FLD: Fourier Latent Dynamics for Structured Motion Representation and Learning

Chenhao Li[1], Elijah Stanger-Jones[1], Steve Heim[1], Sangbae Kim[1]

***Abstract*— Motion trajectories offer reliable references for physics-based motion learning but suffer from sparsity, particularly in regions that lack sufficient data coverage. To address this challenge, we introduce a self-supervised, structured representation and generation method that extracts spatial-temporal relationships in periodic or quasi-periodic motions. The motion dynamics in a continuously parameterized latent space enable our method to enhance the interpolation and generalization capabilities of motion learning algorithms. The motion learning controller, informed by the motion parameterization, operates online tracking of a wide range of motions, including targets unseen during training. By leveraging the identified spatial-temporal structure, our work opens new possibilities for future advancements in general motion representation and learning algorithms.**

I. INTRODUCTION

The availability of reference trajectories, such as motion capture data, has significantly propelled the advancement of motion learning techniques [1, 2, 3, 4, 5, 6, 7]. However, it is difficult to generalize policies using these techniques to motions outside the distribution of the available data [8, 7]. Instead of handling raw motion trajectories in long-horizon, high-dimensional state space, structured representation methods introduce certain inductive biases during training and offer an efficient approach to managing complex movements [9, 10]. These methods focus on extracting the essential features and temporal dependencies of motions, enabling more effective and compact representations [11, 12]. By uncovering and utilizing the underlying patterns and relationships within the motion space, continuous and rich sets of motions can be produced that progress realistically in a smooth and temporally coherent manner [13, 5, 14].

In this work, we present Fourier Latent Dynamics (FLD), a generative extension to Periodic Autoencoder (PAE) [5] that extracts spatial-temporal relationships in periodic or quasi-periodic motions with a novel predictive structure. FLD efficiently represents high-dimensional trajectories by featuring motion dynamics in a continuously parameterized latent space that accommodates essential features and temporal dependencies of natural motions. The enforcement of latent dynamics empowers FLD to enhance the proficiency and generalization capabilities of motion learning algorithms with accurately described motion transitions and interpolations. The motion learning controllers, informed by the latent parameterization space of FLD, demonstrate extended online tracking capability. Supplementary videos and more details for this work are available at `https://sites.google.com/view/iclr2024-fld/home`.

II. RELATED WORK

In contrast to explicitly defined trajectory parameters, self-supervised models such as autoencoders explain motion evolution in a latent space. These representation methods have shown success in controlling non-linear dynamical systems [15], enabling complex decision-making [16], solving long-horizon tasks [17], and imitating motion sequences [18]. A recent practice attempts to identify motion dynamics in a common latent space to foster temporal consistency between different dynamical systems [19]. To consider the correlation between different body parts, a recent work on PAE constructs a latent space using an autoencoder structure and applies a frequency domain conversion as an inductive bias [5]. The extracted latent parameters have been tested as effective full-body state representations in downstream motion learning tasks [14]. Despite such progress, PAE is restricted to representing local frames and is not fully exploited to express overall motions or predict them.

III. PRELIMINARIES

PAE addresses the challenges of learning the structure of the motion space, such as data sparsity and the highly nonlinear nature of the space, by focusing on the periodicity of motions in the frequency domain. We denote trajectory segments of length H in d-dimensional state space preceding time step t by $\mathbf{s}_t = (s_{t-H+1}, \dots, s_t) \in \mathbb{R}^{d \times H}$, as the input to PAE. The autoencoder structure decomposes the input motions into c latent channels that accommodate lower-dimensional embedding $\mathbf{z}_t \in \mathbb{R}^{c \times H}$ of the motion input. A following differentiable Fast Fourier Transform obtains the frequency f_t, amplitude a_t, and offset b_t vectors of the latent trajectories, while the phase vector ϕ_t is computed with a separate fully connected layer. We refer to the original work [5] for more details.

PAE extracts a multi-dimensional latent space from full-body motion data, effectively clustering motions and creating a manifold in which computed feature distances provide a more meaningful similarity measure compared to the original motion space as visualized in Fig. 3.

IV. APPROACH

A. Problem formulation

We consider the state space $\mathcal{S}$ and define a motion sequence $\tau = (s_0, s_1, \dots)$ drawn from a reference dataset $\mathcal{M}$

[1] Biomimetic Robotics Lab, Department of Mechanical Engineering, Massachusetts Institute of Technology, MA 02139, United States {chenhli, elijahsj, sheim, sangbae}@mit.edu

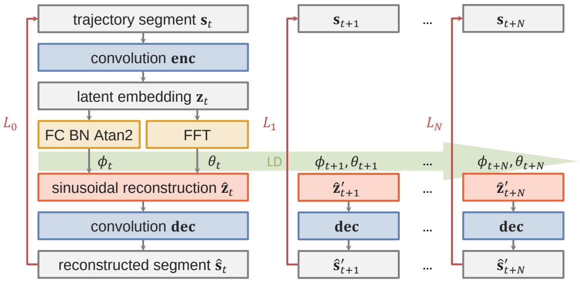
trajectory segment $\mathbf{s}_t$
convolution **enc**
latent embedding $\mathbf{z}_t$
FC BN Atan2
FFT
ϕ_t
θ_t
LD
sinusoidal reconstruction $\hat{\mathbf{z}}_t$
convolution **dec**
reconstructed segment $\hat{\mathbf{s}}_t$
L_0
L_1
L_N
$\mathbf{s}_{t+1}$
$\mathbf{s}_{t+N}$
ϕ_{t+1}, θ_{t+1}
ϕ_{t+N}, θ_{t+N}
$\hat{\mathbf{z}}'_{t+1}$
$\hat{\mathbf{z}}'_{t+N}$
dec
$\hat{\mathbf{s}}'_{t+1}$
$\hat{\mathbf{s}}'_{t+N}$
...

Fig. 1: FLD training pipeline. During training, latent dynamics are enforced to predict proceeding latent states and parameterizations. The prediction loss is computed in the original motion space with respect to the ground truth future states.

as a trajectory of consecutive states $s \in \mathcal{S}$. Our research focuses on creating a physics-based learning controller capable of not only replicating motions prescribed by the reference dataset but also generating motions accordingly in response to novel target inputs, thereby enhancing its generality across a wide range of motions beyond the reference dataset. To this end, we adopt a two-stage training pipeline. In the first stage, an efficient representation model is trained on the reference dataset and a continuously parameterized latent space is obtained where novel motions can be synthesized by sampling the latent encodings. The second stage involves developing an effective learning algorithm that tracks the diverse generated target trajectories.

B. Fourier Latent Dynamics

By inspecting the parameters of the latent trajectories of periodic or quasi-periodic motions encoded by PAE, we observe that the frequency, amplitude, and offset vectors stay nearly time-invariant along the trajectories. We introduce the quasi-constant parameterization assumption.

Assumption 1: A latent trajectory $\mathbf{z} = (\mathbf{z}_t, \mathbf{z}_{t+1}, \dots)$ can be approximated by $\hat{\mathbf{z}} = (\hat{\mathbf{z}}_t, \hat{\mathbf{z}}_{t+1}, \dots)$ with a bounded error $\delta = \|\mathbf{z} - \hat{\mathbf{z}}\|$, where $\hat{\mathbf{z}}_{t'} = \hat{p}(\phi_{t'}, f, a, b)$, $\forall t' \in \{t, t+1, \dots\}$.

Assumption 1 holds with low approximation errors for periodic or quasi-periodic input motion trajectories, which yield constant frequency domain features. Since these latent features are learned, the assumption can be explicitly enforced. In the following context, we denote by ϕ_t the *latent state* and f, a, b the *latent parameterization*. Here, we introduce Fourier Latent Dynamics (FLD), which enforces reconstruction of $\mathbf{z}$ over the complete trajectory by propagating latent dynamics parameterized by a local state ϕ_t and a *constant* set of global parameterization f, a, and b.

We formalize the latent dynamics of FLD and its training process in Fig. 1. For a motion segment $\mathbf{s}_t = (s_{t-H+1}, \dots, s_t)$ whose latent trajectory segment $\mathbf{z}_t$ is parameterized by ϕ_t, f_t, a_t, and b_t, we approximate the proceeding motion segment $\mathbf{s}_{t+i} = (s_{t-H+1+i}, \dots, s_{t+i})$ with the prediction $\hat{\mathbf{s}}'_{t+i}$ decoded from i-step forward propagation $\hat{\mathbf{z}}'_{t+i}$ using the latent dynamics from time step t.

$$\hat{\mathbf{z}}'_{t+i} = \hat{p}(\phi_t + i f_t \Delta t, f_t, a_t, b_t), \quad \hat{\mathbf{s}}'_{t+i} = \mathbf{dec}(\hat{\mathbf{z}}'_{t+i}), \tag{1}$$

where Δt denotes the step time. The latent dynamics in Eq. 1 assumes locally constant latent parameterizations and propagates latent states by advancing i local phase increments. We can compute the prediction loss at time $t+i$. In fact, the local reconstruction process employed by PAE can be viewed as regression on a zero-step forward prediction using the latent dynamics. We can perform regressions on multi-step forward prediction by propagating the latent dynamics and define the total loss for training FLD with the maximum propagation horizon of N and a decay factor α,

$$L^N_{FLD} = \sum_{i=0}^{N} \alpha^i L_i, \quad L_i = \text{MSE}(\hat{\mathbf{s}}'_{t+i}, \mathbf{s}_{t+i}). \tag{2}$$

Training with the FLD loss enforces Assump. 1 in a local range of N steps.

For the following discussions, we consider training FLD on the reference dataset $\mathcal{M}$ and define the latent parameterization space $\Theta \subseteq \mathbb{R}^{3c}$ encompassing the latent frequency, amplitude, and offset. Therefore, each motion trajectory can be exclusively represented by a time-dependent latent state $\phi_t \in \mathbb{R}^c$ that describes the local time indexing and a constant latent parameterization $\theta = (f, a, b) \in \mathbb{R}^{3c}$ that describes the global high-level features of the motion.

C. Motion learning

Given reference trajectories, physics-based motion learning algorithms train a control policy that actuates the joints of the simulated character or robot and reproduces the instructed motion trajectories.

1) Policy training: At the beginning of each episode, a set of latent parameterization $\theta_0 \in \mathbb{R}^{3c}$ is sampled from a skill sampler p_θ. The latent state $\phi_0 \in \mathbb{R}^c$ is uniformly sampled from a fixed range $\mathcal{U}$. The step update of the latent vectors follows the latent dynamics in Eq. 1,

$$\theta_t = \theta_{t-1}, \quad \phi_t = \phi_{t-1} + f_{t-1} \Delta t. \tag{3}$$

At each step, the latent state and the latent parameterization are used to reconstruct a motion segment

$$\hat{\mathbf{s}}_t = (\hat{s}_{t-H+1}, \dots, \hat{s}_t) = \mathbf{dec}(\hat{\mathbf{z}}_t) = \mathbf{dec}(\hat{p}(\phi_t, \theta_t)), \tag{4}$$

whose most recent state $\hat{s}_t$ serves as a tracking target for the learning environment at the current time step.

The latent state and parameterization are provided to the observation space to inform the policy about the motion and the specific frame it should be tracking. Figure 2 provides a schematic overview of the training pipeline.

2) Online tracking: During the inference phase, the policy structure incorporates real-time motion input as tracking targets, irrespective of their periodic or quasi-periodic nature. The latent parameterizations of the intended motion are obtained online using the FLD encoder.

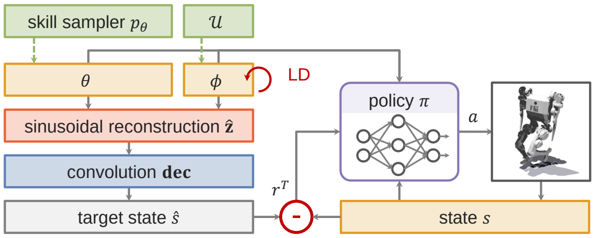

Fig. 2: System overview. During training, the latent states propagate under the latent dynamics and are reconstructed to policy tracking targets $\hat{s}$ at each step. The tracking reward r^T is computed as the distance between the target $\hat{s}$ and the measured states s.

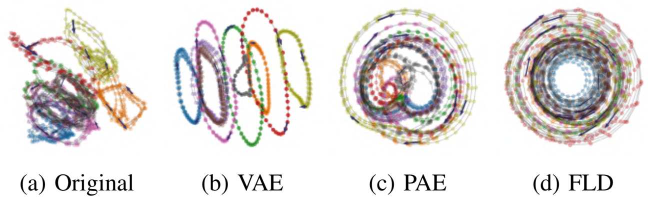
(a) Original (b) VAE (c) PAE (d) FLD

Fig. 3: Latent manifolds for different motions. Each color is associated with a trajectory from a motion type. The arrows denote the state evolution direction. FLD presents the strongest spatial-temporal relationships with explicit latent dynamics enforcement. PAE witnesses a similar but weaker pattern with local sinusoidal reconstruction. In comparison, VAE enables only spatial closeness, and the trajectories of the original states are the least structured.

V. EXPERIMENTS

We evaluate FLD on the MIT Humanoid robot [20], with which we show its applicability to state-of-the-art real-world robotic systems. We use the human locomotion clips collected in [1] retargeted to the joint space of our robot as the reference motion dataset containing slow and fast jog, forward and backward run, slow and fast step in place, left and right turn, and forward stride. Note that the motion labels are not observed during the training of the models and are only used for evaluation. In the motion learning experiments, we use Proximal Policy Optimization (PPO) [21] in Isaac Gym [22].

A. *Structured motion representation*

After the computation of the latent manifold, we project the principal components of the phase features onto a two-dimensional plane, as outlined in [5]. We then compare the latent structure induced by FLD with that by PAE. Additionally, we adopt a Variational Autoencoder (VAE) as a commonly employed method for representing motions in a lower-dimensional space. Lastly, we plot the principal components of the original motion states for comprehensive analysis. We illustrate the latent embeddings acquired by these models in Fig. 3, where each point corresponds to a latent representation of a trajectory segment input.

Notably, FLD demonstrates the most consistent structure akin to concentric cycles, primarily due to the motion-predictive structure within the latent dynamics enforced by Eq. 2. The cycles depicted in the figures represent the primary period of individual motions. The angle around the center (latent state) signifies the timing, while the distance from the center (latent parameterization) represents the high-level features (e.g. velocity, direction, contact frequency, etc.) that remain consistent throughout the trajectory. This pattern reflects the strong temporal regularity captured by Assump. 1, which preserves time-invariant global information regarding the overall motion. As PAE can be viewed as FLD with zero-step latent propagation, in contrast, we observe a weaker pattern in the latent manifold of PAE, where the consistency of high-level features holds only locally. Finally, the reconstruction process employed in VAE training does not impose any specific constraints on the temporal structure of system propagation. Consequently, the resulting latent representation, except for the direct encoding, exhibits the least structured characteristics among the models.

Powered by the latent dynamics, FLD offers a compact representation of high-dimensional motions by employing the time index vector ϕ_t and assuming high-level feature consistency θ throughout each trajectory. Conversely, PAE encodes motion features only locally $\theta_t = \theta(\phi_t)$. The numbers of parameters of different models used to express a trajectory of length $|\tau|$ are listed in Table I.

TABLE I: Motion representation parameters

Original	VAE	PAE	FLD
$d \times \lvert\tau\rvert$	$c \times (\lvert\tau\rvert - H + 1)$	$4c \times (\lvert\tau\rvert - H + 1)$	$4c$

B. *Motion reconstruction and prediction*

We demonstrate the generality of FLD in reconstructing and predicting *unseen* motions during training. Figure 4 (left) illustrates a representative validation with a diagonal run motion. At time $t = 65$, FLD undertakes motion reconstruction and prediction for future state transitions based on the most recent information $\mathbf{s}_t$, as elaborated in Sec. IV-B. For comparison, we train a PAE and a feed-forward (FF) model with fully connected layers with the same input and output structure as FLD.

It is evident that the motion predicted by FLD aligns with the actual trajectories. Particularly in joint position evolution which presents strong sinusoidal periodicity, it exhibits the lowest relative error e. The superiority of FLD is especially pronounced in long-term prediction regions, where the other models accumulate significantly larger compounding errors. The effectiveness of FLD in accurately predicting motion for an extended horizon is attributed to the latent dynamics enforced with an appropriate propagation horizon N in Eq. 2. In the extreme case of $N = 0$ (PAE), the relative error is larger due to the weaker temporal propagation structure. The result on the diagonal run trajectory demonstrates the ability of FLD to accurately predict future states despite not being exposed to this specific motion during training. This showcases the generalization capability of FLD, as it effectively

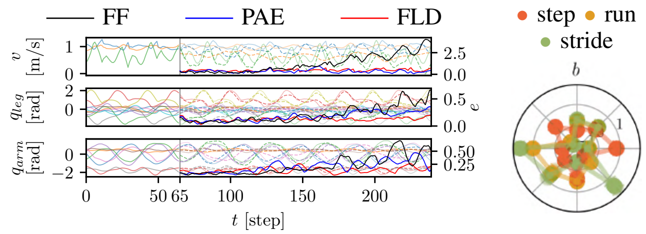

Fig. 4: Motion reconstruction and prediction of a diagonal run trajectory (left). The solid and dashed curves denote the ground truth and predicted state evolution. The relative prediction error (vivid) of FF, PAE and FLD is depicted with the axis indicating e on the right. Latent offset (right) of step in place, forward run, and forward stride. Each radius denotes a latent channel.

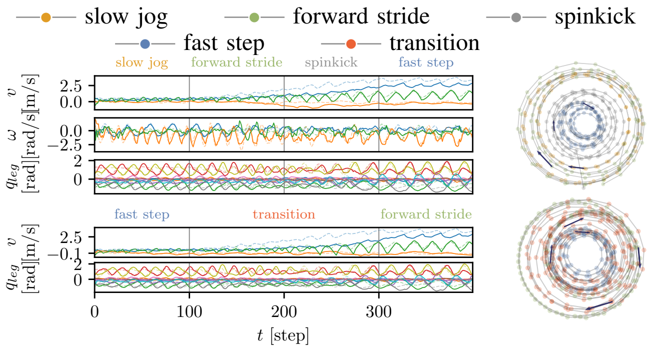

Fig. 5: Motion tracking (top) and motion transition (bottom). The dashed curves denote the user-specified tracking target, and the solid ones denote the measured system states. The corresponding latent manifolds are depicted on the right side.

captures the underlying dynamics and temporal relationships inherent in the training dataset, which are prevalent and can be adapted to unseen motions. In comparison, the FF model training fails to understand the spatial-temporal structure in the motions and results in strong overfitting to the training dataset, thus limiting its generality. Moreover, the dedicated FF model solely propagates the states through autoregression and does not provide any data representation.

With the embedded motion-predictive structure, the enhanced generality achieved by FLD is attributed to the well-shaped latent representation space, where sensible distances between motion patterns are established. Figure 4 (right) depicts the latent offsets of step in place, forward run, and forward stride, where the parameterization of the intermediate motion (forward run) is distributed in between.

C. Motion tracking

Following Sec. IV-C, we can learn a motion tracking controller that employs FLD parameterization space. We perform an online tracking experiment where real-time user input of various motion types is provided to the controller as tracking targets.

In the first example (Fig. 5, top), we switch the input motion to a different type every 100 time steps (indicated by vertical grey lines). We observe that the controller achieves accurate user input tracking, as evidenced by the close alignment between the dictated (dashed) and measured (solid) states, except for the spinkick motion. Moreover, the controller demonstrates the ability to transition between different tracking targets smoothly. By considering the tracking of an arbitrary motion as a process of wandering between continuously parameterized periodic priors, FLD dynamically extracts essential characteristics of local approximates. To further understand the performance of FLD and the learning agent on tracking motions that fall into the gaps between trajectories captured in the reference dataset, we construct in the second example (Fig. 5, bottom) a transition phase where the target motion parameterizations are obtained from linear interpolation between the source and target motions. In particular, the interpolated movements exhibit a gradual evolution of high-level motion features, providing a clear and structured transition from high-frequency, low-velocity stepping to low-frequency, high-velocity striding sequences. This gradual evolution of motion features in the interpolated trajectories suggests that FLD is capable of capturing and preserving the essential temporal and spatial relationships of the underlying motions. It bridges the gap between different motion types and velocities, generating coherent and natural motion sequences that smoothly transition from one to another.

VI. Conclusion

In this work, we present FLD, a novel self-supervised, structured representation and generation method that extracts spatial-temporal relationships in periodic or quasi-periodic motions. FLD efficiently represents high-dimensional trajectories by featuring motion dynamics in a continuously parameterized latent space that accommodates essential features and temporal dependencies of natural motions. Compared with models without explicitly enforced temporal structures, FLD significantly reduces the number of parameters required to express non-linear trajectories and generalizes accurate state transition prediction to unseen motions. The enhanced generality by FLD is further confirmed with the high-level understanding of motion similarity by the latent parameterization space. The motion learning controllers, informed by the latent parameterization space, demonstrate extended online tracking capability. By leveraging the identified spatial-temporal structure, FLD opens up possibilities for future advancements in motion representation and learning algorithms.

References

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

[2] K. Bergamin, S. Clavet, D. Holden, and J. R. Forbes, “Drecon: data-driven responsive control of physics-based characters,” *ACM Transactions On Graphics (TOG)*, vol. 38, no. 6, pp. 1–11, 2019.

[3] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa, "Amp: Adversarial motion priors for stylized physics-based character control," *ACM Transactions on Graphics (ToG)*, vol. 40, no. 4, pp. 1–20, 2021.

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

[5] S. Starke, I. Mason, and T. Komura, "Deepphase: Periodic autoencoders for learning motion phase manifolds," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–13, 2022.

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

[7] C. Li, S. Blaes, P. Kolev, M. Vlastelica, J. Frey, and G. Martius, "Versatile skill control via self-supervised adversarial imitation of unlabeled mixed motions," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2944–2950.

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

[9] J. Min and J. Chai, "Motion graphs++ a compact generative model for semantic motion analysis and synthesis," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 6, pp. 1–12, 2012.

[10] S. Lee, S. Lee, Y. Lee, and J. Lee, "Learning a family of motor skills from a single motion clip," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–13, 2021.

[11] Y. Lee, K. Wampler, G. Bernstein, J. Popović, and Z. Popović, "Motion fields for interactive character locomotion," in *ACM SIGGRAPH Asia 2010 papers*, 2010, pp. 1–8.

[12] S. Levine, J. M. Wang, A. Haraux, Z. Popović, and V. Koltun, "Continuous character control with low-dimensional embeddings," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 4, pp. 1–10, 2012.

[13] S. Starke, Y. Zhao, T. Komura, and K. Zaman, "Local motion phases for learning multi-contact character movements," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 54–1, 2020.

[14] P. Starke, S. Starke, T. Komura, and F. Steinicke, "Motion in-betweening with phase manifolds," *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, vol. 6, no. 3, pp. 1–17, 2023.

[15] M. Watter, J. Springenberg, J. Boedecker, and M. Riedmiller, "Embed to control: A locally linear latent dynamics model for control from raw images," *Advances in neural information processing systems*, vol. 28, 2015.

[16] D. Ha and J. Schmidhuber, "World models," *arXiv preprint arXiv:1803.10122*, 2018.

[17] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," *arXiv preprint arXiv:1912.01603*, 2019.

[18] G. Berseth, F. Golemo, and C. Pal, "Towards learning to imitate from a single video demonstration," *arXiv preprint arXiv:1901.07186*, 2019.

[19] N. H. Kim, Z. Xie, and M. Panne, "Learning to correspond dynamical systems," in *Learning for Dynamics and Control*. PMLR, 2020, pp. 105–117.

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

[21] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[22] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," in *Conference on Robot Learning*. PMLR, 2022, pp. 91–100.