ModSkill: Physical Character Skill Modularization

Yiming Huang[1], Zhiyang Dou[1,2], Lingjie Liu[1]
[1]University of Pennsylvania, [2]The University of Hong Kong
{ymhuang9,zydou,lingjie.liu}@seas.upenn.edu

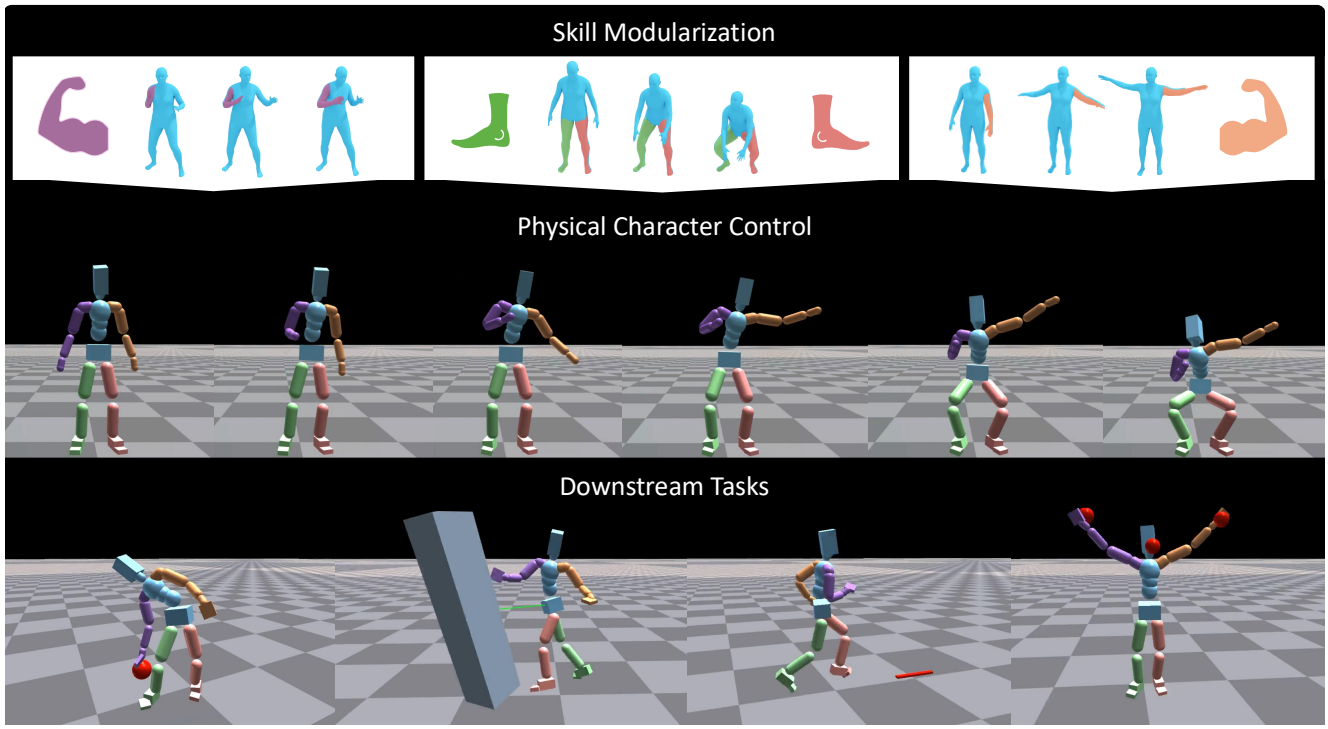

Figure 1. We propose a modularized skill learning framework, ModSkill, which incorporates body-part-level inductive bias for motor skill acquisition. ModSkill decouples full-body motion into skill embeddings for controlling individual body parts. Learned from large-scale motion datasets, these modular skills can be combined to control a simulated character to perform diverse motions, such as the Usain Bolt pose, and seamlessly reused for various downstream tasks.

Abstract

Human motion is highly diverse and dynamic, posing challenges for imitation learning algorithms that aim to generalize motor skills for controlling simulated characters. Prior methods typically rely on a universal full-body controller for tracking reference motion (tracking-based model) or a unified full-body skill embedding space (skill embedding). However, these approaches often struggle to generalize and scale to larger motion datasets. In this work, we introduce a novel skill learning framework, ModSkill, that decouples complex full-body skills into compositional, modular skills for independent body parts, leveraging body structure-inspired inductive bias to enhance skill learning performance. Our framework features a skill modularization attention mechanism that processes policy observations into modular skill embeddings that guide low-level controllers for each body part. We further propose Generative Adaptive Sampling for Active Skill Learning, using large motion generation models to adaptively enhance policy learning in challenging tracking scenarios. Results show that this modularized skill learning framework, enhanced by generative sampling, outperforms existing methods in precise full-body motion tracking and enables reusable skill embeddings for diverse goal-driven tasks.

1. Introduction

Physically simulated characters are widely used in animation [23, 59], VR/AR [36, 62], and robotic tasks [12, 13]. When combined with large-scale human motion capture

data [38], imitation learning policies allow such characters to perform a wide range of motion skills. [19, 35, 37, 53].

Previous approaches to physical character skill learning can be typically classified into two categories. *Tracking-based methods* train controllers to imitate reference motions by tracking target pose sequences from motion clips. Recently, progressive learning techniques have been utilized to gradually extract more complex full-body motion skills from diverse data into a set of expert controllers [19, 35]. The motor skills learned from this mixture of experts can be distilled into a compact universal motion representation that offers broader coverage of human motion [37]. However, such methods still fall short in addressing scalability challenges with larger datasets, requiring more experts and increasing manual effort for skill extraction. On the other hand, *Skill embedding methods* employ hierarchical frameworks that pre-train compact skill embedding spaces, which are then repurposed for high-level tasks guided by carefully designed, task-specific rewards [6, 18, 47, 52, 70]. As of yet, the expressivity limitations of these latent skill spaces hinder their ability to capture the diverse range of full-body human motion skills present in larger motion datasets. Moreover, the inherent diversity of human motion poses substantial generalization challenges, making existing methods susceptible to overfitting.

In this work, we argue that human motion is inherently modular, as evidenced by neuroscience and evolutionary developmental biology [3, 14, 21, 51]. Motivated by this modularity, we pursue *Skill Modularization* to introduce an inductive bias to the skill learning process that decouples full-body motion for independent control of body parts. Compared to full-body skills, modular skills for individual body parts are not only more compact but also exhibit a compositional quality that can generate highly diverse full-body motion [16, 22]. By emphasizing these compact part-level skill spaces, we can effectively learn discriminative features (Sec. 4.3) that enhance motion skill learning.

Building on this intuition, we introduce ModSkill, a novel modularized skill learning framework that utilizes body-part inductive bias to effectively decouple full-body motor skills in large-scale motion datasets [38] into reusable, part-specific skills that each guide an independent low-level controller. Our approach incorporates a skill modularization attention mechanism that effectively captures relationships between part-specific observations and produces spherical modular skill embeddings for each controller, guiding the corresponding body parts of the simulated agent. The attention mechanism encourages full-body consistency by enabling information sharing across body parts during data-driven imitation learning.

Imitation policies trained with DRL methods like PPO [49] often overfit to training samples, limiting generalization. To address this, we further propose a novel *Active Skill Learning* scheme using a Generative Adaptive Sampling strategy. This approach uses pre-trained large motion generation models [54] to generate synthetic samples for challenging motions. Unlike prior efforts [6, 35], where resampled motion clips consistently come from a fixed motion dataset, our method applies a powerful generative model to provide prior-level resampling capabilities, thereby enhancing the skill learning process with more diverse motion samples. We demonstrate the effectiveness of this sampling scheme for skill learning (Sec. 4.4).

ModSkill, with its modularity and compositionality, effectively learns diverse motor skills that are reusable for various downstream tasks. We conduct extensive experiments to demonstrate that ModSkill achieves state-of-the-art performance both on full-body tracking-based tasks and across a wide range of generative, goal-driven benchmarks, including steering, reaching, striking, and VR tracking. In summary, our contributions are three-fold:

- We propose a modularized skill learning framework that integrates a body-part level inductive bias, i.e., skill modularization, within an attention mechanism for extracting part-specific skill embeddings that guide independent low-level controllers for each body part.
- We introduce an Active Skill Learning scheme with a Generative Adaptive Sampling strategy using the generative motion prior of large motion generation models that enhances motion imitation performance.
- Our controllers achieve superior performance in precise motion tracking, and the learned part-wise skills are effectively reusable for generative downstream tasks.

2. Related Work

Physics-based Motion Imitation. Reproducing diverse and realistic human motions with physics-based characters has been a longstanding area of focus [7, 11, 24, 26–29, 41–44, 60, 70]. Given the diversity of human motion, many approaches focus on task-specific scenarios requiring only a subset of motor skills[5, 11, 61, 64, 70]. To generalize motion controllers, adversarial motion priors have been introduced [46]. Mixtures of experts have also been widely used, where each expert can focus on a specific task, bridging the gap between task-specific and generalized motion imitation [34, 35, 45, 56, 63]. However, relying on multiple experts introduces scalability challenges, as adding more experts may require increasing manual effort for progressive skill learning. Furthermore, prior methods predominantly focus on full-body skills, which may limit the expressiveness of the controller. In contrast, we address challenges in motion imitation by decoupling full-body motion into part-specific motor skills, leveraging the inductive bias of modularization to improve skill learning.

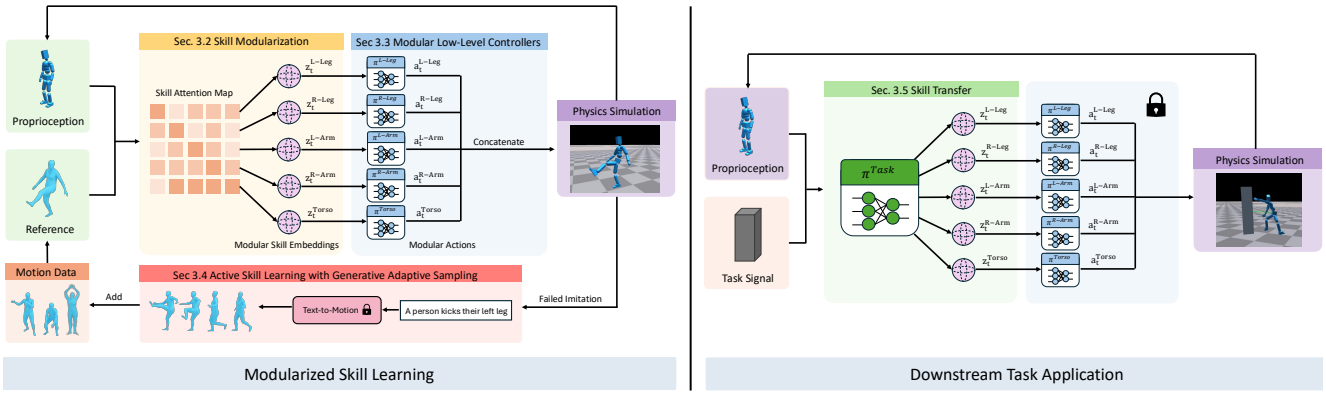

Figure 2. **Left**: We extract modular skills from a large-scale motion dataset using a motion imitation objective, enabling low-level controllers to control various body parts of a physically simulated character. Active skill learning, through adaptive sampling from an off-the-shelf motion generation model, further enhances policy performance. **Right**: The learned modular skills can be transferred to downstream tasks by freezing the low-level controllers and training a high-level policy with task-specific rewards.

Physics-based Skill Embedding. Adversarial learning and motion imitation enable the development of reusable motion skills for diverse downstream tasks. In hierarchical learning settings, adversarial methods have been used to map structured latent embeddings to reusable low-level motor skills [6, 18, 47, 52]. While these skills transfer well to specialized tasks with tailored motion data, they struggle to scale to more diverse motions. On the other hand, efforts have been made to model probabilistic latent spaces that can capture motor skills from larger datasets [25, 40, 65, 74]. Moreover, prior work has shown that precise motion imitation policy networks can be distilled into universal motion representations, enabling coverage of large-scale motion datasets [19, 37, 56]. In contrast, our method focuses on decoupling motor skill learning during motion imitation by emphasizing the formulation of modular, part-level skills, allowing us to achieve effective and reusable skills without the need for additional distillation.

Kinematics-based Motion Generation. The rich latent space of kinematics-based motion generation models enables the generation of diverse motion patterns from multi-modal conditions [4, 9, 17, 54, 71]. Prior work has shown that synthetic motion data provides valuable supervision for training and refining generative models [2, 8, 66]. Furthermore, when integrated with a physics-based controller, motion generation models can effectively guide task planning and execution [48, 55, 69]. In this work, we propose to leverage the expressive power of large motion generation models to adaptively create synthetic examples of challenging motion imitation scenarios, enhancing policy learning and generalization.

Part-level Motion Learning. Prior work has shown advantages of body-part-level motion learning in both kinematics-based motion generation models [15, 16, 57, 72, 73] and physical controllers [1, 20, 22, 50]. Lee et al. [22] use an assembler module to combine motion signals from different sources before directing a single controller. PMP [42] adapts adversarial motion priors to body parts to extract specialized style rewards from different motion datasets. Similarly, Xu et al. [67] leverage multiple discriminators for training a control policy to imitate body part reference motions from different sources. However, both methods still rely on a single controller for full-body control, limiting motor skill decoupling across body parts. PartwiseMPC [20] takes a step further by decoupling motion planning, enabling planning for both independent body parts and the whole body, improving generalization to unseen environments. In contrast, our approach introduces compact, modularized skill embeddings and low-level controllers for individual body parts, offering reusable modular skills for both precise motion imitation and generative, goal-driven tasks.

3. ModSkill

In this paper, we present ModSkill, a modularized framework for extracting body-part-level motor skills from large-scale motion datasets through imitation learning. As illustrated in Fig. 2, our policy network consists of two key components for modularized skill learning: 1) a skill modularization attention mechanism (Sec. 3.2) that produces spherical embeddings to capture body-part-specific skills, and 2) a set of low-level skill-conditioned controllers (Sec. 3.3) that control the movement of individual body parts. To further improve policy performance, we introduce a generative adaptive sampling strategy that incorporates synthetic data from motion generation models into policy training (Sec. 3.4). Additionally, we show that the modular skills learned in our framework can be effectively transferred to downstream tasks via a high-level, task-specific policy (Sec. 3.5).

3.1. Preliminaries

Given a reference motion sequence of T frames, $s^r_{1:T}$, our policy network, denoted as π_{ModSkill}, is tasked to control a SMPL-based simulated humanoid agent [30, 35] to imitate the reference motion. We model the policy network as a Markov Decision Process (MDP), $M = \langle S, A, T, R, \gamma \rangle$, where S, A, T, R, γ represent the state space, action space, transition dynamics, reward function, and discount factor.

States and Actions. Following prior work [35], the state s_t at time t is composed of the proprioceptive state s^p_t and the reference state s^r_t. Specifically, s^p_t describes the current simulated configuration of the agent and is defined as:

$$s^p_t := (\theta_t, p_t, v_t, \omega_t) \tag{1}$$

where θ_t, p_t, v_t, and ω_t are the simulated joint rotations, positions, velocities, and angular velocities, respectively. The reference state s^r_t encodes the target joint poses for time step $t+1$, and the differences between the target joint poses and velocities of time step $t+1$ and the corresponding simulated values at the current time step, t:

$$s^r_t := (\hat{\theta}_{t+1} \ominus \theta_t, \hat{p}_{t+1} - p_t, \hat{v}_{t+1} - v_t, \hat{\omega}_{t+1} - \omega_t, \hat{\theta}_{t+1}, \hat{p}_{t+1}) \tag{2}$$

where $\ominus$ denotes rotation difference. Here, $\hat{\theta}_{t+1}$, $\hat{p}_{t+1}$, $\hat{v}_{t+1}$, and $\hat{\omega}_{t+1}$ represent the reference joint rotations, positions, velocities, and angular velocities, respectively. Both s^r_t and s^p_t are canonicalized with respect to the agent's local coordinate frame. PD (proportional-derivative) controllers are used, enabling the agent to follow desired joint angles and velocities based on the reference motion. The action space specifies PD control targets for each actuated joint, without using residual forces [68] or residual control [33].

Reward. We define the reward term as the sum of an imitation reward, $r_{\text{imitation}}$, a full-body discriminator reward, r_{amp}, to encourage *natural and consistent* full-body motion from body-part-level motor skills [46], and an energy penalty reward, r_{energy}, to encourage smoother motion [62]:

$$r := r_{\text{imitation}} + r_{\text{amp}} + r_{\text{energy}} \tag{3}$$

Specifically, the imitation reward $r_{\text{imitation}}$ encourages the humanoid agent to imitate the reference motion by minimizing the difference between the translation $(p_t, \hat{p}_t)$, rotation $(\theta_t, \hat{\theta}_t)$, linear velocity $(v_t, \hat{v}_t)$, and angular velocity $(\omega_t, \hat{\omega}_t)$ of the simulated character and the target motion:

$$\begin{aligned} r_{\text{imitation}} &:= w_p e^{-\lambda_p \|p_t - \hat{p}_t\|} + w_\theta e^{-\lambda_\theta \|\theta_t - \hat{\theta}_t\|} \\ &+ w_v e^{-\lambda_v \|v_t - \hat{v}_t\|} + w_\omega e^{-\lambda_\omega \|\omega_t - \hat{\omega}_t\|} \end{aligned} \tag{4}$$

where $w_{\{\cdot\}}$, $\lambda_{\{\cdot\}}$ denote the corresponding weights.

3.2. Skill Modularization

To achieve modularized skill learning, we partition the rigid bodies of the simulated agent into K sets, each corresponding to a distinct body part. In this work, we set $K = 5$, corresponding to the set of body parts, $\mathcal{P} := \{$Left Leg (L-Leg), Right Leg (R-Leg), Left Arm (L-Arm), Right Arm (R-Arm), Torso$\}$. Note that the specific partitioning is not restrictive, and can be adapted to suit different use cases or system configurations (see Supplementary).

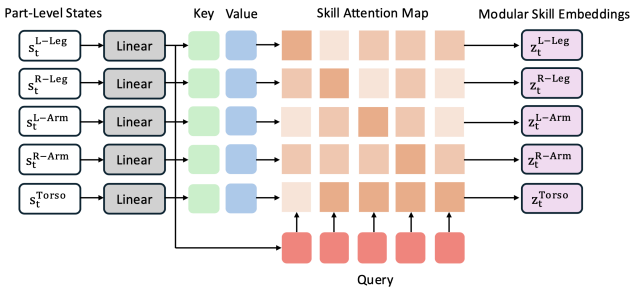

Figure 3. **Skill Modularization Attention Mechanism:** Given partial states for each body part, attention between body parts allows information sharing for producing modular skill embeddings.

Let the state of all joints for body part $k \in \mathcal{P}$ at time t be denoted as s^k_t. Our policy network incorporates an attention mechanism to enable high-level information sharing across body parts for effective skill modularization and encouraging full-body consistency. As shown in Fig. 3, for each body part $k \in \mathcal{P}$, the corresponding state s^k_t is projected into three vectors: key $\mathbf{K}^k_t$, query $\mathbf{Q}^k_t$, and value $\mathbf{V}^k_t$. The attention scores between the query $\mathbf{Q}^k_t$ of the current body part k and the keys $\mathbf{K}^{k'}_t$ from all body parts $k' \in \mathcal{P}$ are computed by calculating the scaled dot-product between the query and each key, and then passed through a softmax function to obtain the attention weights:

$$\alpha^{k,k'}_t = \text{softmax}\left(\frac{\mathbf{Q}^k_t \cdot \mathbf{K}^{k'}_t}{\sqrt{d}}\right) \tag{5}$$

where d is the dimension of the query and key vectors. The attention weights $\alpha^{k,k'}_t$ indicate the relative importance of the states of different body parts when computing the skill embedding for body part k. The skill embedding z^k_t for each body part is then obtained by computing a weighted sum of the value vectors $\mathbf{V}^{k'}_t$ from all body parts, with the attention weights serving as the coefficients:

$$z^k_t = \sum_{k'} \alpha^{k,k'}_t \mathbf{V}^{k'}_t \tag{6}$$

z^k_t is normalized with respect to $\|z^k_t\|$ to lie on the unit sphere. This normalization ensures that the skill embeddings are constrained within a consistent space, allowing for more stable learning [47]. By sharing information via this attention mechanism, each body part can focus on different aspects of the overall state for modularized skill learning.

3.3. Modular Low-Level Controllers

To further enhance the flexibility and effectiveness of our policy network, we implement modular low-level con-

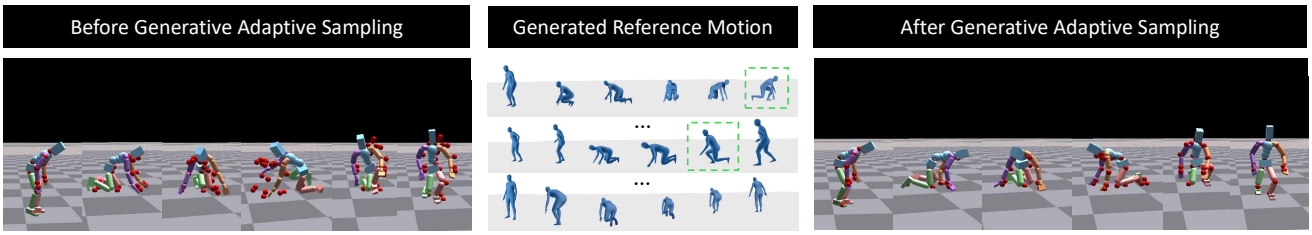

Figure 4. **Generative Adaptive Sampling**: When the policy struggles to track a challenging motion, such as crawling (left), we adaptively generate new samples (middle) using motion generation models. These generated sequences introduce diverse variations, including changes in motion trajectories and leg usage (highlighted in green), which allow the policy to effectively learn motor skills and successfully track the challenging scenario (right). Motion frames are displayed sequentially from left to right. Red spheres indicate target joint locations.

trollers that operate alongside the skill modularization attention mechanism. For each body part k, we designate a low-level controller $\pi^k = \mathcal{N}(\mu(z_t^k), \sigma)$, which models a Gaussian distribution with fixed diagonal covariance. The skill embeddings, z_t^k, generated by the attention mechanism serve as the input for these controllers, which process the information to produce targeted actions a_t^k for the actuated rigid bodies corresponding to body part k. The produced actions for each controller are concatenated to form the full-body PD target, denoted as a_t, for controlling the simulated agent. Using the same setup in [46], we incorporate a full-body discriminator D_{amp} that computes a real/fake value based on the current full-body proprioception of the humanoid. This style signal encourages the formulation of natural full-body motion from modular part-level skills. By decoupling the action prediction into specialized modules, we enhance the flexibility of the network, enabling effective imitation for a wide range of motions.

3.4. Generative Adaptive Sampling

Active skill learning enables efficient policy training by directing the learning process toward more informative regions of the skill space. Denote $\mathcal{S}$ as the skill space and p_θ as the distribution of trajectories produced by the policy network, parameterized by θ. The objective of skill learning can be defined as maximizing the expected return: $J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)}[R(\tau)]$, where $\tau = \{(s_t, a_t)\}_{t=1}^T$ denotes a sampled trajectory, and $R(\tau)$ represents the cumulative reward. Prior methods [6, 35, 37] adaptively sample motion sequences based on their failure probability. However, this approach operates on a fixed dataset $\mathcal{D}_{\text{train}}$, which may suffer from overfitting to challenging motion sequences.

Instead of relying solely on the fixed training set, we propose a Generative Adaptive Sampling strategy. For each failed sample τ_{fail}, we synthesize N motion sequences $\{\tilde{\tau}_i\}_{i=1}^N$ using an off-the-shelf text-to-motion diffusion model [54], conditioned on the corresponding HumanML3D text descriptions $\mathcal{T}(\tau_{\text{fail}})$ [9]:

$$\tilde{\tau}_i \sim p_\phi(\tau | \mathcal{T}(\tau_{\text{fail}})) \tag{7}$$

where $p_\phi(\tau|\mathcal{T})$ denotes the conditional generative distribution of the diffusion model parameterized by ϕ. This process ensures that the generated motions maintain semantic consistency with the failed motion while introducing diverse variations. To mitigate the impact of unrealistic motion artifacts from synthetic data, we utilize the adversarial discriminator to filter out "fake" samples:

$$\mathcal{F}(\tilde{\tau}_i) = \begin{cases} \tilde{\tau}_i & \text{if } D_{\text{amp}}(\tilde{\tau}_i) \geq 0.5 \\ \emptyset & \text{if } D_{\text{amp}}(\tilde{\tau}_i) < 0.5 \end{cases} \tag{8}$$

where $\emptyset$ indicates that the sample is discarded. Finally, we integrate the filtered synthetic motion samples into the training process by augmenting the dataset: $\mathcal{D}_{\text{train}} \leftarrow \mathcal{D}_{\text{train}} \cup \{\mathcal{F}(\tilde{\tau}_i)\}_{i=1}^N$. Qualitative results in Fig. 4 demonstrate that this expanded dataset provides more balanced and comprehensive motion samples, enabling the policy to generalize better across diverse skill variations.

3.5. Skill Transfer for Downstream Tasks

After π_{ModSkill} converges, a high-level policy $\pi_{\text{Task}}(z_t^{k_1}, ..., z_t^{k_K} | s_t^p, s_t^g)$ can be trained to apply the learned modular skills to downstream tasks, where s_t^p and s_t^g represent the proprioceptive state and task-specific goal signal, respectively, and $z_t^{k_1}, ..., z_t^{k_K}$ indicate the corresponding spherical skill embeddings for each body part $k_1, ..., k_K \in \mathcal{P}$. Similar to the low-level controllers, each high-level task policy is modeled as a Gaussian distribution with a fixed diagonal covariance: $\mathcal{N}(\mu_{\text{Task}}(s_t^p, s_t^g), \sigma_{\text{Task}})$. When training the high-level policy, we freeze the low-level controllers to preserve the learned modular motor skills. In this work, we demonstrate the effectiveness and reusability of our modular part-level skills on a set of generative motion tasks. Details regarding the setup of each task are provided in the supplementary material.

4. Evaluation

Experiment Settings. *Motion Tracking Task:* We evaluate ModSkill's performance on the full-body motion tracking task, comparing it against state-of-the-art motion trackers

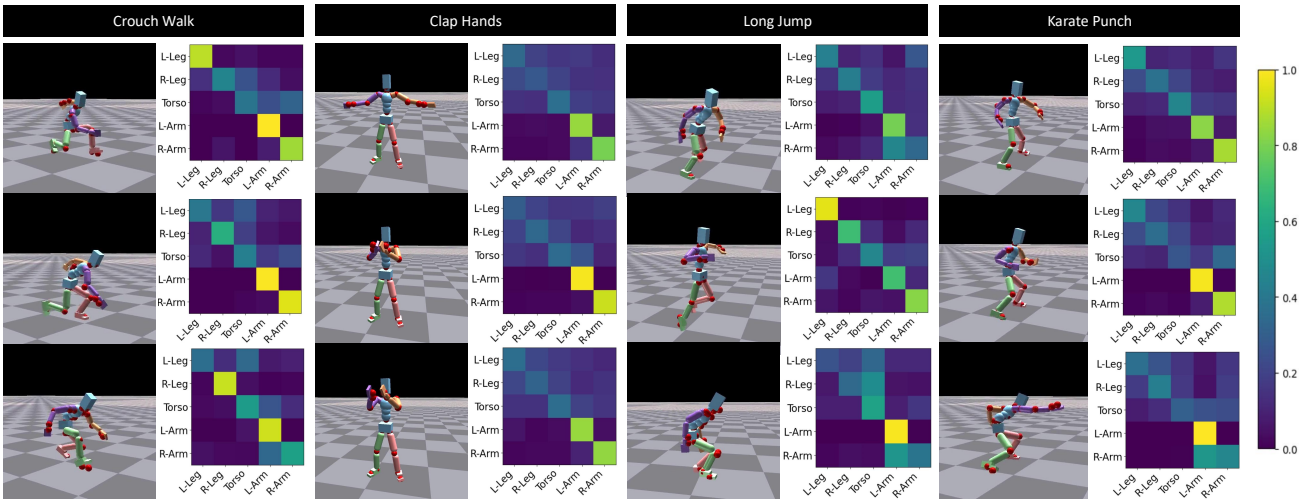

Figure 5. Qualitative results of motion tracking on AMASS-Test (left) and the corresponding attention maps during skill modularization (right). Red spheres indicate the target joint locations in the simulation. The attention maps reveal that, when extracting skill embeddings, body parts are capable of focusing on their own specific information for precise movements (e.g., hands during clapping) while also attending to holistic information across body parts for full-body consistency (e.g., stabilizing lower body when landing from a jump).

UHC [32], PHC [35], PHC+ and PULSE [37]. *Motion Skill Embedding Task:* We also highlight the reusability of our modular skill embeddings by applying ModSkill to generative tasks such as reaching, steering, striking, and VR tracking, and comparing its performance with reusable skill representations: PULSE [37] and ASE [47]. Following [37], we adapt ASE to produce per-frame latent skill embeddings for a fair comparison. Notably, while the original ASE uses a full-body adversarial motion prior, we will also compare the impact of partwise adversarial motion priors [1] on learning skill embeddings and high-level tasks (ASE-PMP). We adopt the same body part partition as our framework to formulate ASE-PMP.

Datasets. For training and testing the full-body and VR tracking policies, we utilize the cleaned AMASS training set and test set, respectively [35]. For the strike and reach tasks, we sample initial states from the AMASS training set. For speed tasks, we follow [59] to use a subset of AMASS of only locomotion for initial state sampling.

Metrics. For motion imitation and VR tracking, we report the global end-effector mean per-joint position error ($E_{\text{g-mpjpe}}$) and root-relative end-effector mean per-joint position error (E_{mpjpe}) in millimeters. We also compare physics-based metrics, including acceleration error (E_{acc}) in mm/frame^2 and velocity error (E_{vel}) in mm/frame. Following [35], we define the success rate (Succ) as the percentage of time the average per-joint error remains within 0.5 meters of the reference motion. For VR tracking, the success rate is based on only three body points (Head, Left Hand, Right Hand). For generative tasks (reach, steer, strike), we compare the undiscounted return normalized by the maximum possible reward per episode. We also compute the Average Pairwise Distance (APD) [6, 10, 31, 58] to measure the diversity of motion sequences produced for each task, where a larger value indicates higher diversity.

Implementation Details. All physics simulations are conducted using Isaac Gym [39]. Our policy network is trained with 3072 parallel environments on a single NVIDIA A6000 GPU for approximately 2e9 steps. For PULSE, we use the original model settings. For ModSkill and the ASE baselines, all low-level controller networks are four-layer perceptrons (MLPs) with dimensions [2048, 1536, 1024, 512]. Discriminators and encoders for adversarial skill learning are two-layer MLPs with dimensions [1024, 512]. Each high-level policy for downstream tasks is a three-layer MLP with dimensions [2048, 1024, 512]. The latent dimension of the skill embeddings is 64. The controllers operate at 30 Hz, while the simulation runs at 60 Hz. Additional details are provided in the supplementary material.

4.1. Motion Tracking

Tab. 1 and Tab. 2 show the performance of our method on the AMASS train and test sets for the full-body motion tracking task and VR tracking task, respectively. For both tracking tasks, our modular policy network outperforms baselines, reducing tracking errors on both training and test motion sequences. Fig. 5 shows qualitative results of full-body tracking on AMASS test, along with attention maps from skill modularization. Each row of an attention map represents a queried body part, while the columns correspond to other body parts serving as keys. The attention weights reflect how information from each key contributes to the queried part's skill embedding. For precise, localized

actions such as hand clapping or slow steps, body parts primarily attend to their own features. In contrast, for motions requiring full-body coordination, such as landing or punching from a squat, multiple body parts share information to maintain consistency and stability. Our experiments show that by modularizing skills, a single network can achieve better imitation accuracy and generalization capabilities.

Table 1. Full-body tracking results on AMASS train and test.

	AMASS-Train					AMASS-Test				
Method	Succ ↑	$E_{\text{g-mpjpe}}$ ↓	E_{mpjpe} ↓	E_{acc} ↓	E_{vel} ↓	Succ ↑	$E_{\text{g-mpjpe}}$ ↓	E_{mpjpe} ↓	E_{acc} ↓	E_{vel} ↓
UHC	97.0%	36.4	25.1	4.4	5.9	96.4%	50.0	31.2	9.7	12.1
PHC	98.9%	37.5	26.9	3.3	4.9	97.1%	47.5	40.0	6.8	9.1
PULSE	99.8%	39.2	35.0	3.1	5.2	97.1%	54.1	43.5	7.0	10.3
PHC+	**100%**	26.1	21.1	2.6	3.9	99.2%	36.1	24.1	6.2	8.1
Ours	99.6%	**25.5**	**20.4**	**2.1**	**3.4**	**99.3%**	**32.2**	**22.7**	**4.4**	**6.3**

Table 2. VR tracking results on AMASS train and test.

	AMASS-Train					AMASS-Test				
Method	Succ ↑	$E_{\text{g-mpjpe}}$ ↓	E_{mpjpe} ↓	E_{acc} ↓	E_{vel} ↓	Succ ↑	$E_{\text{g-mpjpe}}$ ↓	E_{mpjpe} ↓	E_{acc} ↓	E_{vel} ↓
ASE	18.6%	128.7	87.9	40.9	33.3	8.0%	114.3	99.2	57.7	44.0
ASE-PMP	7.2%	159.7	155.7	142.2	123.0	1.5%	161.7	126.1	151.1	96.4
PULSE	**99.5%**	57.8	51.0	3.9	7.1	**93.4%**	88.6	67.1	9.1	14.9
Ours	99.3%	**52.9**	**47.9**	**3.7**	**6.4**	**93.4%**	**83.2**	**65.7**	**8.8**	**13.4**

4.2. Skill Embedding Downstream Tasks

As shown in Fig. 6, the modular skills learned by our framework can be effectively applied to downstream tasks. In Tab. 3, we record the normalized return for the downstream tasks, steering, reach and strike, with 0 being the minimum possible return value, and 1 being the maximum. Compared to ASE baselines, our method creates a more expressive skill space, leading to superior normalized returns. Our approach matches PULSE in performance, but, unlike PULSE, our policy network does not require additional distillation from a motion tracking policy to obtain effective skill embeddings. Instead, the learned modular skills can be directly applied to a variety of downstream tasks while preserving accurate motion imitation capabilities. In contrast, PULSE suffers from a significant decrease in motion tracking accuracy compared to the motion tracking policy, PHC+, used for distillation (see Tab. 1). We further compare the diversity of motion sequences produced by PULSE and ModSkill for each task using the APD score. Specifically, we calculate the mean and standard deviation of this metric over 10 iterations, each sampling 10,000 motion sequences per task. Tab. 4 shows that ModSkill achieves higher motion diversity across the three generative tasks, demonstrating the effectiveness of modular skills.

4.3. Skill Interpolation and Composition

In Fig. 7, we present the t-SNE visualization of body-part skill embeddings for motions from AMASS-Test. We uniformly sample skill embeddings for each body part every 10 frames, resulting in ∼3000 samples. For clarity, we label a

Table 3. Normalized returns for steering, reach, and strike.

	Steering	Reach	Strike
ASE	$0.60^{\pm 0.001}$	$0.10^{\pm 0.003}$	$0.12^{\pm 0.006}$
ASE-PMP	$0.31^{\pm 0.002}$	$0.06^{\pm 0.002}$	$0.13^{\pm 0.004}$
PULSE	$0.92^{\pm 0.002}$	$0.77^{\pm 0.002}$	$\mathbf{0.88}^{\pm 0.002}$
ModSkill	$\mathbf{0.93}^{\pm 0.001}$	$\mathbf{0.79}^{\pm 0.002}$	$\mathbf{0.88}^{\pm 0.003}$

Table 4. APD of motions produced for steering, reach, and strike.

	Steering	Reach	Strike
PULSE	$123.74^{\pm 0.13}$	$87.97^{\pm 0.11}$	$88.94^{\pm 0.28}$
ModSkill	$\mathbf{128.13}^{\pm 0.14}$	$\mathbf{116.12}^{\pm 1.60}$	$\mathbf{94.86}^{\pm 0.94}$

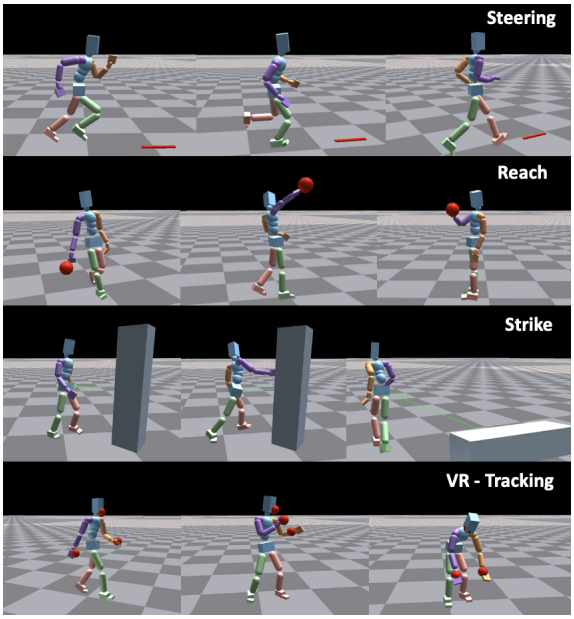

Figure 6. Our modular skill embeddings are flexible and informative, achieving natural human-like behavior in downstream tasks.

subset of samples for five different types of motions. We observe consistent structures for the same motion across body parts, with symmetric structures emerging for motions with alternating patterns, e.g., walking and jumping. We further demonstrate the structure of modular skill embeddings through interpolation. In Fig. 8, we begin with the sequence of skill embeddings that controls the simulated character to cross both arms in front of its chest, and then gradually interpolate the left-hand skill embeddings towards those corresponding to a left-hand waving motion, while keeping the sequence of skill embeddings for all other body parts unchanged. The resulting motion exhibits a smooth transition, and further analysis of the left hand's linear velocity in the local coordinate frame over simulated time steps confirms this smoothness. Furthermore, due to the modularity of the controllers, our framework allows for flexible integration of new modular components. For example, as shown in Fig. 9, we can combine articulated hand controllers with our pre-trained body part controllers to control a humanoid with dexterous hands to imitate a motion sequence without retraining the entire framework from scratch.

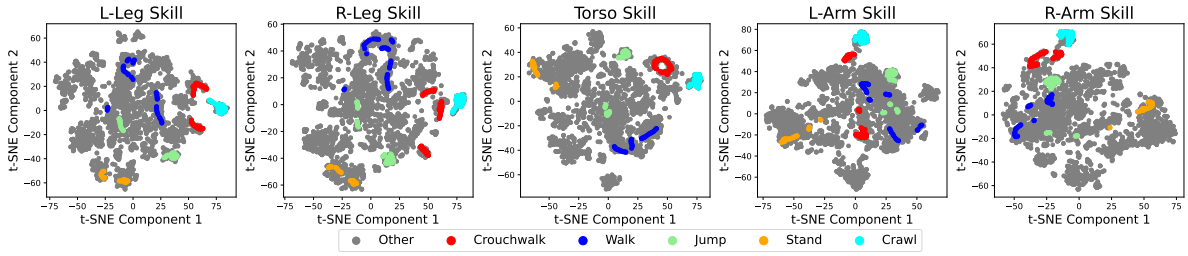

Figure 7. Modular Skill Embedding t-SNE Visualization on AMASS-Test.

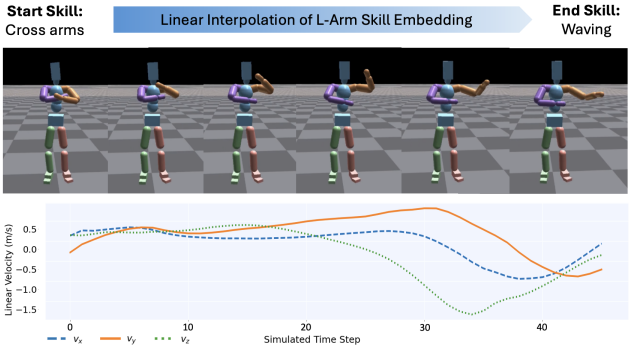

Figure 8. **Skill Interpolation**: Interpolating left-hand skill embeddings from "crossing arms" to "waving" shows a smooth transition in motion (top row) and linear velocity in the local coordinate frame (bottom row) for the left-hand of the humanoid.

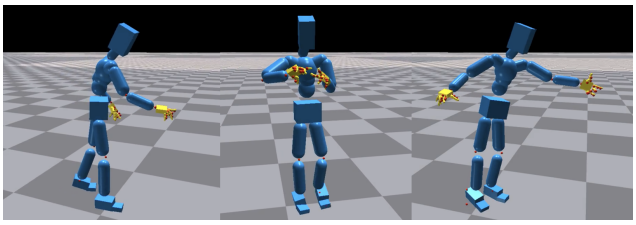

Figure 9. ModSkill enables flexible integration of additional modular controllers without needing to retrain from scratch. By combining articulated hand controllers (highlighted in yellow) with our pre-trained body part controllers, we can control a humanoid with dexterous hands. Red spheres represent target joint locations.

4.4. Ablations

We ablate the effectiveness of each component within our framework with respect to the full-body motion tracking task. In Tab. 5, we present motion tracking results on the AMASS test set with respect to the following configurations: with/without modular controllers for individual body parts, with/without a skill modularization attention mechanism to extract modular skill embeddings, and with/without generative adaptive sampling. We observe that incorporating all components yields the best performance. Notably, even without the attention mechanism, our model outperforms PHC+ in motion tracking on unseen motion sequences, highlighting the effectiveness of modular motor skills. The addition of the attention mechanism not only improves performance but also enables the formulation of reusable modular motor skills for applications to downstream tasks. Furthermore, generative adaptive sampling enhances policy performance on unseen motions.

Table 5. **Ablation on components of ModSkill.** We evaluate on AMASS-Test for the full-body motion tracking task to demonstrate the effectiveness of each component. Modular: whether to use low-level controllers for each body part, Attention: whether to use a skill modularization attention layer. Generative: whether to use generative adaptive sampling.

Modular	Attention	Generative	Succ ↑	$E_{\text{g-mpjpe}}$ ↓	E_{mpjpe} ↓	E_{acc} ↓	E_{vel} ↓
×	×	×	96.4%	41.1	28.4	5.4	7.2
×	✓	×	98.5%	37.4	27.2	5.1	7.1
✓	×	×	98.6%	35.9	23.8	4.4	6.5
✓	✓	×	99.3%	32.4	23.2	4.5	**6.3**
✓	✓	✓	**99.3%**	**32.2**	**22.7**	**4.4**	**6.3**

5. Conclusion

In this paper, we introduce a novel modularized skill learning framework, ModSkill, which decouples complex full-body skills into compositional, modular skills for independent body parts. The framework incorporates a Skill Modularization Attention layer that transforms policy observations into modular skill embeddings, guiding independent low-level controllers for each body part. Additionally, our Active Skill Learning approach with Generative Adaptive Sampling utilizes large motion generation models to adaptively enhance policy learning in challenging tracking scenarios. Our results demonstrate that this modularized framework, enhanced by generative sampling, outperforms existing methods in achieving precise full-body motion tracking and enables reusable skill embeddings for diverse, goal-driven tasks.

Limitations. While modularized skill learning offers advantages in creating compact, compositional skill spaces, it introduces an inherent tradeoff in search complexity due to the formulation of multiple modular components. Future work could explore methods for improving the efficiency of modular skill composition. Additionally, incorporating environment-aware interaction priors and strategies for sim-to-real transfer would be promising directions for extending the framework.

References

[1] Jinseok Bae, Jungdam Won, Donggeun Lim, Cheol-Hui Min, and Young Min Kim. Pmp: Learning to physically interact with environments using part-wise motion priors. In *ACM SIGGRAPH 2023 Conference Proceedings*, New York, NY, USA, 2023. Association for Computing Machinery. 3, 6

[2] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737, 2023. 3

[3] Werner Callebaut and Diego Rasskin-Gutman. *Modularity: understanding the development and evolution of natural complex systems*. MIT press, 2005. 2

[4] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 3

[5] Jieming Cui, Tengyu Liu, Nian Liu, Yaodong Yang, Yixin Zhu, and Siyuan Huang. Anyskill: Learning open-vocabulary physical skill for interactive agents. In *The First Workshop on Populating Empty Cities – Virtual Humans for Robotics and Autonomous Driving at CVPR 2024, 2nd Round*, 2024. 2

[6] Zhiyang Dou, Xuelin Chen, Qingnan Fan, Taku Komura, and Wenping Wang. C·ase: Learning conditional adversarial skill embeddings for physics-based characters. *SIGGRAPH Asia 2023 Conference Papers*, 2023. 2, 3, 5, 6

[7] Petros Faloutsos, Michiel Van de Panne, and Demetri Terzopoulos. Composable controllers for physics-based character animation. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 251–260, 2001. 2

[8] Nate Gillman, Michael Freeman, Daksh Aggarwal, Chia-Hong Hsu, Calvin Luo, Yonglong Tian, and Chen Sun. Self-correcting self-consuming loops for generative model training. In *Proceedings of the 41st International Conference on Machine Learning*, pages 15646–15677. PMLR, 2024. 3

[9] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, 2022. 3, 5

[10] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic scene-aware motion prediction. In *Proceedings of the International Conference on Computer Vision 2021*, 2021. 6

[11] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *ACM SIGGRAPH 2023 Conference Proceedings*, New York, NY, USA, 2023. Association for Computing Machinery. 2

[12] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris M. Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *8th Annual Conference on Robot Learning*, 2024. 1

[13] Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. 2024. 1

[14] Arend Hintze and Christoph Adami. Evolution of complex modular biological networks. *PLoS computational biology*, 4(2):e23, 2008. 2

[15] Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. Como: Controllable motion generation through language guided pose code editing. In *European Conference on Computer Vision*. Springer, Cham, 2024. 3

[16] Deok-Kyeong Jang, Soomin Park, and Sung-Hee Lee. Motion puzzle: Arbitrary motion style transfer by body part. *ACM Transactions on Graphics*, 41(3):1–16, 2022. 2, 3

[17] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[18] Jordan Juravsky, Yunrong Guo, Sanja Fidler, and Xue Bin Peng. Padl: Language-directed physics-based character control. In *SIGGRAPH Asia 2022 Conference Papers (SA '22 Conference Papers)*, 2022. 2, 3

[19] Jordan Juravsky, Yunrong Guo, Sanja Fidler, and Xue Bin Peng. Superpadl: Scaling language-directed physics-based control with progressive supervised distillation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 3

[20] N. Khoshsiyar, R. Gou, T. Zhou, S. Andrews, and M. van de Panne. Partwisempc: Interactive control of contact-guided motions. *Computer Graphics Forum*, n/a(n/a):e15174. 3

[21] Hiroaki Kitano. Computational systems biology. *Nature*, 420(6912):206–210, 2002. 2

[22] Seyoung Lee, Jiye Lee, and Jehee Lee. Learning virtual chimeras by dynamic motion reassembly. *ACM Trans. Graph.*, 41(6), 2022. 2, 3

[23] Sergey Levine and Jovan Popovic. Physically Plausible Simulation for Character Animation. In *Eurographics/ ACM SIGGRAPH Symposium on Computer Animation*. The Eurographics Association, 2012. 1

[24] Chenhao Li, Elijah Stanger-Jones, Steve Heim, and Sangbae Kim. Fld: Fourier latent dynamics for structured motion representation and learning. *arXiv preprint arXiv:2402.13820*, 2024. 2

[25] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel van de Panne. Character controllers using motion vaes. *ACM Trans. Graph.*, 39(4), 2020. 3

[26] C Karen Liu, Aaron Hertzmann, and Zoran Popović. Learning physics-based motion style with nonlinear inverse optimization. *ACM Transactions on Graphics (TOG)*, 24(3): 1071–1081, 2005. 2

[27] Libin Liu and Jessica Hodgins. Learning to schedule control fragments for physics-based characters using deep q-learning. *ACM Transactions on Graphics (TOG)*, 36(3):1–14, 2017.

[28] Libin Liu and Jessica Hodgins. Learning basketball dribbling skills using trajectory optimization and deep reinforcement learning. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018.

[29] Libin Liu, KangKang Yin, Michiel Van de Panne, Tianjia Shao, and Weiwei Xu. Sampling-based contact-rich motion control. In *ACM SIGGRAPH 2010 papers*, pages 1–10. 2010. 2

[30] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 4

[31] Qiujing Lu, Yipeng Zhang, Mingjian Lu, and Vwani Roychowdhury. Action-conditioned on-demand motion generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 6

[32] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. In *Advances in Neural Information Processing Systems*, 2021. 6

[33] Zhengyi Luo, Shun Iwase, Ye Yuan, and Kris Kitani. Embodied scene-aware human pose estimation. In *Advances in Neural Information Processing Systems*, 2022. 4

[34] Zhengyi Luo, Ye Yuan, and Kris M. Kitani. From universal humanoid control to automatic physically valid character creation. *ArXiv*, abs/2206.09286, 2022. 2

[35] Zhengyi Luo, Jinkun Cao, Alexander W. Winkler, Kris Kitani, and Weipeng Xu. Perpetual humanoid control for real-time simulated avatars. In *International Conference on Computer Vision (ICCV)*, 2023. 2, 4, 5, 6

[36] Zhengyi Luo, Jinkun Cao, Rawal Khirodkar, Alexander Winkler, Kris Kitani, and Weipeng Xu. Real-time simulated avatar from head-mounted sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 571–581, 2024. 1

[37] Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander Winkler, Jing Huang, Kris M. Kitani, and Weipeng Xu. Universal humanoid motion representations for physics-based control. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3, 5, 6

[38] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019. 2

[39] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, N. Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac gym: High performance gpu-based physics simulation for robot learning. *ArXiv*, abs/2108.10470, 2021. 6

[40] Josh Merel, Leonard Hasenclever, Alexandre Galashov, Arun Ahuja, Vu Pham, Greg Wayne, Yee Whye Teh, and Nicolas Heess. Neural probabilistic motor primitives for humanoid control. In *International Conference on Learning Representations*, 2019. 3

[41] Liang Pan, Zeshi Yang, Zhiyang Dou, Wenjia Wang, Buzhen Huang, Bo Dai, Taku Komura, and Jingbo Wang. Tokenhsi: Unified synthesis of physical human-scene interactions through task tokenization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5379–5391, 2025. 2

[42] Xue Bin Peng, Glen Berseth, Kangkang Yin, and Michiel Van De Panne. Deeploco: dynamic locomotion skills using hierarchical deep reinforcement learning. *ACM Trans. Graph.*, 36(4), 2017. 3

[43] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Trans. Graph.*, 37(4):143:1–143:14, 2018.

[44] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. Sfv: reinforcement learning of physical skills from videos. *ACM Trans. Graph.*, 37(6), 2018. 2

[45] Xue Bin Peng, Michael Chang, Grace Zhang, Pieter Abbeel, and Sergey Levine. *MCP: learning composable hierarchical control with multiplicative compositional policies*. Curran Associates Inc., Red Hook, NY, USA, 2019. 2

[46] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: adversarial motion priors for stylized physics-based character control. *ACM Trans. Graph.*, 40(4), 2021. 2, 4, 5

[47] Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Trans. Graph.*, 41(4), 2022. 2, 3, 4, 6

[48] Jiawei Ren, Mingyuan Zhang, Cunjun Yu, Xiao Ma, Liang Pan, and Ziwei Liu. Insactor: Instruction-driven physics-based characters. *NeurIPS*, 2023. 3

[49] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. 2

[50] Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman. Local motion phases for learning multi-contact character movements. *ACM Trans. Graph.*, 39(4), 2020. 3

[51] Francesca Sylos-Labini, Valentina La Scaleia, Germana Cappellini, Arthur Dewolf, Adele Fabiano, Irina A Solopova, Vito Mondì, Yury Ivanenko, and Francesco Lacquaniti. Complexity of modular neuromuscular control increases and variability decreases during human locomotor development. *Communications Biology*, 5(1):1256, 2022. 2

[52] Chen Tessler, Yoni Kasten, Yunrong Guo, Shie Mannor, Gal Chechik, and Xue Bin Peng. Calm: Conditional adversarial latent models for directable virtual characters. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–9, 2023. 2, 3

[53] Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion. In *ACM Transactions On Graphics (TOG)*. ACM New York, NY, USA, 2024. 2

[54] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3, 5

[55] Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit H. Bermano, and Michiel van de Panne. Closd: Closing the loop between simulation and diffusion for multi-task character control, 2024. 3

[56] Nolan Wagener, Andrey Kolobov, Felipe Vieira Frujeri, Ricky Loynd, Ching-An Cheng, and Matthew Hausknecht. MoCapAct: A multi-task dataset for simulated humanoid control. In *Advances in Neural Information Processing Systems*, pages 35418–35431, 2022. 2, 3

[57] Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Tlcontrol: Trajectory and language control for human motion synthesis. In *European Conference on Computer Vision*. Springer, Cham, 2024. 3

[58] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20460–20469, 2022. 6

[59] Jingbo Wang, Zhengyi Luo, Ye Yuan, Yixuan Li, and Bo Dai. Pacer+: On-demand pedestrian animation controller in driving scenarios. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 6

[60] Wenjia Wang, Liang Pan, Zhiyang Dou, Jidong Mei, Zhouyingcheng Liao, Yuke Lou, Yifan Wu, Lei Yang, Jingbo Wang, and Taku Komura. Sims: Simulating stylized human-scene interactions with retrieval-augmented script generation. In *ICCV*, 2025. 2

[61] Yinhuai Wang, Qihan Zhao, Runyi Yu, Ailing Zeng, Jing Lin, Zhengyi Luo, Hok Wai Tsui, Jiwen Yu, Xiu Li, Qifeng Chen, Jian Zhang, Lei Zhang, and Ping Tan. Skillmimic: Learning reusable basketball skills from demonstrations, 2024. 2

[62] Alexander Winkler, Jungdam Won, and Yuting Ye. Questsim: Human motion tracking from sparse sensors with simulated avatars. In *SIGGRAPH Asia 2022 Conference Papers*, New York, NY, USA, 2022. Association for Computing Machinery. 1, 4

[63] Jungdam Won, Deepak Gopinath, and Jessica Hodgins. A scalable approach to control diverse behaviors for physically simulated characters. *ACM Trans. Graph.*, 39(4), 2020. 2

[64] Jungdam Won, Deepak Gopinath, and Jessica Hodgins. Control strategies for physically simulated characters performing two-player competitive sports. *ACM Trans. Graph.*, 40(4), 2021. 2

[65] Jungdam Won, Deepak Gopinath, and Jessica Hodgins. Physics-based character controllers using conditional vaes. *ACM Trans. Graph.*, 41(4), 2022. 3

[66] Xianghui Xie, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Template free reconstruction of human-object interaction with procedural interaction generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[67] Pei Xu, Xiumin Shang, Victor Zordan, and Ioannis Karamouzas. Composite motion learning with task control. *ACM Transactions on Graphics*, 42(4), 2023. 3

[68] Ye Yuan and Kris Kitani. Residual force control for agile human behavior imitation and extended motion synthesis. In *Advances in Neural Information Processing Systems*, 2020. 4

[69] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[70] Haotian Zhang, Ye Yuan, Viktor Makoviychuk, Yunrong Guo, Sanja Fidler, Xue Bin Peng, and Kayvon Fatahalian. Learning physically simulated tennis skills from broadcast videos. *ACM Trans. Graph.*, 42(4), 2023. 2

[71] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[72] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. Finemogen: Fine-grained spatio-temporal motion generation and editing. *NeurIPS*, 2023. 3

[73] Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In *European Conference on Computer Vision*, pages 18–38. Springer, Cham, 2024. 3

[74] Qingxu Zhu, He Zhang, Mengting Lan, and Lei Han. Neural categorical priors for physics-based character control. *ACM Trans. Graph.*, 42(6), 2023. 3