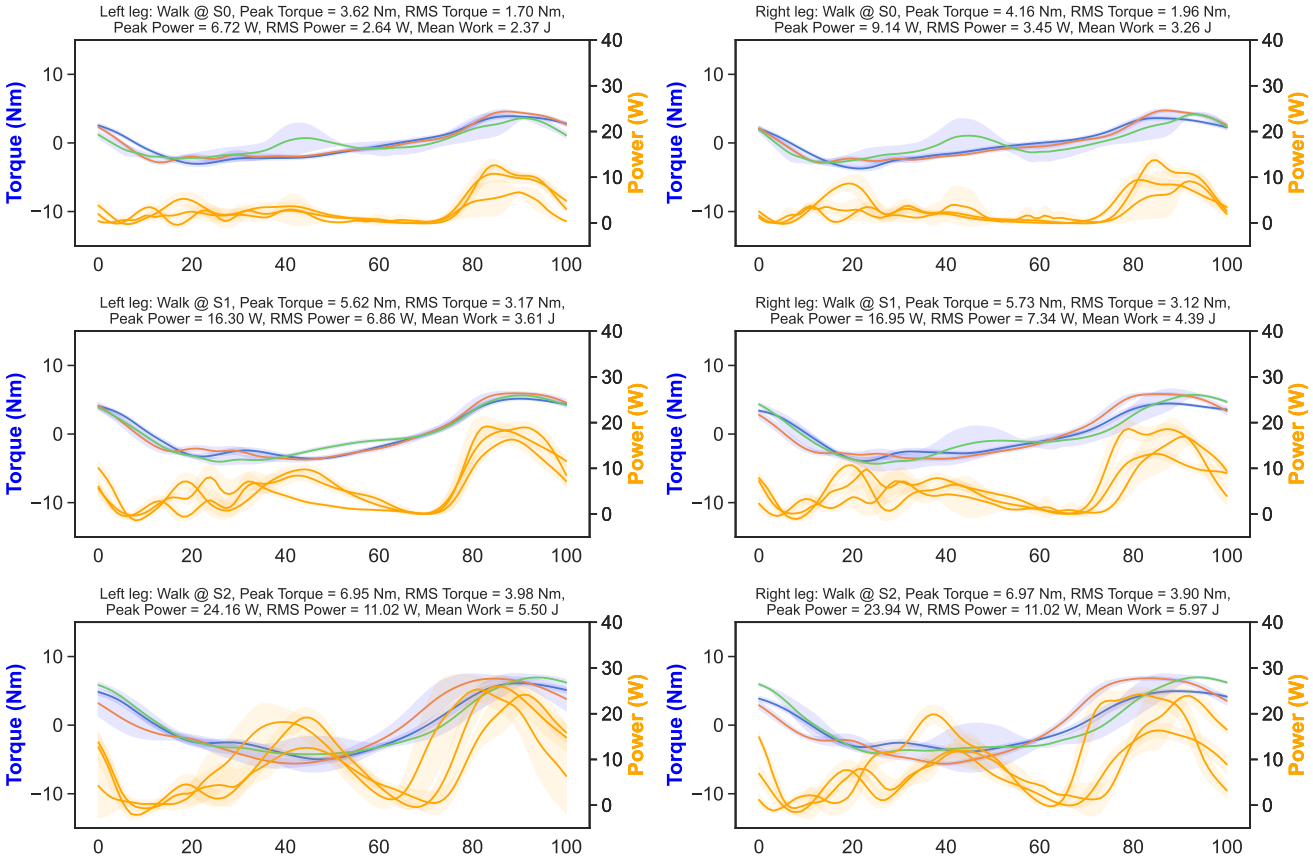

Fig. 10. this figure is too late, move earlier to result section. also plot with matlab, this figure quality looks not good (Will improve the figure quality) A picture to demonstrate one subject wearing three different Hip Exo and all three subjects wearing the same Hip Exo.

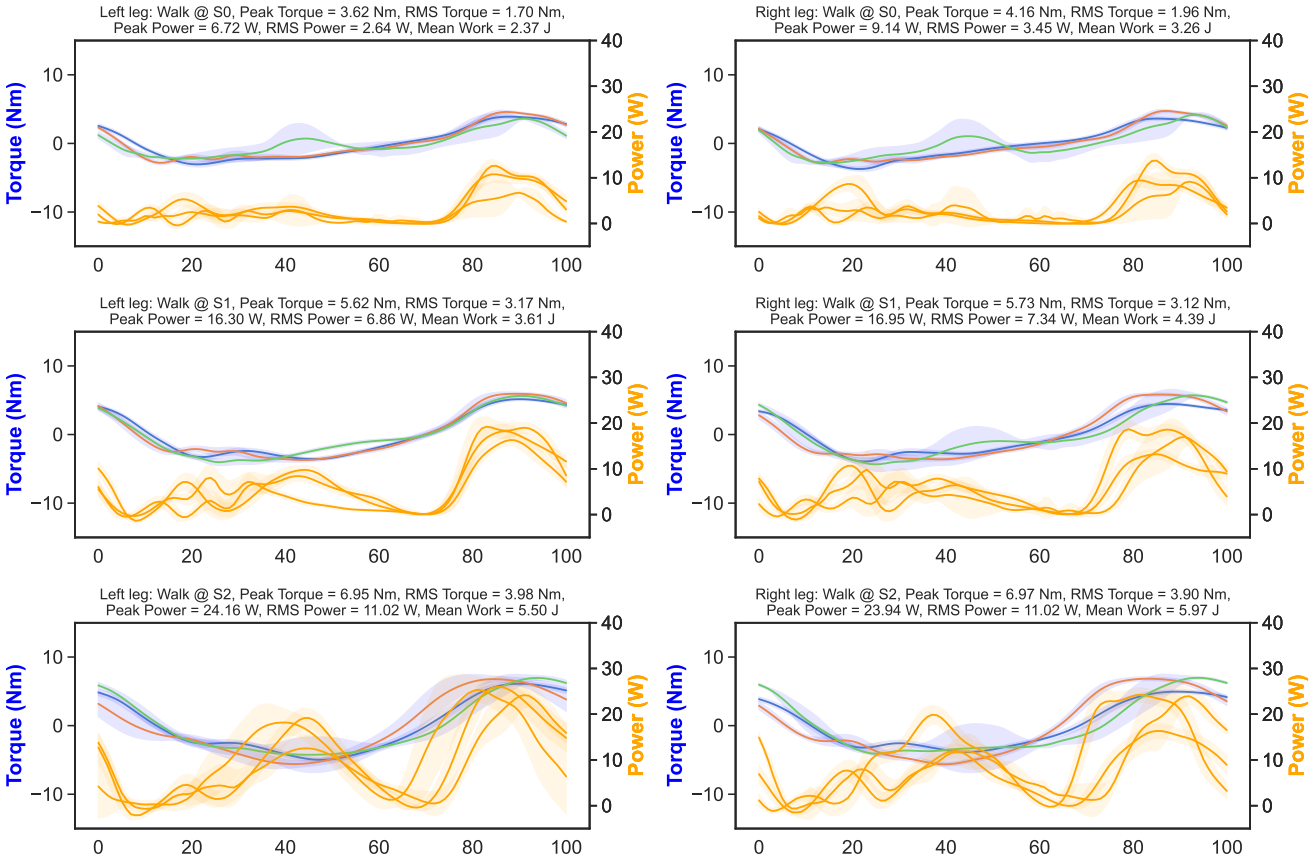

Fig. 11. this figure is too late, move earlier to result section. also plot with matlab, this figure quality looks not good (Will improve the figure quality) A picture to demonstrate one subject wearing three different Hip Exo and all three subjects wearing the same Hip Exo.

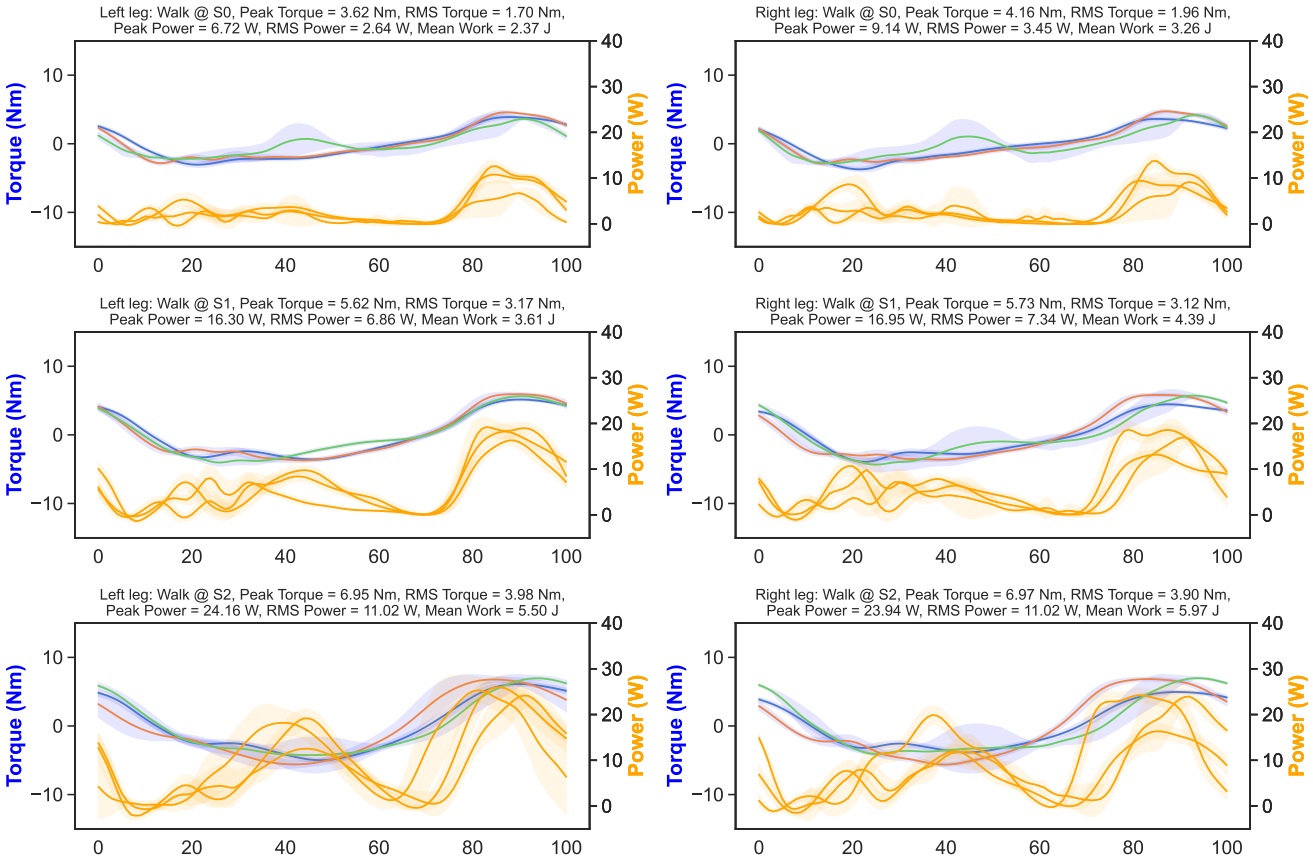

Fig. 12. (Will improve the figure quality) A picture to demonstrate one subject wearing three different Hip Exo and all three subjects wearing the same Hip Exo.

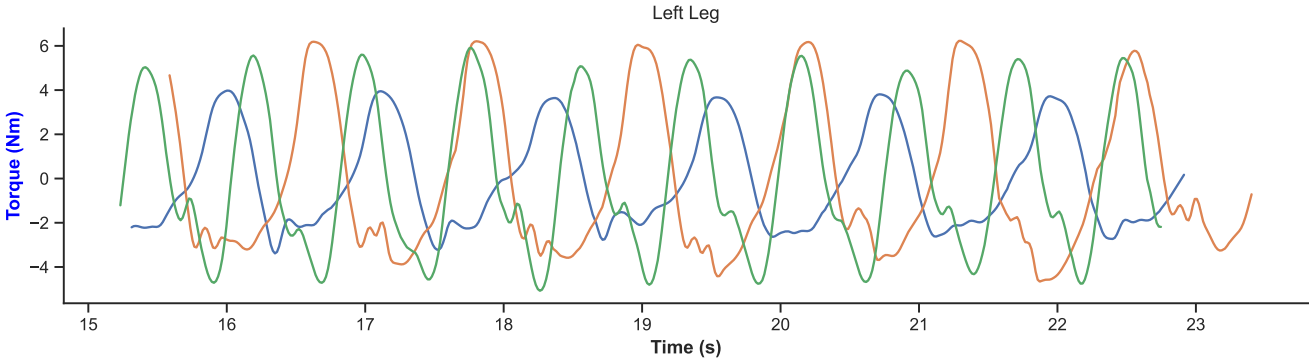

Fig. 13. (Will improve the figure quality) A picture to demonstrate one subject wearing three different Hip Exo and all three subjects wearing the same Hip Exo.

with a musculoskeletal model via deep reinforcement learning," *Journal of neuroengineering and rehabilitation*, vol. 20, no. 1, p. 34, 2023.

[35] J. Lin, R. D. Gregg, and P. B. Shull, "Improving task-agnostic energy shaping control of powered exoskeletons with task/gait classification," *IEEE Robotics and Automation Letters*, 2024.

[36] J. Zhang, P. Fiers, K. A. Witte, R. W. Jackson, K. L. Poggensee, C. G. Atkeson, and S. H. Collins, "Human-in-the-loop optimization of exoskeleton assistance during walking," *Science*, vol. 356, no. 6344, pp. 1280–1284, 2017.

[37] P. Slade, M. J. Kochenderfer, S. L. Delp, and S. H. Collins, "Personalizing exoskeleton assistance while walking in the real world," *Nature*, vol. 610, no. 7931, pp. 277–282, 2022.

[38] Y. Wen, J. Si, A. Brandt, X. Gao, and H. H. Huang, "Online reinforcement learning control for the personalization of a robotic knee prosthesis," *IEEE transactions on cybernetics*, vol. 50, no. 6, pp. 2346–2356, 2019.

[39] U. H. Lee, V. S. Shetty, P. W. Franks, J. Tan, G. Evangelopoulos, S. Ha, and E. J. Rouse, "User preference optimization for control of ankle exoskeletons using sample efficient active learning," *Science Robotics*, vol. 8, no. 83, p. eadg3705, 2023.

[40] X. Zhang, E. Tricomi, F. Missiroli, N. Lotti, C. Bokranz, D. Nicklas, and L. Masia, "Enhancing gait assistance control robustness of a hip exosuit by means of machine learning," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7566–7573, 2022.

[41] 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.

[42] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2017, pp. 23–30.

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

[44] 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.

[45] R. Martín-Martín, M. A. Lee, R. Gardner, S. Savarese, J. Bohg, and A. Garg, "Variable impedance control in end-effector space: An action space for reinforcement learning in contact-rich tasks," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 1010–1017.

[46] O. Nachum, S. S. Gu, H. Lee, and S. Levine, "Data-efficient hierarchical

reinforcement learning," *Advances in neural information processing systems*, vol. 31, 2018.
[47] Z. Hou, W. Yang, R. Chen, P. Feng, and J. Xu, "A hierarchical compliance-based contextual policy search for robotic manipulation tasks with multiple objectives," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 4, pp. 5444–5455, 2022.
[48] M. K. Ishmael, D. Archangeli, and T. Lenzi, "A powered hip exoskeleton with high torque density for walking, running, and stair ascent," *IEEE/ASME transactions on mechatronics*, vol. 27, no. 6, pp. 4561–4572, 2022.
[49] G. B. Margolis, G. Yang, K. Paigwar, T. Chen, and P. Agrawal, "Rapid locomotion via reinforcement learning," *The International Journal of Robotics Research*, vol. 43, no. 4, pp. 572–587, 2024.
[50] C. Berg, V. Caggiano, and V. Kumar, "Sar: Generalization of physiological agility and dexterity via synergistic action representation," *Autonomous Robots*, vol. 48, no. 8, p. 28, 2024.
[51] Y. Feng, X. Xu, and L. Liu, "Musclevae: Model-based controllers of muscle-actuated characters," in *SIGGRAPH Asia 2023 Conference Papers*, 2023, pp. 1–11.
[52] Z. Li, X. Cheng, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Reinforcement learning for robust parameterized locomotion control of bipedal robots," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 2811–2817.
[53] G. Feng, H. Zhang, Z. Li, X. B. Peng, B. Basireddy, L. Yue, Z. Song, L. Yang, Y. Liu, K. Sreenath *et al.*, "Genloco: Generalized locomotion controllers for quadrupedal robots," in *Conference on Robot Learning*. PMLR, 2023, pp. 1893–1903.
[54] 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.
[55] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
[56] J. Xing, I. Geles, Y. Song, E. Aljalbout, and D. Scaramuzza, "Multi-task reinforcement learning for quadrotors," *IEEE Robotics and Automation Letters*, 2024.
[57] S. Sodhani, A. Zhang, and J. Pineau, "Multi-task reinforcement learning with context-based representations," in *International Conference on Machine Learning*. PMLR, 2021, pp. 9767–9779.
[58] J. He, K. Li, Y. Zang, H. Fu, Q. Fu, J. Xing, and J. Cheng, "Efficient multi-task reinforcement learning with cross-task policy guidance," *Advances in Neural Information Processing Systems*, vol. 37, pp. 117 997–118 024, 2024.
[59] Z. Hou, T. Ma, W. Wang, and H. Yu, "Contextual policy search for task-level adaptation in physical human–robot interaction," *IEEE/ASME Transactions on Mechatronics*, 2025.
[60] K. Mülling, J. Kober, O. Kroemer, and J. Peters, "Learning to select and generalize striking movements in robot table tennis," *The International Journal of Robotics Research*, vol. 32, no. 3, pp. 263–279, 2013.
[61] C. Yang, K. Yuan, Q. Zhu, W. Yu, and Z. Li, "Multi-expert learning of adaptive legged locomotion," *Science Robotics*, vol. 5, no. 49, p. eabb2174, 2020.
[62] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. PMLR, 2018, pp. 1861–1870.
[63] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 3803–3810.
[64] Y. Li and S. S. Ge, "Human–robot collaboration based on motion intention estimation," *IEEE/ASME Transactions on Mechatronics*, vol. 19, no. 3, pp. 1007–1014, 2013.
[65] A. Kanazawa, J. Kinugawa, and K. Kosuge, "Adaptive motion planning for a collaborative robot based on prediction uncertainty to enhance human safety and work efficiency," *IEEE Transactions on Robotics*, vol. 35, no. 4, pp. 817–832, 2019.
[66] S. Fritz and M. Lusardi, "White paper:"walking speed: the sixth vital sign"," *Journal of geriatric physical therapy*, vol. 32, no. 2, pp. 2–5, 2009.