TeCH: Temporal Distance Modeling via Contrastive Representation Learning for Humanoid Whole-Body Control

Author Names Omitted for Anonymous Review. Paper-ID 18

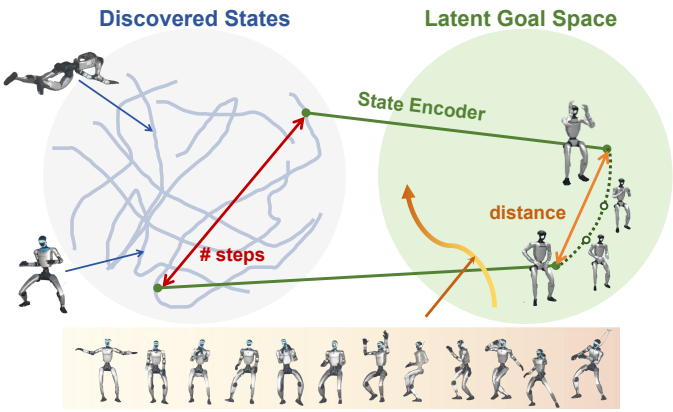

Fig. 1. Overviewv of **TeCH**: Self-explored robot states are embedded via a state encoder into a latent manifold, with latent distance serving as a proxy for transition step count; the learned representation enables both continuous motion tracking and discontinuous goal reaching.

***Abstract*—Humanoid robots are ideal general AI platforms owing to their environmental adaptability, while their high degrees of freedom complicate scalable motor learning. Traditional on-policy reinforcement learning methods such as PPO rely on handcrafted rewards, suffering from prohibitive training costs, limited state coverage, and poor generalization. To address these drawbacks, unsupervised representation learning has been widely adopted for humanoid control. Existing solutions mainly include Forward-Backward (FB) representation learning and contrastive metric learning-based TemporaL Distance Representation (TLDR). Limited by the linear MDP assumption, FB fails in complex contact-rich scenarios and yields unstable representations. This paper presents a novel unsupervised humanoid control method TeCH built upon the contrastive learning framework of TLDR. Our method constructs dense rewards via latent temporal distance modeling to guide structured exploration, delivering a concise, stable training pipeline and well-organized latent space. It enables high-precision zero-shot motion tracking and goal reaching with prominent data efficiency, and can be seamlessly deployed on physical robots. Extensive simulated and real-world experiments validate the superior tracking accuracy and generalization of our method, demonstrating its great potential for practical humanoid control.**

I. Introduction

Humanoid robots serve as ideal physical platforms for general AI due to their natural compatibility with human environments. Their high degrees of freedom support flexible whole-body loco-manipulation, yet complicate scalable motor learning. Traditional methods using handcrafted rewards perform well only on limited tasks and lack robustness and generalization in complex scenarios. To address this, researchers shift toward learning general behavioral priors to obtain reusable low-level motor skills, instead of optimizing for separate tasks.

In recent years, reinforcement learning has advanced sim-to-real motor skill transfer for humanoid robots. Centered on whole-body motion tracking, these methods enable robots to learn dynamic, reusable skills for locomotion and loco-manipulation tasks. However, most prevailing approaches adopt on-policy optimizers such as PPO with manually designed tracking or task-specific rewards, yielding two key drawbacks. First, they demand massive high-dimensional motion data and large-scale parallel environments, leading to prohibitive training costs. For example, SONIC [23] requires 128 GPUs for three consecutive days of training. Second, constrained dataset coverage fails to span the full control space of high-degree-of-freedom humanoids. Lacking structured and smooth representations, existing methods limit skill generalization and reuse, and cannot efficiently generate robust responses against external disturbances.

To tackle these limitations, representation learning-based unsupervised reinforcement learning methods have emerged for humanoid control. A representative branch adopts successor feature-based value decomposition, exemplified by Forward-Backward (FB) representation learning [35], which learns composable latent representations for unsupervised skill discovery and behavior reuse. With high sample efficiency, strong reusability, and compatibility with large-scale replay buffers, off-policy unsupervised reinforcement learning is promising for constructing humanoid behavioral foundation models. Introduced to humanoid control by BFM-Zero [18], FB learning acquires smooth, composable motion primitives via structured exploration, delivering favorable skill interpolation, robustness, and generalization by modeling behavioral manifolds. Nevertheless, FB suffers from critical limitations in high-dimensional humanoid systems. It builds on the linear Markov Decision Process (linear MDP) assumption in latent embedding space, which fails under complex contact dynamics, causing representation instability and degradation. Overall, existing unsupervised skill learning methods inevitably trade off representation expressiveness, training stability, and task scalability.

Another branch of unsupervised reinforcement learning leverages contrastive metric learning to learn state representations. It models temporal correlations and behavioral distances to produce dense, stable optimization signals. As a representative paradigm learned via contrastive learning, TemporaL Distance Representation (TLDR) [1] differs from FB methods:

it abandons value function decomposition and directly learns temporal reachability and system dynamics using temporal distance objectives.

On this basis, this paper proposes **TeCH** (as shown in Figure 1), an unsupervised control framework for humanoid robots built upon the temporal distance representation [1]. This method constructs dense progress rewards using distance variations in the latent space, which endows autonomous exploration with clearer directionality and naturally establishes correspondences between states and goals. Benefiting from more straightforward optimization objectives and a concise, stable training pipeline, **TeCH** delivers better scalability and optimization stability in global goal control tasks. We also observe that **TeCH** can construct well-structured latent spaces to support zero-shot motion tracking and goal reaching with enhanced precision. It achieves exceptional data efficiency during training and demonstrates promising performance when deployed on physical humanoid robots.As the two mainstream unsupervised reinforcement learning paradigms, both FB and TLDR frameworks have been proven capable of transferring learned capabilities to real-world humanoid robots. The main contributions of this paper are summarized as follows:

- **Methodological Innovation**: We present a novel method **TeCH** for humanoid control based on temporal distance learning. It produces dense latent rewards for structured exploration and improves tracking accuracy in unsupervised reinforcement learning.
- **Comparative Benchmark**: We conduct the first systematic comparison of unsupervised RL on humanoid robots, which provides practical insights for future design.
- **Experimental Validation**: Extensive simulated and real-robot experiments validate that **TeCH** achieves high-precision whole-body tracking and goal reaching for practical deployment.

II. Related Works

A. Humanoid Whole-body Control

In recent years, learning-based methods have achieved remarkable progress in humanoid whole-body control. Though reinforcement learning can generate complex dynamic behaviors effectively in simulation [28, 21, 22, 34, 33], sim-to-real transfer remains a critical barrier to real-robot deployment. Various strategies including domain randomization and system identification have been proposed, yet most existing works focus on single-task learning for specific behaviors such as walking, running and standing up [29, 30, 3, 32, 39, 15, 27].

Since 2025, multi-task and general-purpose humanoid control have attracted growing attention, aiming to build a single policy for diverse tasks [12, 13, 41, 40, 38, 4, 11, 14, 5, 6, 8]. These methods typically train tracking policies in simulation and distill them into unified multi-task policies via CVAE-based latent learning or diffusion models [12, 38, 40, 4, 41, 19]. Nevertheless, they are inherently limited by motion dataset quality and generally adopt on-policy reinforcement learning. Distinct from such paradigms, our method leverages off-policy reinforcement learning to directly learn general policies. It yields a richer, more structured skill space and is free from constraints brought by motion datasets.

B. Unsupervised Reinforcement Learning

Unsupervised reinforcement learning enables learning of diverse, controllable behaviors without task-specific rewards. Mutual-information-based skill discovery methods, including DIAYN [7], VIC [9], and VALOR [16], improve skill distinguishability by maximizing mutual information between latent variables and state trajectories. However, these methods are primarily validated in simple locomotion settings and typically adopt on-policy training, limiting their scalability to high-dimensional humanoid systems.

Common contrastive and representation learning approaches, such as CIC [20] and Proto-RL [37], learn structured state and skill embeddings via InfoNCE and prototype matching for more stable representations. Meanwhile, intrinsic motivation methods including RND [2], ICM [26], and SMM [17] promote exploration by rewarding state novelty and coverage. Nevertheless, these paradigms lack explicit goal-progress modeling and deliver weak supervision for goal-reaching behaviors.

Forward–Backward (FB) representation learning [36] offers structured behavior modeling by decomposing value functions into forward state embeddings and backward skill embeddings, supporting composable skill representation and zero-shot generalization. Despite these merits, FB relies on strict successor-feature decomposition assumptions and requires joint optimization of dual mappings, with insufficient validation on high-dimensional, contact-rich humanoid control.

As another effective contrastive learning paradigm, TemporaL Distance Representation (TLDR) [1] abandons forward-backward factorization. It explicitly models temporal reachability via a temporal-distance-aware encoder ϕ_ψ, which embeds temporally adjacent states closely and produces dense progress-aligned rewards r_{tldr}. This design yields a simpler, more stable off-policy optimization objective. In this work, we extend TLDR to real humanoid robots and conduct systematic evaluations on general motion tracking and goal-reaching tasks.

III. Pre-training of TeCH

TeCH comprise two core components: an unsupervised pre-training phase and a zero-shot inference pipeline (Section IV). In this chapter, we outline their unsupervised pre-training workflow, covering the training procedure in simulation environments and the design specifics for transfer to real humanoid robots. We provide the problem formulation in Section III-A, and introduce the unsupervised reinforcement learning framework of the **TeCH** in Section III-B and Section III-C which is based on the temporal distance representation framework.

For **TeCH**, the learning objectives are twofold: 1) a latent representation space that captures the temporal structure of behavior trajectories. In this space, states or trajectories with temporal proximity and similar behavior patterns are mapped

to adjacent positions, empowering the policy to generalize based on trajectory-level behavior semantics (Section III-B); 2) a policy conditioned on latent task representations. We can map macroscopic robot poses into this structured space to achieve unified control over diverse motion targets and behavior patterns (Section III-C).

To achieve zero-shot sim-to-real transfer of the policy, we follow [18] introduce an adversarial objective during training to align humanoid robot behaviors with the human data distribution (Section III-D); we also add auxiliary training objectives to enforce critical constraints for sim-to-real transfer, such as joint limit compliance and bounded action change rates. The full pre-training pipeline and pseudocode of **TeCH** are further elaborated in Section III-E.

A. Problem Formulation

We formulate the real-world humanoid robot control problem as a Partially Observable Markov Decision Process (POMDP), defined as a five-tuple (S, O, A, P, γ), where S denotes the full state space, O the observation space, A the action space, $P(s_{t+1} \mid s_t, a_t)$ the state transition dynamics, and $\gamma \in (0, 1)$ the discount factor.

For a humanoid robot with 29 degrees of freedom, the action $a \in A \subset \mathbb{R}^{29}$ consists of target values for a proportional-derivative (PD) controller over all degrees of freedom. The privileged state information $s \in \mathbb{R}^{463}$ includes root height, body pose, body orientation, linear velocity, and angular velocity. The observable state is defined as $o_t = \{q_t - \bar{q}, \dot{q}_t, \omega_t^{\text{root}}/4, g_t\} \in \mathbb{R}^{64}$, where $q_t \in \mathbb{R}^{29}$ denotes joint positions normalized with respect to the nominal pose $\bar{q}$, $\dot{q}_t \in \mathbb{R}^{29}$ denotes joint velocities, $\omega_t^{\text{root}} \in \mathbb{R}^3$ denotes root angular velocity, and $g_t \in \mathbb{R}^3$ denotes the projected gravity vector in the root frame. We further define the observation history as $o_{t,H} = \{o_{t-H}, a_{t-H}, \ldots, o_t\} \in \mathbb{R}^{93 \cdot H + 64}$, which consists of proprioceptive states and actions. Except for root height, all state components are normalized with respect to the current heading direction and root position. During the pretraining stage, the agent has access to an unlabeled motion dataset $\mathcal{M} = \{\tau\}$, where each trajectory contains both observation and privileged state sequences, i.e.,$\tau = (o_1, s_1, \ldots, o_{l(\tau)}, s_{l(\tau)})$, where $l(\tau)$ denotes the length of trajectory τ.

B. Temporal Distance Representation Learning

Formally, given a state transition (s_t, s_{t+1}), we construct pseudo-goals via temporal rolling:

$$g_t = \text{roll}(s_{t+1}), \tag{1}$$

This temporal rolling operation extends future states beyond single-step transitions, moving beyond immediate adjacency and modeling longer-range temporal dependencies. As a result, the learned representation is optimized under a contrastive learning paradigm to capture multi-step environmental dynamics, rather than merely fitting instantaneous state transitions. We then map all observed states and constructed pseudo-goals into a unified latent space via a shared encoder:

$$\phi_x = \phi_\psi(s_t), \quad \phi_y = \phi_\psi(s_{t+1}), \quad \phi_g = \phi_\psi(g_t). \tag{2}$$

The encoder is trained with a dual-objective contrastive loss to regularize latent space geometry:

$$\mathcal{L}_{\text{TE}}(\psi, \lambda) = -\mathcal{J}_{\text{dist}}(\phi_x, \phi_g) + \lambda\, \mathcal{C}_{\text{step}}(\phi_x, \phi_y), \tag{3}$$

where the core term $\mathcal{J}_{\text{dist}}$ serves as the contrastive objective that enforces explicit feature separation between temporally distant state pairs in the latent space, distinguishing reachable and unreachable temporal state relationships. Complementarily, the step-wise smoothness term $\mathcal{C}_{\text{step}}$ acts as a regularization constraint between consecutive states, preventing abrupt feature shifts and maintaining continuous latent evolution. Jointly optimized, this dual contrastive objective shapes a physically consistent latent geometry, where latent-space distances explicitly reflect the temporal reachability and multi-step transition logic of real robot trajectories.

C. Goal Progress Reward with Temporal Distance

For the policy π_θ, the key objective is to measure progress toward a latent goal z. Let $\bar{z}(\cdot)$ denote the projected latent representation. The reward is defined as:

$$r_{\text{tldr}}(s_t, s_{t+1}, z) = \|z - \bar{z}(s_t)\|_2 - \|z - \bar{z}(s_{t+1})\|_2. \tag{4}$$

This formulation has a clear intuition: if the next state is closer to the goal in the latent space, the reward is **positive**; otherwise, it is negative. Consequently, the policy is continuously driven to optimize toward the direction of goal proximity, resulting in a dense and directionally informative learning signal.

D. Transfer *TeCH* to Real Robots

Deploying unsupervised reinforcement learning on humanoid robots faces issues such as limited real-world observations, sim-to-real dynamics gaps, unstable motions and unnatural behaviors from unregulated exploration. To resolve these problems and enhance sim-to-real transfer, we adopt four training components. We apply asymmetric training to handle partial observability, and leverage domain randomization to reduce simulation overfitting. Auxiliary objectives are introduced to ensure motion stability and safety, while a latent-conditioned style discriminator regularizes policies to generate human-like behaviors. The detailed implementation of all above modules follows BFM-Zero [18].

E. Pre-training Pipeline

As illustrated in Figure **??**, the pretraining pipeline consists of three iterative modules: *online interaction*, *experience replay*, and *multi-objective optimization*. These components form a closed training loop that continuously operates across parallel simulation environments.

During online interaction, the policy executes actions conditioned on a latent variable z and collects transitions of the form $(s_t, o_{t-k:t}, a_t, s_{t+1}, z_t)$, which are stored in an online replay buffer. Importantly, the latent variable z is not drawn

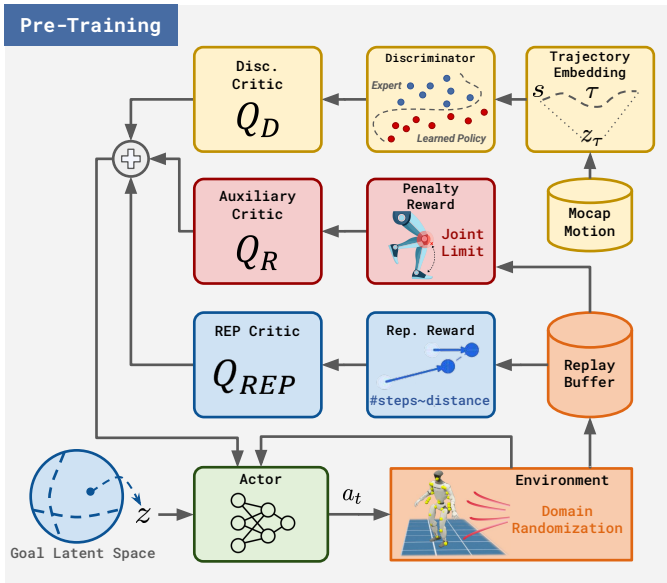

Fig. 2. This pre-training pipeline employs three parallel critics to jointly optimize the policy: a discriminative critic for humanoid motion regularization, an auxiliary critic providing sim-to-real transfer reward, and a representation critic that regularizes state embeddings to align with latent goal.

from a single source but is instead constructed from a mixture of three complementary signals, jointly balancing exploration diversity, goal reachability, and behavioral priors:

1) Randomly sampled latent variables (e.g., from Gaussian or hyperspherical priors), which expand the coverage of the latent task space and encourage behavioral diversity;
2) Latent representations associated with online samples, which align learning with goals that are currently reachable by the policy;
3) Latent representations derived from expert trajectories, which inject human-motion priors and provide stable anchors in the latent space.

Notably, **TeCH** enables a more flexible exploration strategy by leveraging its learned temporal distance metric. In addition to the mixed-goal sampling scheme adopted by [18]—including random goals, replayed high-value goals, and expert goals—**TeCH** further estimates the temporal reachability between the current state and candidate goal states. This reachability estimate is then used to dynamically adjust goal-sampling probabilities throughout training.

Specifically, during the early stages of training, **TeCH** assigns higher sampling probabilities to temporally distant and infrequently visited goals, thereby encouraging broader exploration. As the policy gradually stabilizes, the sampling distribution shifts toward regions associated with higher expected returns, allowing learning to focus on refining task-relevant skills and improving overall policy performance.

During the **experience replay** stage, the learner simultaneously samples transition batches from the online replay buffer and trajectory segments from unlabeled expert demonstrations. These samples are used to construct the discriminator learning signal and the representation learning objective, respectively.

The training then proceeds to the **multi-objective optimization** stage, where parameters are updated sequentially according to Discriminator $\rightarrow$ Encoder(ϕ_ψ) $\rightarrow$ Main Critic and Auxiliary Critic $\rightarrow$ Actor $\rightarrow$ Target Network Soft Update. This closed-loop procedure is repeated until convergence.

From an optimization perspective, the Actor is driven by three complementary learning signals: the representation objective $\boldsymbol{Q_{REP}}$ (induced by r_{tldr}), the style objective $\boldsymbol{Q_D}$, and the auxiliary objective $\boldsymbol{Q_R}$. Combining these terms yields the final Actor objective:

$$\mathcal{L}(\pi) = -\mathbb{E}_{\substack{(o_{t,H},s_t)\sim\mathcal{D} \\ a_t=\pi(o_{t,H},z),\, z\sim\nu}} \Big[\lambda_{REP}\boldsymbol{Q_{REP}} + \lambda_D \boldsymbol{Q_D} + \lambda_R \boldsymbol{Q_R}\Big]. \tag{5}$$

The complete pretraining procedure of **TeCH** is summarized in Algorithm **??**. For clarity, implementation details related to parallel environment simulation and distributed training are omitted.

IV. Zero-shot Inference in TeCH

During the testing stage, **TeCH** can solve a variety of tasks in a *zero-shot* manner, without any additional task-specific learning, planning, or fine-tuning. We first embed different downstream tasks into the latent space, and consider two representative types of tasks.

- **Goal Reaching**: Given discontinuous goal states, we map them into multiple policy conditions z, enabling the robot to switch between different target poses.
- **Trajectory Tracking**: We aggregate future states within a reference trajectory window and map them into a sequence of latent conditions $\{z_t\}$, allowing the policy to progressively track the target motion.

TeCH does not belong to the successor-feature-based unsupervised reinforcement learning paradigm and therefore does not provide a unified value-based inference interface. As a result, it typically cannot directly support reward optimization tasks such as those in [18]. In contrast, **TeCH** introduces temporal-distance-aware representations that explicitly encode the temporal reachability relationship with respect to the current state. Consequently, in zero-shot tracking scenarios, **TeCH** often exhibits faster and tighter trajectory tracking behavior.

Goal Reaching. For the goal reaching task, the corresponding latent vector is formulated as:

$$z_g = \phi(s_g), \tag{6}$$

where $\phi(\cdot)$ denotes the temporal distance encoder of **TeCH**.

Tracking. For motion trajectory tracking, given a trajectory $\tau = \{s_1, \ldots, s_n\}$, the sequence of latent variables $\{z_t\}$ for the policy is defined as:

$$z_t = \sum_{t'=t}^{t+H} w_{t'}\, \phi(s_{t'}), \tag{7}$$

where H represents the look-ahead time window, and $w_{t'}$ are discount weights satisfying $w_{t'} \geq 0$ and $\sum_{t'=t}^{t+H} w_{t'} = 1$.

V. Experiments

A. Experimental Setup

In this section, we evaluate **TeCH** comprehensively in both simulation and real-world robotic platforms. We train both methods in the Unitree G1 humanoid simulation environment [43] based on IsaacLab [25]. The simulation frequency is set to 200 Hz, while the control frequency is 50 Hz. For the motion dataset, we use the LAFAN1 dataset [10], which has been retargeted to the Unitree G1 robot and contains approximately 40 motion clips of several minutes each.

B. Simulation Results

1) Motion Tracking: This evaluation measures the imitation accuracy of motion sequences. We evaluate the policies on both the training dataset (LAFAN1 [10], 40 clips, approximately 2–3 hours in total) and the test dataset (100-Style [24], 3–4k clips, approximately 18–19 hours in total), where all motions are retargeted to the G1 robot. We use joint position error as the evaluation metric:

$$E_{\text{mae}}(e, m) = \frac{1}{|e|} \sum_{t=1}^{|e|} \|q_t(e) - q_t(m)\|, \tag{8}$$

where m denotes the target motion trajectory. The final results are reported as the average error over all motion clips.

Unlike prior frameworks such as Sonic that rely on task-specific motion tracking rewards and imitation learning, BFM-Zero and **TeCH** explore unsupervised reinforcement learning for humanoid control, learning a unified latent space of reusable skills that enables discontinuous goal reaching and hierarchical planning.

Nevertheless, we compare against the state-of-the-art general-purpose motion tracking controller SONIC [23]. We evaluate the publicly released model under the same environment and test sets, reporting E_{mae} and the standard deviation across motion clips. For fairness, BFM-Zero, **TeCH**, and SONIC [23] are evaluated without early termination, i.e., metrics are computed over full trajectories. Since SONIC typically struggles to recover quickly after severe failures or falls, we additionally report SONIC (TER.), where metrics are computed only over segments before termination.

TABLE I
Motion tracking error on training and test datasets for **TeCH** and baselines(mean $\pm$ std).

Method	Motion Tracking Error (E_{mae}) (↓)	
	LAFAN1	100-Style
SONIC[b]	0.2916 ± 0.1887	0.1522 ± 0.1049
SONIC (TER.)[a]	0.1081 ± 0.0213	0.1355 ± 0.0351
BFM-Zero[b]	0.1510 ± 0.0255	0.1674 ± 0.0459
TeCH[b]	0.1318 ± 0.0329	0.1474 ± 0.0405

[a] SONIC (TER.) is evaluated only on segments before termination.
[b] SONIC, BFM-Zero, and **TeCH** are evaluated on full trajectories.

From the tracking results, we observe that SONIC [23] performs poorly in recovery after falls and during ground transitions, which significantly degrades its overall tracking performance on the test set. Under fully out-of-distribution conditions (e.g., external perturbations, falls, or random initializations), BFM-Zero and **TeCH** are generally able to recover balance more naturally and continue stable tracking, whereas SONIC often fails to return to a stable motion regime. We attribute this advantage primarily to off-policy unsupervised exploration. Compared to methods relying on large-scale motion imitation supervision, unsupervised RL frameworks like **TeCH** cover a much broader range of non-standard states and dynamic transitions during training, leading to a smoother and more continuous skill space.

Even in terms of tracking accuracy alone, **TeCH** achieve performance comparable to SONIC (TER.), as shown in Table I. Notably, SONIC already outperforms GMT [4], Any2Track [42], and BeyondMimic [19] in its original evaluation. We also emphasize the significant difference in training efficiency: SONIC [23] uses 128 GPUs, large-scale motion datasets, and massive parallel environments for training. In contrast, our method is trained with a single GPU and only several hours of LAFAN1 data, reducing GPU-hours and environment samples by nearly two orders of magnitude.

For a more granular comparison of the two unsupervised RL methods, we present qualitative motion tracking results in Figure 3. **TeCH** achieves highly accurate motion tracking, with generated trajectories closely aligned with reference motions at the joint level, indicating that the learned representation effectively captures local pose structure. Moreover, **TeCH** consistently outperforms BFM-Zero, particularly in mitigating global root rotation drift, as further supported by the quantitative results in Figure 4, especially in fast-rotation scenarios. This performance gap stems from limitations of BFM-Zero, which learns representations via indirect forward-backward factorization, introducing approximation errors in the successor measure M and constraining its ability to model global orientation, long-horizon phase information, and cumulative root drift. In contrast, **TeCH** formulates tracking objectives more directly, resulting in improved global pose consistency while preserving precise local joint accuracy.

2) Goal Reaching: We manually extract 21 "stable" poses from the training dataset (i.e., LAFAN1), defined as states with zero velocity. To evaluate the proximity between the agent and the target pose, we define the joint error as:

$$E_{\text{mae}}(e, g) = \frac{1}{|e|} \sum_{t=1}^{|e|} \|q_t(e) - q(g)\|, \tag{9}$$

where e denotes an episode and $q(\cdot)$ represents the joint configuration (i.e., a 29-dimensional vector). The final results are reported as the average over all target poses. Each episode has a fixed horizon of $H = 500$.

Figure 5(a) compares the joint-level mean absolute error between BFM-Zero and **TeCH** on the goal-reaching task. The results show that both methods achieve comparable accuracy, while **TeCH** reaches the target more efficiently with a slight degradation in precision. Figure 5(b) shows three randomly

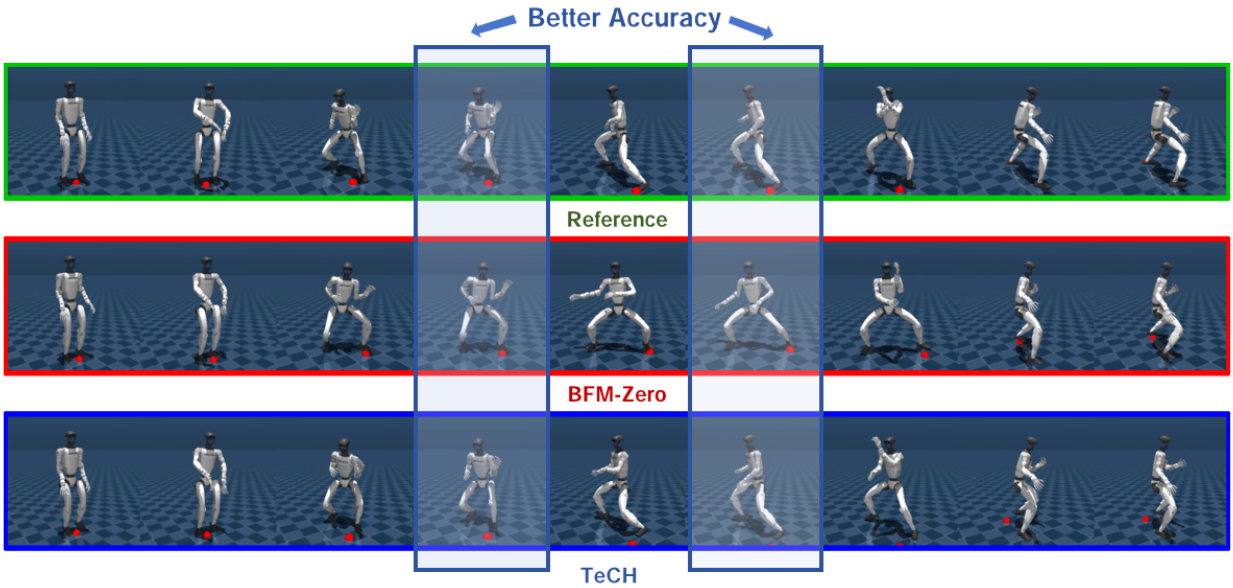

Fig. 3. Qualitative comparison of motion tracking between BFM-Zero and **TeCH** in simulation: We can see both methods can effectively track in domain motions, and **TeCH** is slightly better.

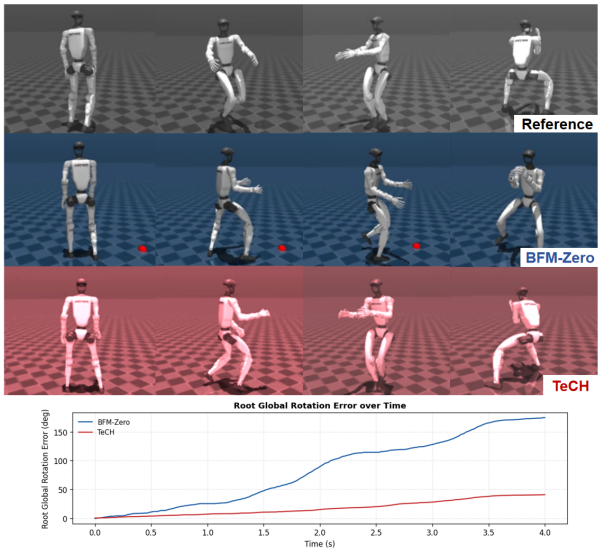

Fig. 4. Comparison of Root Global Rotation Error between BFM-Zero and **TeCH** under Out-of-Distribution Fast 360-Degree Rotation Task.

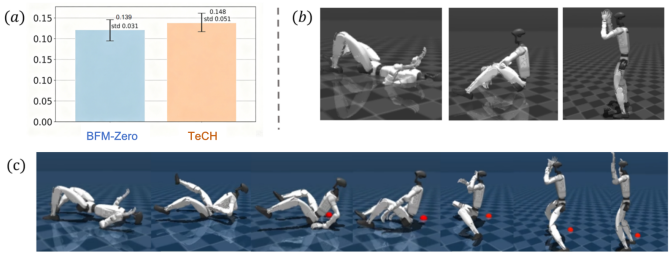

Fig. 5. Comparison of joint error and qualitative results for goal reaching: (a) average joint MAE over 21 target poses; (b) example target poses; (c) qualitative results of **TeCH** on the goal reaching task.

selected target poses, illustrating their diversity and complexity. Figure 5(c) presents qualitative results, where the robot successfully transitions from lying down to sitting and then to standing, demonstrating robust recovery behaviors.

Notably, **TeCH** reaches target goals produces coherent trajectories even in challenging recovery scenarios. In contrast, tracking-based methods such as SONIC [23] struggle with discontinuous goal transitions, highlighting the limitation of purely tracking-driven objectives for discontinuous goal-reaching.

C. Real-World Results of *TeCH*

As shown in Figure 6 and Figure 7, **TeCH** also demonstrates strong performance in both motion tracking and goal reaching when deployed on real hardware. When large external pushes perturb the robot, **TeCH** can autonomously stand back up after a fall just like BFM-Zero, whereas the baseline SONIC [23] cannot recover from falling states. This is due to our unified latent goal space encoding full-body recovery motions during pre-training, allowing the policy to generate self-righting trajectories without separate fall-recovery fine-tuning. However, the two URL methods exhibit distinct behavioral characteristics in real-world tests. Compared to BFM-Zero, **TeCH** tends to execute target motions more directly and responds more aggressively, sometimes resulting in slightly less smooth behaviors. This is due to our representation loss optimizing for minimal transition steps between latent goals, which prioritizes fast state transfer over motion smoothness; in contrast, BFM-Zero's motion-imitation discriminator loss biases the policy toward slower, gradual movements extracted from mocap datasets.

D. Random Rollout

We further investigate the quality of unconditioned random rollouts sampled from our latent action space $\mathcal{Z}$, as visualized

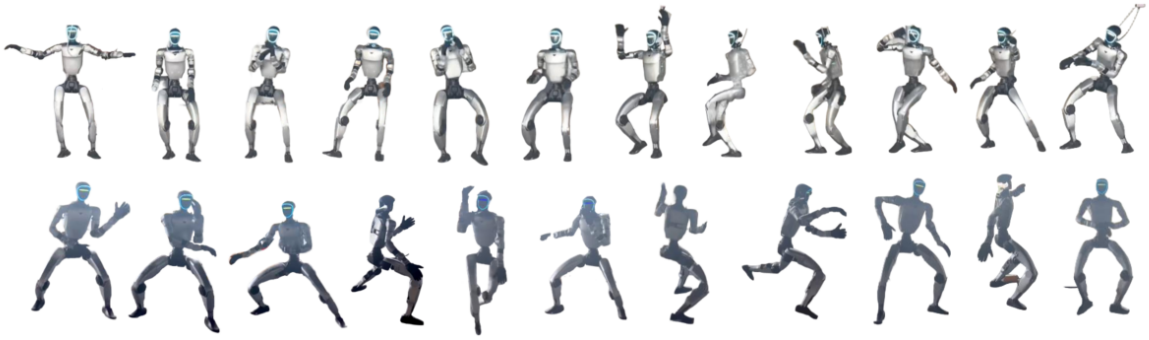

Fig. 6. **TeCH** performs robust motion tracking on high-dynamic motions in real robot tests.

Fig. 7. **TeCH** completes real-world goal reaching tasks with stable movements and smooth behavioral transitions.

in Figure 8. We observe that fully random latent samples yield coherent, human-like humanoid motion sequences without any task guidance or motion priors. The humanoid avatar smoothly transitions across diverse balanced poses, ranging from deep squatting to wide-stance weight shifting and outstretched limb movements, exhibiting natural human-like fluidity and avoiding the jerky joint spasms or floating artifacts prevalent in poorly regularized latent representations.

E. Latent Interpolation

As illustrated by the green dotted trajectory on the right side of Figure 1, linear latent-space interpolation between two target latent embeddings produces smooth, physically consistent humanoid motion sequences. The intermediate robot poses along the interpolation path evolve continuously and gradually without distorted or fragmented unnatural motions. This verifies that our learned latent goal space owns valid continuous structure: straightforward linear interpolation in the embedding space directly maps to smooth sequential state transitions of the robot, supporting both motion tracking and goal-reaching planning.

VI. Ablation Studies

A. Latent Space Dimensionality

The performance of unsupervised behavior foundation models largely depends on the representational capacity of the latent space $Z \subseteq \mathbb{R}^D$: when D is too small, task embeddings and successor features may fail to distinguish different skills; when D is too large, optimization becomes more difficult and sample-inefficient, and training instability may arise due to redundant representations and increased estimation variance.

To analyze the effect of this hyperparameter, we keep all other training settings fixed (network architecture, learning rate, domain randomization, and discriminator reward weights) and evaluate both BFM-Zero and **TeCH** in the Unitree G1 simulation environment. We sweep the latent dimension $D \in \{32, 64, 128, 256, 384, 512, 1024\}$, and use Earth Mover's Distance (EMD) [31] on the LAFAN1 dataset as the metric for action distribution matching (lower is better, indicating closer alignment to the reference motion distribution).

As shown in Figure 9, the EMD of both methods decreases as D increases, indicating improved action distribution fitting, but with diminishing returns. The performance largely saturates when $D \geq 256$. Across all settings, **TeCH** consistently achieves lower EMD than BFM-Zero, suggesting that it can more accurately capture the target motion distribution under the same capacity. When D is sufficiently large, BFM-Zero with successor-feature factorization also achieves comparable performance. Considering the trade-off between performance, computational cost, and training stability, we use $D = 256$ as the default latent dimensionality in all main experiments.

*B. **TeCH** Encoder Learning Rate*

For encoder learning in **TeCH**, a large learning rate may cause instability in the latent geometry, while a small learning rate may lead to sparse progress signals and delayed policy updates. To study its effect on motion distribution fitting, we fix $D = 256$ and all other hyperparameters, and sweep the encoder learning rate $lr \in \{5\times10^{-5}, 3\times10^{-6}, 10^{-5}, 3\times10^{-6}, 10^{-6}, 8\times10^{-7}, 10^{-9}\}$. We report EMD evaluated during training (lower is better).

As shown in Figure 10, a large learning rate ($lr \approx 3\times10^{-5}$) leads to rapid early decrease in EMD but quickly saturates at a higher level with large fluctuations, indicating insufficient refinement of the temporal distance structure. Smaller learning rates (3×10^{-6} and 10^{-6}) yield smoother convergence and better performance, reaching a stable range after around $100,\mathrm{M}$ steps, but remain suboptimal. The best performance

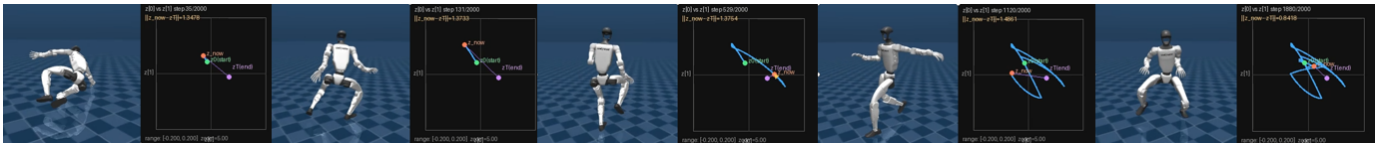

Fig. 8. Random latent samples produce natural, artifact-Free humanoid motion sequences without task supervision or motion prior

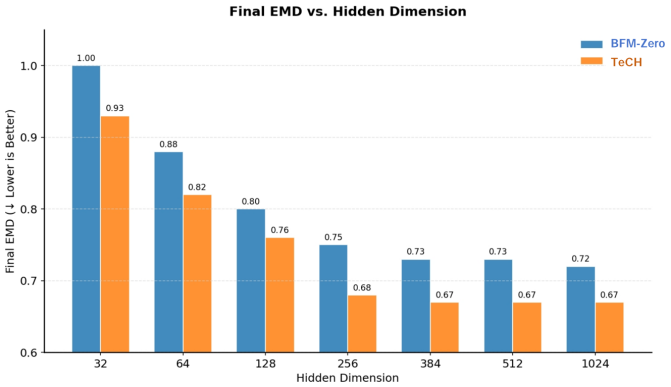

Fig. 9. EMD comparison of BFM-Zero and **TeCH** under different latent dimensions D.

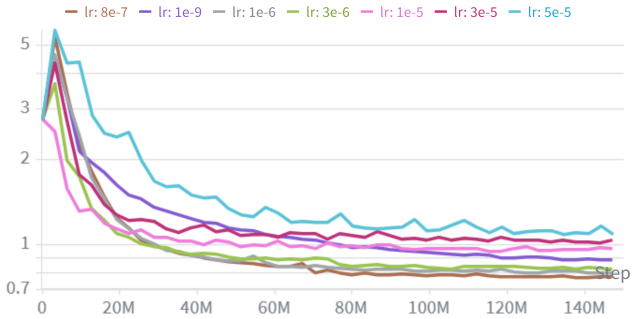

Fig. 10. **TeCH** encoder learning rate ablation: EMD over training steps under different lr values.

is achieved at $lr = 8{\times}10^{-7}$, which attains the lowest EMD within the standard training budget. A further reduction to 10^{-9} improves performance slowly but significantly slows convergence. Overall, there is a clear trade-off between convergence speed and final performance, and we use $lr = 8{\times}10^{-7}$ as the default setting in all experiments.

C. Online Latent Goal Update Frequency

In addition to the encoder learning rate, **TeCH** also relies on periodically refreshing the latent goal z during online interaction (corresponding to the `update-z-every` hyperparameter). This interval determines how many environment steps a sampled latent variable is kept before resampling. A short interval may weaken long-horizon consistency, while a long interval may reduce skill diversity and overly concentrate replay data in latent space. We sweep `update-z-every` $\in \{5, 10, 50, 100, 200\}$ while fixing $D = 256$ and the optimal learning rate above.

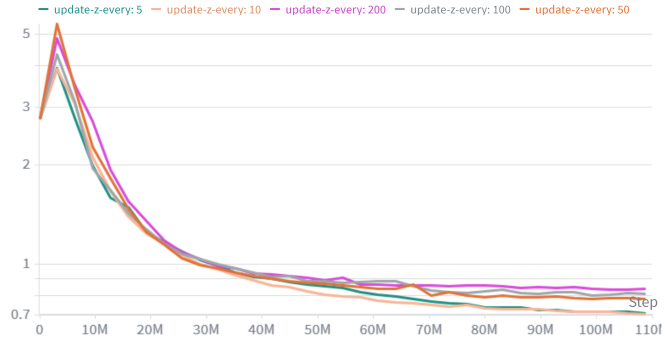

Fig. 11. **TeCH** ablation on online latent goal update frequency: EMD over training steps under different update intervals.

As shown in Figure 11, all curves decrease rapidly in the early stage and begin to diverge after approximately $40\,\mathrm{M}$ steps. Updating z every 5 or 10 steps leads to the fastest convergence and the smoothest curves, achieving the lowest EMD (around 0.70) at $110\,\mathrm{M}$ steps. The setting `update-z-every`$= 50$ performs slightly worse, with a final EMD of approximately 0.78, while 100 and 200 result in 0.81 and 0.83–0.85, respectively. Overall, more frequent z resampling improves action distribution fitting by increasing behavioral coverage in parallel environments while maintaining stable temporal representations. However, the performance gain saturates as the interval increases from 50 to 100 and 200. The difference between `update-z-every`$= 5$ and $= 10$ becomes negligible after $80\,\mathrm{M}$ steps. Combining all ablation results, we use $lr = 8{\times}10^{-7}$ and `update-z-every`$= 10$ as default settings in the main experiments.

VII. Conclusion

In conclusion, this paper presents a novel unsupervised humanoid control method built upon the contrastive learning based TLDR framework. Different from existing FB representation that relies on strict linear MDP assumptions, **TeCH** directly models temporal reachability in latent space to produce stable, dense reward signals and supports effective off-policy training. Combined with a series of practical training strategies including asymmetric training, domain randomization, auxiliary objectives and style regularization, our method well addresses partial observability, sim-to-real gaps and unstable or unnatural motions. Extensive experiments across simulation and physical robots verify that our approach achieves high-precision zero-shot tracking and goal reaching with strong data efficiency and generalization ability, demonstrating promising potential for real-world humanoid robot deployment.

References

[1] Junik Bae, Kwanyoung Park, and Youngwoon Lee. Tldr: Unsupervised goal-conditioned rl via temporal distance-aware representations, 2024. URL https://arxiv.org/abs/2407.08464.

[2] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation, 2018. URL https://arxiv.org/abs/1810.12894.

[3] Zixuan Chen, Xialin He, Yen-Jen Wang, Qiayuan Liao, Yanjie Ze, Zhongyu Li, S. Shankar Sastry, Jiajun Wu, Koushil Sreenath, Saurabh Gupta, and Xue Bin Peng. Learning smooth humanoid locomotion through lipschitz-constrained policies. *CoRR*, abs/2410.11825, 2024.

[4] Zixuan Chen, Mazeyu Ji, Xuxin Cheng, Xuanbin Peng, Xue Bin Peng, and Xiaolong Wang. GMT: general motion tracking for humanoid whole-body control. *CoRR*, abs/2506.14770, 2025.

[5] Xuxin Cheng, Yandong Ji, Junming Chen, Ruihan Yang, Ge Yang, and Xiaolong Wang. Expressive whole-body control for humanoid robots. *arXiv preprint arXiv:2402.16796*, 2024.

[6] Pranay Dugar, Aayam Shrestha, Fangzhou Yu, Bart van Marum, and Alan Fern. Learning multi-modal whole-body control for real-world humanoid robots, 2025. URL https://arxiv.org/abs/2408.07295.

[7] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function, 2018. URL https://arxiv.org/abs/1802.06070.

[8] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans, 2024. URL https://arxiv.org/abs/2406.10454.

[9] Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control, 2016. URL https://arxiv.org/abs/1611.07507.

[10] Félix G. Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher J. Pal. Robust motion in-betweening. *ACM Trans. Graph.*, 39(4):60, 2020.

[11] Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. *arXiv preprint arXiv:2403.04436*, 2024.

[12] Tairan He, Wenli Xiao, Toru Lin, Zhengyi Luo, Zhenjia Xu, Zhenyu Jiang, Jan Kautz, Changliu Liu, Guanya Shi, Xiaolong Wang, Linxi Fan, and Yuke Zhu. HOVER: versatile neural whole-body controller for humanoid robots. *CoRR*, abs/2410.21229, 2024.

[13] Tairan He, Jiawei Gao, Wenli Xiao, Yuanhang Zhang, Zi Wang, Jiashun Wang, Zhengyi Luo, Guanqi He, Nikhil Sobanbab, Chaoyi Pan, Zeji Yi, Guannan Qu, Kris Kitani, Jessica K. Hodgins, Linxi Fan, Yuke Zhu, Changliu Liu, and Guanya Shi. ASAP: aligning simulation and real-world physics for learning agile humanoid whole-body skills. *CoRR*, abs/2502.01143, 2025.

[14] 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 *Conference on Robot Learning*, pages 1516–1540. PMLR, 2025.

[15] Xialin He, Runpei Dong, Zixuan Chen, and Saurabh Gupta. Learning getting-up policies for real-world humanoid robots, 2025. URL https://arxiv.org/abs/2502.12152.

[16] Jongmin Kim, Youngwoon Lee, Pieter Abbeel, and Guanya Shi. Variational curriculum reinforcement learning for unsupervised discovery of skills, 2023. URL https://arxiv.org/abs/2303.16342.

[17] Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching, 2020. URL https://arxiv.org/abs/1906.05274.

[18] Yitang Li, Zhengyi Luo, Tonghe Zhang, Cunxi Dai, Anssi Kanervisto, Andrea Tirinzoni, Haoyang Weng, Kris Kitani, Mateusz Guzek, Ahmed Touati, Alessandro Lazaric, Matteo Pirotta, and Guanya Shi. Bfm-zero: A promptable behavioral foundation model for humanoid control using unsupervised reinforcement learning, 2025. URL https://arxiv.org/abs/2511.04131.

[19] Qiayuan Liao, Takara E. Truong, Xiaoyu Huang, Guy Tevet, Koushil Sreenath, and C. Karen Liu. Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion, 2025. URL https://arxiv.org/abs/2508.08241.

[20] Hao Liu and Pieter Abbeel. Cic: Contrastive intrinsic control for unsupervised skill discovery, 2022. URL https://arxiv.org/abs/2202.00161.

[21] Zhengyi Luo, Jinkun Cao, Alexander Winkler, Kris Kitani, and Weipeng Xu. Perpetual humanoid control for real-time simulated avatars. In *ICCV*, pages 10861–10870. IEEE, 2023.

[22] 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 *ICLR*. OpenReview.net, 2024.

[23] Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Sirui Chen, Fernando Castañeda, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, Xingye Da, Runyu Ding, Cyrus Hogg, Lina Song, Edy Lim, Eugene Jeong, Tairan He, Haoru Xue, Wenli Xiao, Zi Wang, Simon Yuen, Jan Kautz, Yan Chang, Umar Iqbal, Linxi "Jim" Fan, and Yuke Zhu. Sonic: Supersizing motion tracking for natural humanoid whole-body control, 2025. URL https://arxiv.org/abs/2511.07820.

[24] Ian Mason. 100STYLE: A motion capture dataset of 100 locomotion styles, 2023. URL https://www.ianxmason.com/100style/.

[25] Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Ritvik Singh,

Yunrong Guo, Hammad Mazhar, Ajay Mandlekar, Buck Babich, Gavriel State, Marco Hutter, and Animesh Garg. Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics Autom. Lett.*, 8(6):3740–3747, 2023.

[26] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction, 2017. URL https://arxiv.org/abs/1705.05363.

[27] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. *CoRR*, abs/1710.06537, 2017. URL http://arxiv.org/abs/1710.06537.

[28] 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, 2018.

[29] Ilija Radosavovic, Tete Xiao, Bike Zhang, Trevor Darrell, Jitendra Malik, and Koushil Sreenath. Real-world humanoid locomotion with reinforcement learning. *Sci. Robotics*, 9(89), 2024.

[30] Ilija Radosavovic, Bike Zhang, Baifeng Shi, Jathushan Rajasegaran, Sarthak Kamat, Trevor Darrell, Koushil Sreenath, and Jitendra Malik. Humanoid locomotion as next token prediction. *CoRR*, abs/2402.19469, 2024.

[31] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[32] Younggyo Seo, Carmelo Sferrazza, Haoran Geng, Michal Nauman, Zhao-Heng Yin, and Pieter Abbeel. Fasttd3: Simple, fast, and capable reinforcement learning for humanoid control. *CoRR*, abs/2505.22642, 2025.

[33] Agon Serifi et al. Vmp: Versatile motion priors for robustly tracking motion on physical characters. *Computer Graphics Forum*, 43(8), 2024. doi: 10.1111/cgf.15175. URL https://doi.org/10.1111/cgf.15175.

[34] Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting. *ACM Trans. Graph.*, 43(6):209:1–209:21, 2024.

[35] Ahmed Touati and Yann Ollivier. Learning one representation to optimize all rewards. In *NeurIPS*, pages 13–23, 2021.

[36] Ahmed Touati and Yann Ollivier. Learning one representation to optimize all rewards, 2021. URL https://arxiv.org/abs/2103.07945.

[37] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Reinforcement learning with prototypical representations, 2021. URL https://arxiv.org/abs/2102.11271.

[38] Kangning Yin, Weishuai Zeng, Ke Fan, Zirui Wang, Qiang Zhang, Zheng Tian, Jingbo Wang, Jiangmiao Pang, and Weinan Zhang. Unitracker: Learning universal whole-body motion tracker for humanoid robots. *CoRR*, abs/2507.07356, 2025.

[39] Kevin Zakka, Baruch Tabanpour, Qiayuan Liao, Mustafa Haiderbhai, Samuel Holt, Jing Yuan Luo, Arthur Allshire, Erik Frey, Koushil Sreenath, Lueder A. Kahrs, Carmelo Sferrazza, Yuval Tassa, and Pieter Abbeel. Mujoco playground. *CoRR*, abs/2502.08844, 2025.

[40] Weishuai Zeng, Shunlin Lu, Kangning Yin, Xiaojie Niu, Minyue Dai, Jingbo Wang, and Jiangmiao Pang. Behavior foundation model for humanoid robots, 2025. URL https://arxiv.org/abs/2509.13780.

[41] Zhikai Zhang, Chao Chen, Han Xue, Jilong Wang, Sikai Liang, Yun Liu, Zongzhang Zhang, He Wang, and Li Yi. Unleashing humanoid reaching potential via real-world-ready skill space. *CoRR*, abs/2505.10918, 2025.

[42] Zhikai Zhang, Jun Guo, Chao Chen, Jilong Wang, Chenghuai Lin, Yunrui Lian, Han Xue, Zhenrong Wang, Maoqi Liu, Jiangran Lyu, Huaping Liu, He Wang, and Li Yi. Track any motions under any disturbances, 2025. URL https://arxiv.org/abs/2509.13833.

[43] . G1, 2025. URL https://www.unitree.com/cn/g1/.