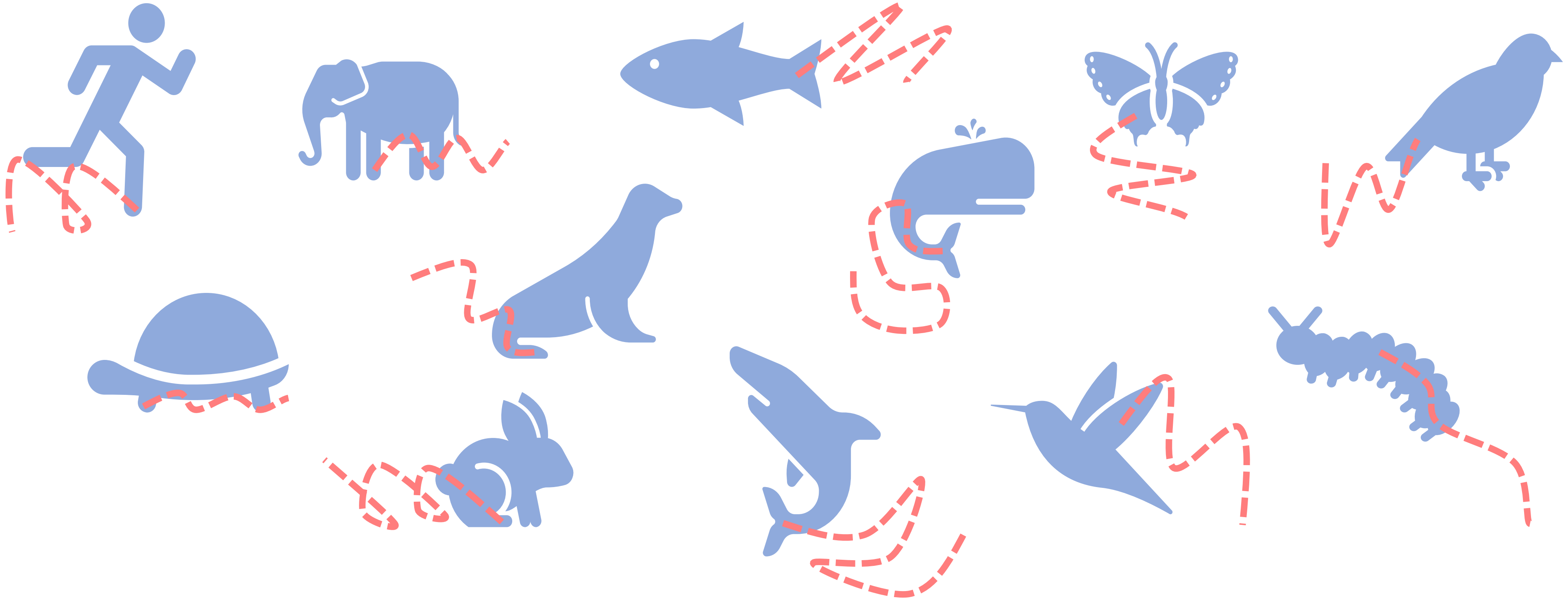
Periodic Skill Discovery

Jonghae Park, Daesol Cho, Jusuk Lee, Dongseok Shim, Inkyu Jang, H. Jin Kim

Lab for Autonomous Robotics Research

Seoul National University

Locomotion in nature : *Inherently* Periodic

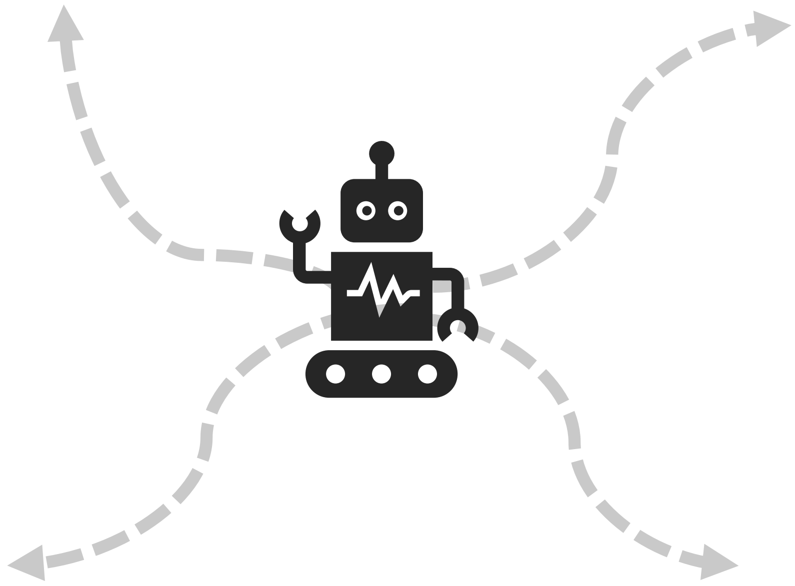


All forms of locomotion skills share a *periodic structure*

Unsupervised Skill Discovery

Unsupervised Skill Discovery

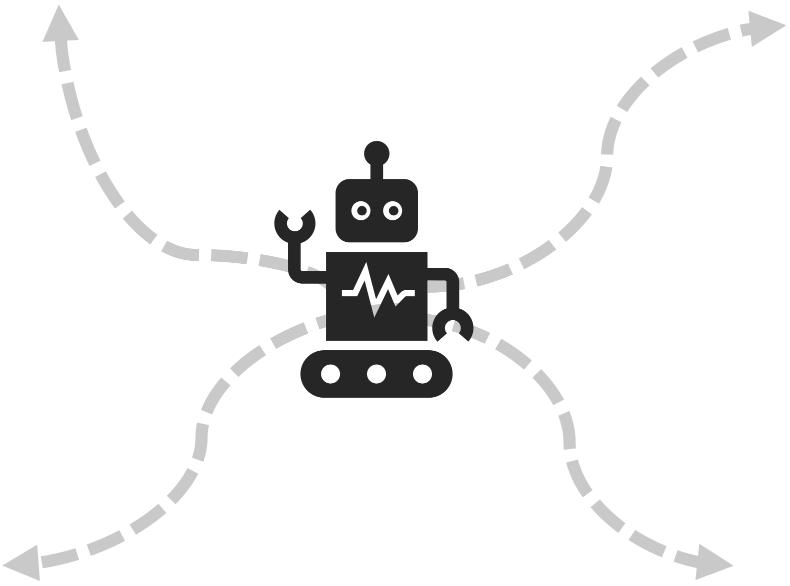
(1) Unsupervised skill learning



Learn useful skills from the environment
without any external rewards

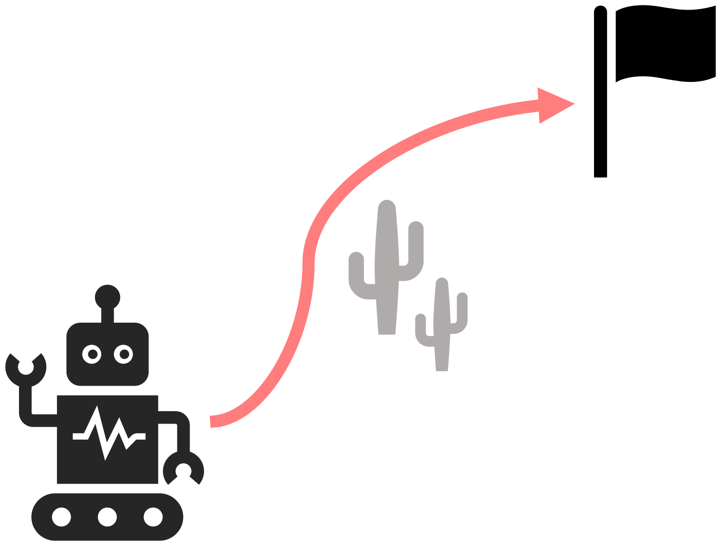
Unsupervised Skill Discovery

(1) Unsupervised skill learning



Learn useful skills from the environment without any external rewards

(2) Solving downstream task efficiently



Leverage the learned skills for finetuning or high-level planning

Prior works often fail to learn *multi-timescale* behaviors

Prior works often fail to learn *multi-timescale* behaviors

(1) Mutual Information (MI) - based skill discovery (e.g., DIAYN, DADS, …)

Prior works often fail to learn *multi-timescale* behaviors

(1) Mutual Information (MI) - based skill discovery (e.g., DIAYN, DADS, …)

$$I(S;Z) = -H(Z\,|\,S) + H(Z) = \mathbb{E}_{z,\tau}[\log p(z\,|\,s)] - \mathbb{E}_z[\log p(z)]$$
$$\geq \mathbb{E}_{z,\tau}[\log q_\theta(z\,|\,s)] + (\text{constant}) \simeq \mathbb{E}_{z,\tau}\left[-\frac{1}{2\sigma^2}\|z - \mu_\theta(s)\|_2^2\right] + (\text{constant})$$

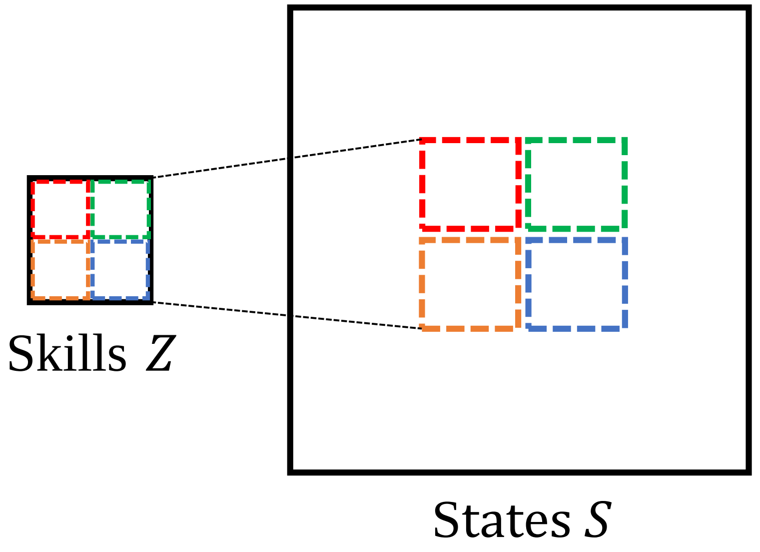
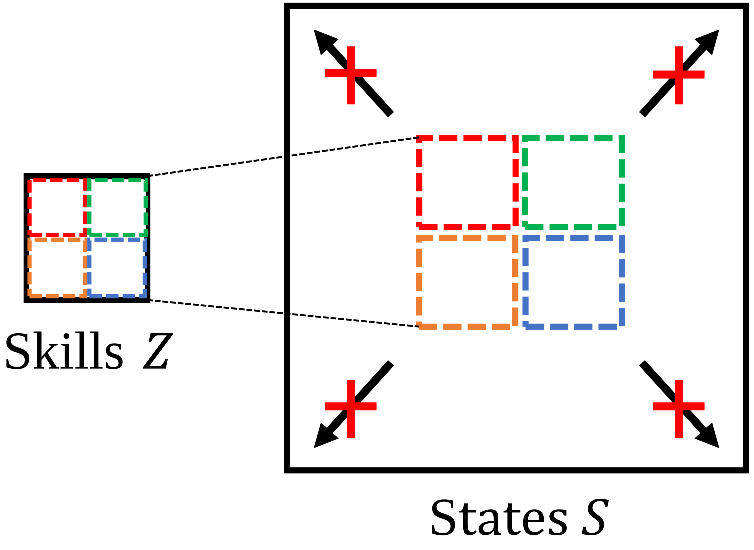


Figure adapted from: Park *et al.*, METRA: Scalable Unsupervised RL with Metric-Aware Abstraction (2023)

Prior works often fail to learn *multi-timescale* behaviors

(1) Mutual Information (MI) - based skill discovery (e.g., DIAYN, DADS, …)

$$
\begin{aligned}
I(S;Z) &= -H(Z \mid S) + H(Z) = \mathbb{E}_{z,\tau}[\log p(z \mid s)] - \mathbb{E}_z[\log p(z)] \\
&\geq \mathbb{E}_{z,\tau}[\log q_\theta(z \mid s)] + (\text{constant}) \simeq \mathbb{E}_{z,\tau}\left[-\frac{1}{2\sigma^2}\|z - \mu_\theta(s)\|_2^2\right] + (\text{constant})
\end{aligned}
$$



"No additional motivation for exploration"

"Do not consider temporal aspects of skills"

Prior works often fail to learn *multi-timescale* behaviors

(2) Distance - maximizing skill discovery (e.g., LSD, CSD, METRA, …)

Prior works often fail to learn *multi-timescale* behaviors

(2) Distance - maximizing skill discovery (e.g., LSD, CSD, METRA, …)

$$\mathcal{J}_{\text{DSD}} := \mathbb{E}_{(z,\tau)\sim\mathcal{D}}\left[(\phi(s_{t+1}) - \phi(s_t))^\top z\right] \quad \text{s.t.} \quad \|\phi(x) - \phi(y)\| \leq d(x,y) \quad \forall x, y \in \mathcal{D}$$

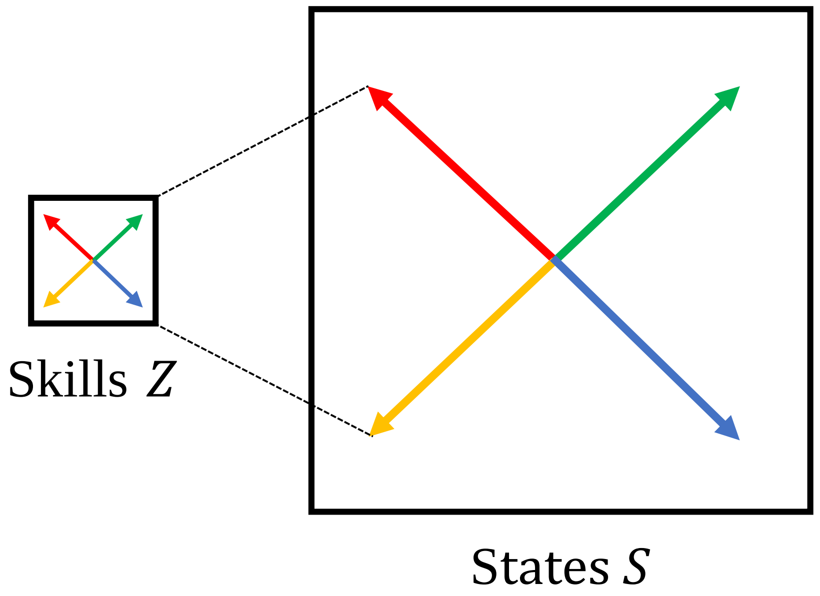
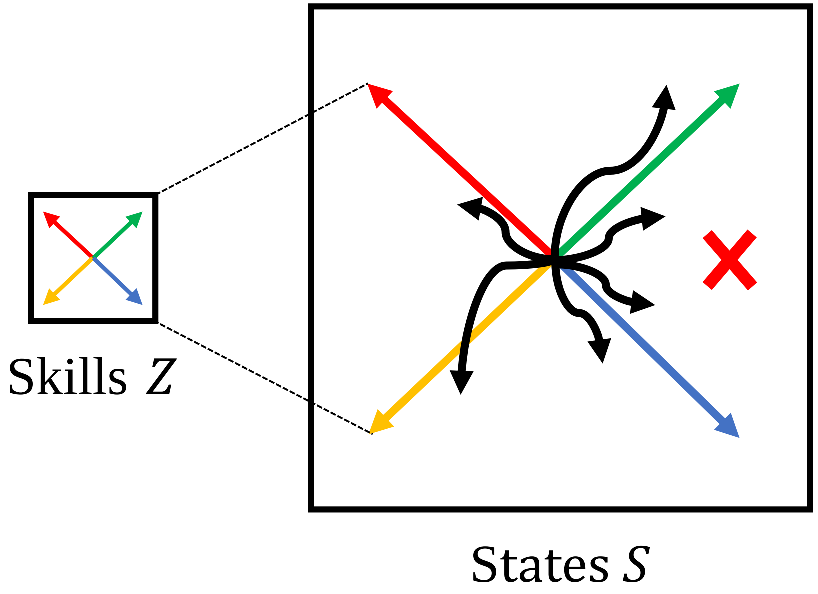


Figure adapted from: Park *et al.*, METRA: Scalable Unsupervised RL with Metric-Aware Abstraction (2023)

Prior works often fail to learn *multi-timescale* behaviors

(2) Distance - maximizing skill discovery (e.g., LSD, CSD, METRA, …)

$$\mathcal{J}_{\text{DSD}} := \mathbb{E}_{(z,\tau)\sim\mathcal{D}}\left[(\phi(s_{t+1}) - \phi(s_t))^\top z\right] \quad \text{s.t.} \quad \|\phi(x) - \phi(y)\| \leq d(x,y) \quad \forall x, y \in \mathcal{D}$$



“Prefers *hard-to-achieve* behaviors”

“Little incentive to adjust the temporal patterns of skills”

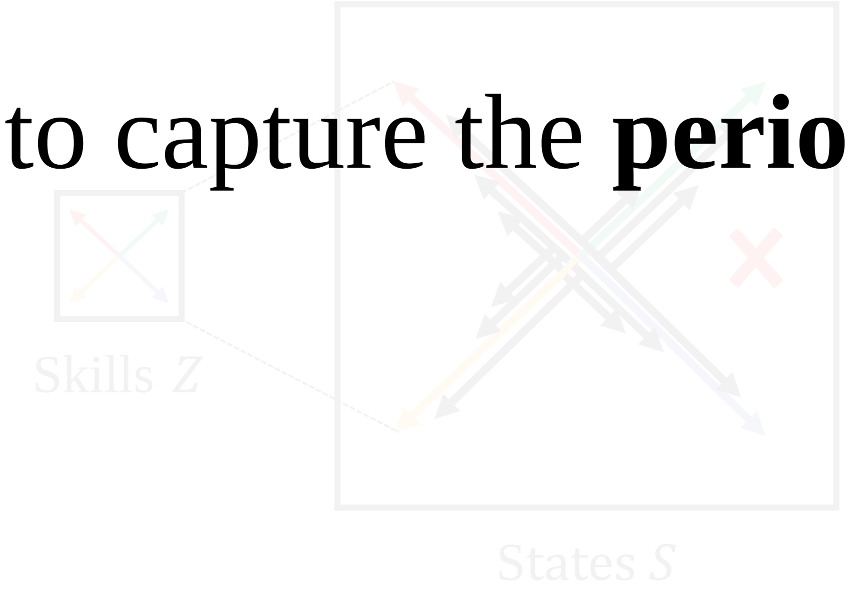
Prior works often fail to learn *multi-timescale* behaviors

(2) Distance - maximizing skill discovery (e.g., LSD, CSD, METRA, …)

$$\mathcal{J}_{\text{DSD}} := \mathbb{E}_{(z,\tau)\sim\mathcal{D}}\left[(\phi(s_{t+1}) - \phi(s_t))^\top z\right] \quad \text{s.t.} \quad \|\phi(x) - \phi(y)\| \leq d(x,y) \quad \forall x, y \in \mathcal{D}$$

Both approaches often fail

to capture the **periodic structure** of behaviors.



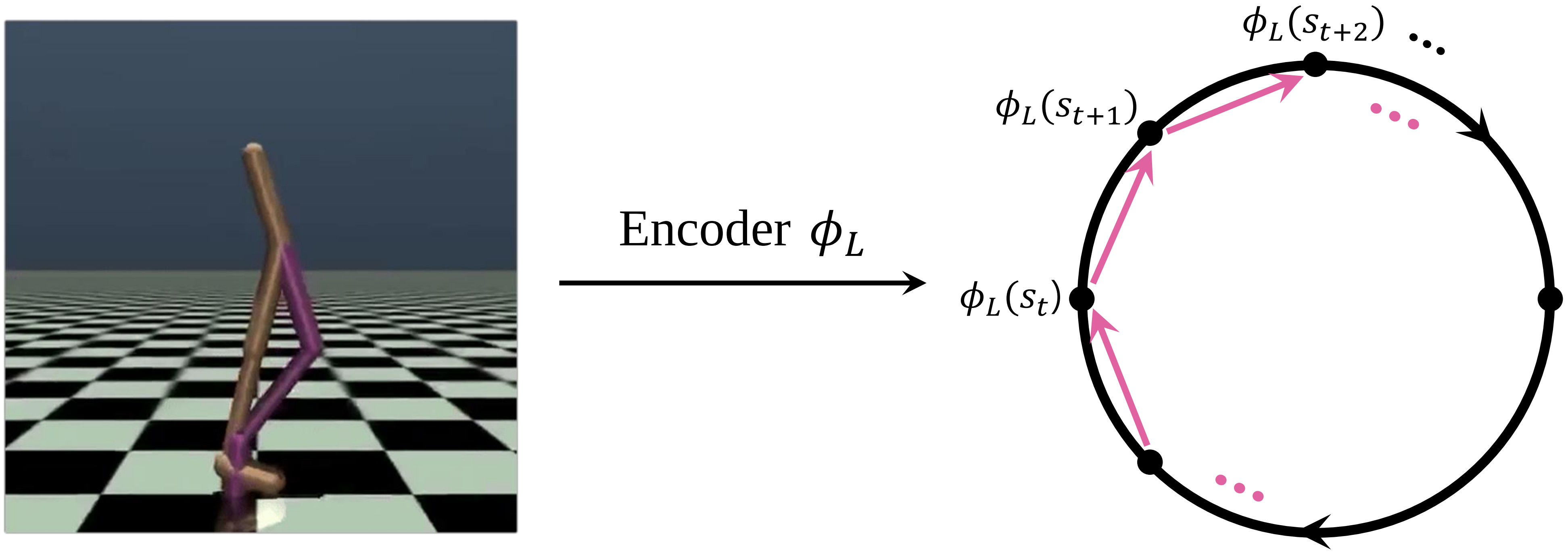
"prefers *hard-to-achieve* behaviors"

"little incentive to adjust the temporal patterns of skills"

Intuition

Intuition

Construct a **circular latent space** to capture **periodic** behavior



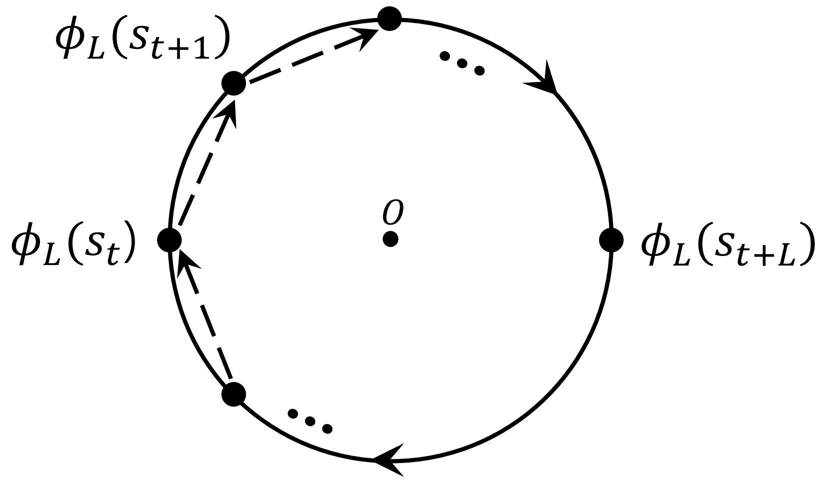
Representations for Periodicity

Representations for Periodicity

To learn a **representation** that returns to its initial state every **2L** timesteps..

Representations for Periodicity

To learn a **representation** that returns to its initial state every **2L** timesteps..

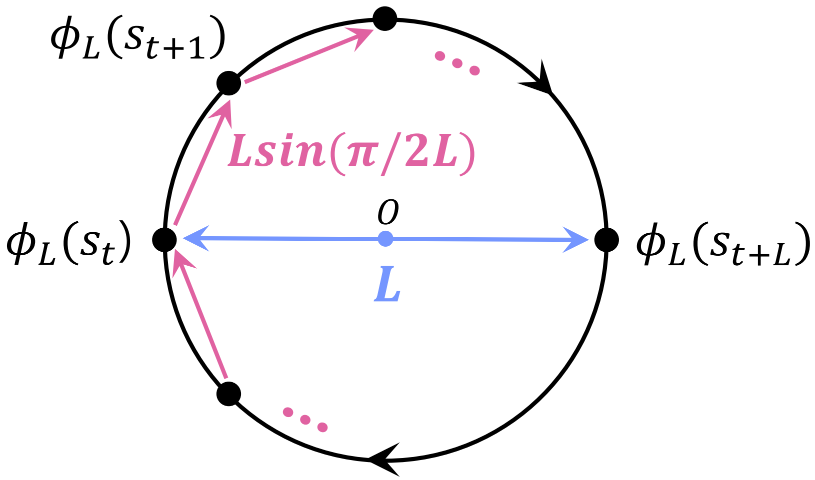


$$\phi_L(s_t) = \phi_L(s_{t+2L})$$

$$\phi_L(s_t) = -\phi_L(s_{t+L}) \quad \forall t \in \{0, 1, 2, \ldots\}$$

Representations for Periodicity

To learn a **representation** that returns to its initial state every **2L** timesteps..



$$\mathcal{J}_{\text{PSD},\phi} := \mathbb{E}_{p(\tau,L)}\Big[\|\phi_L(s_{t+L}) - \phi_L(s_t)\|_2 - k\,\|\phi_L(s_{t+L}) + \phi_L(s_t)\|_2\Big]$$

$$\text{s.t.}\quad \|\phi_L(s_{t+L}) - \phi_L(s_t)\|_2 \;\le\; L,$$

$$\|\phi_L(s_{t+1}) - \phi_L(s_t)\|_2 \;\le\; L\,\sin(\pi/2L)$$

"Constructs a 2L-gon inscribed in a circle of diameter L"

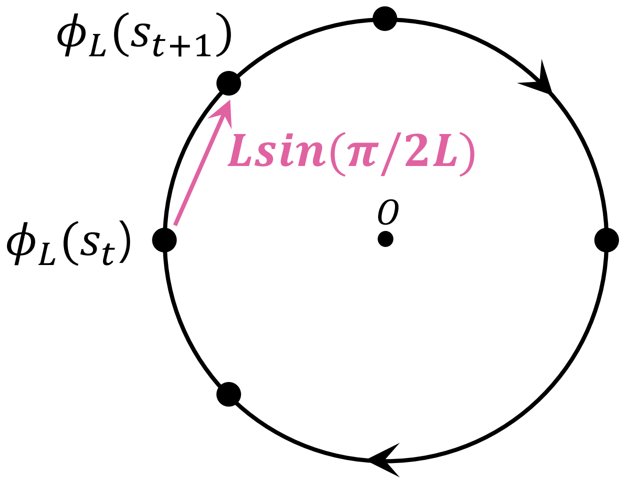
Single-step Intrinsic reward

Single-step Intrinsic reward

To learn a policy $\pi(a \mid s, L)$ that repeats every **2L** timesteps..

Single-step Intrinsic reward

To learn a policy $\pi(a \mid s, L)$ that repeats every **2L** timesteps..

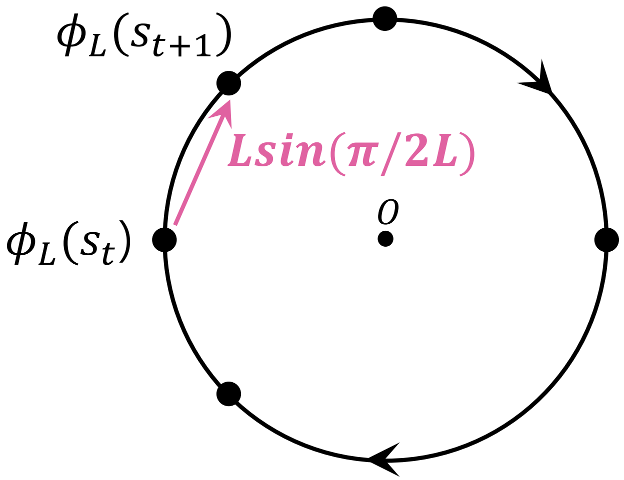


$$\Delta \;:=\; \|\phi_L(s_{t+1}) - \phi_L(s_t)\|_2 \;-\; L\sin(\pi/2L)$$

$$r_{\text{PSD}}(s_t, s_{t+1}, L) := \exp\left(-\kappa\,\Delta^2\right)$$

Single-step Intrinsic reward

To learn a policy $\pi(a \mid s, L)$ that repeats every **2L** timesteps..



$$\Delta \;:=\; \|\phi_L(s_{t+1}) - \phi_L(s_t)\|_2 \;-\; L\sin\big(\pi/2L\big)$$

$$r_{\text{PSD}}(s_t, s_{t+1}, L) := \exp\big(-\kappa\,\Delta^2\big)$$

“Single-step intrinsic reward encouraging **2L-periodicity**”

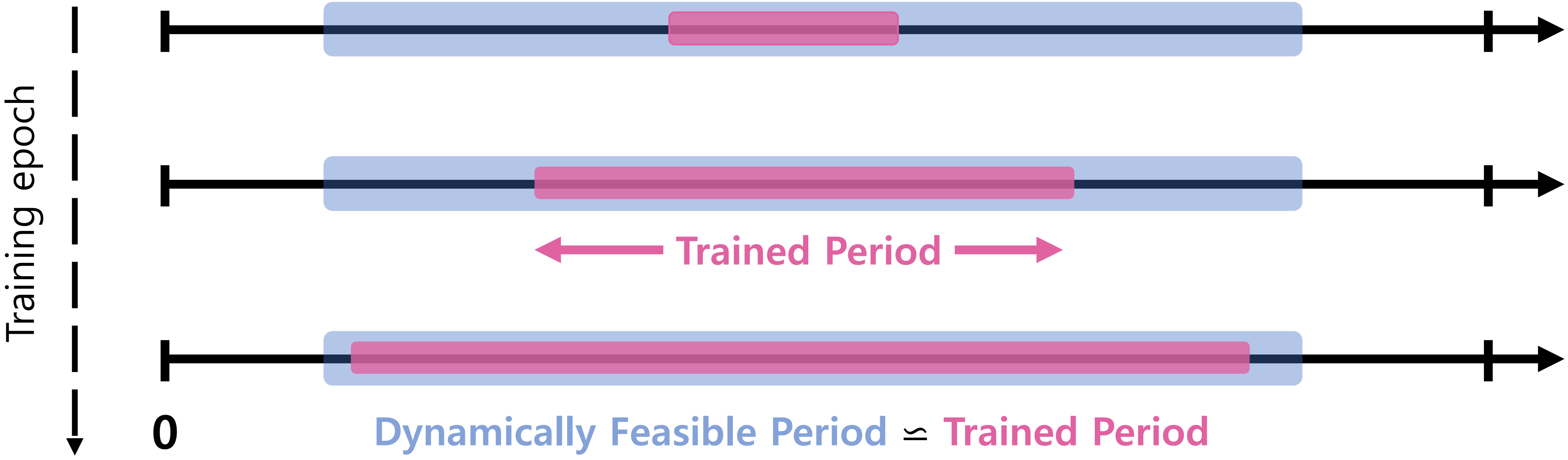
How can PSD learn maximally diverse periods?

How can PSD learn maximally diverse periods?

Propose increasingly harder periods to expand into the dynamically feasible range

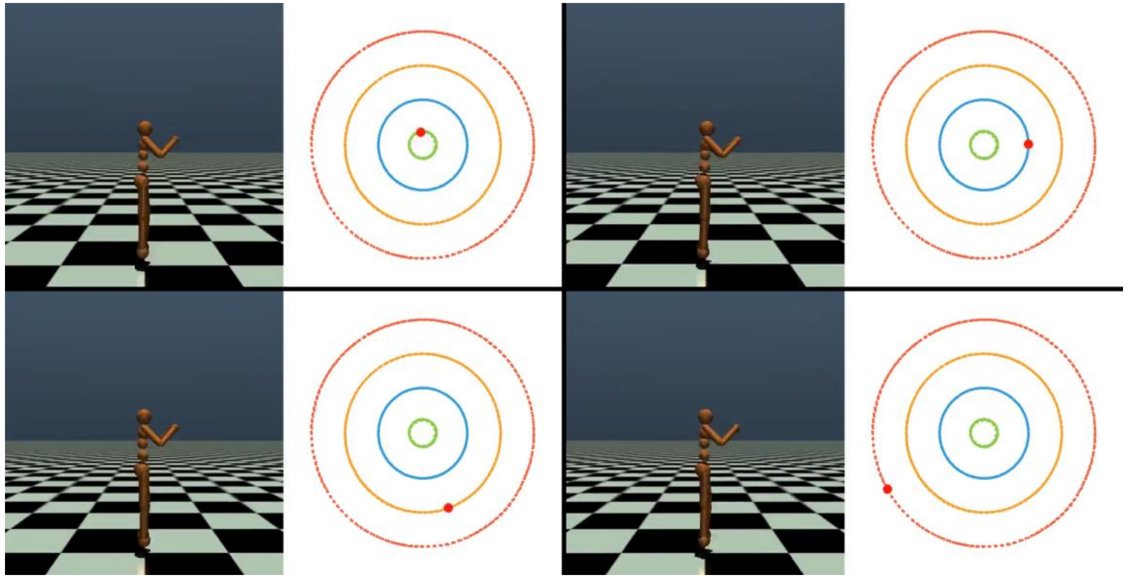
How can PSD learn maximally diverse periods?

Propose increasingly harder periods to expand into the dynamically feasible range

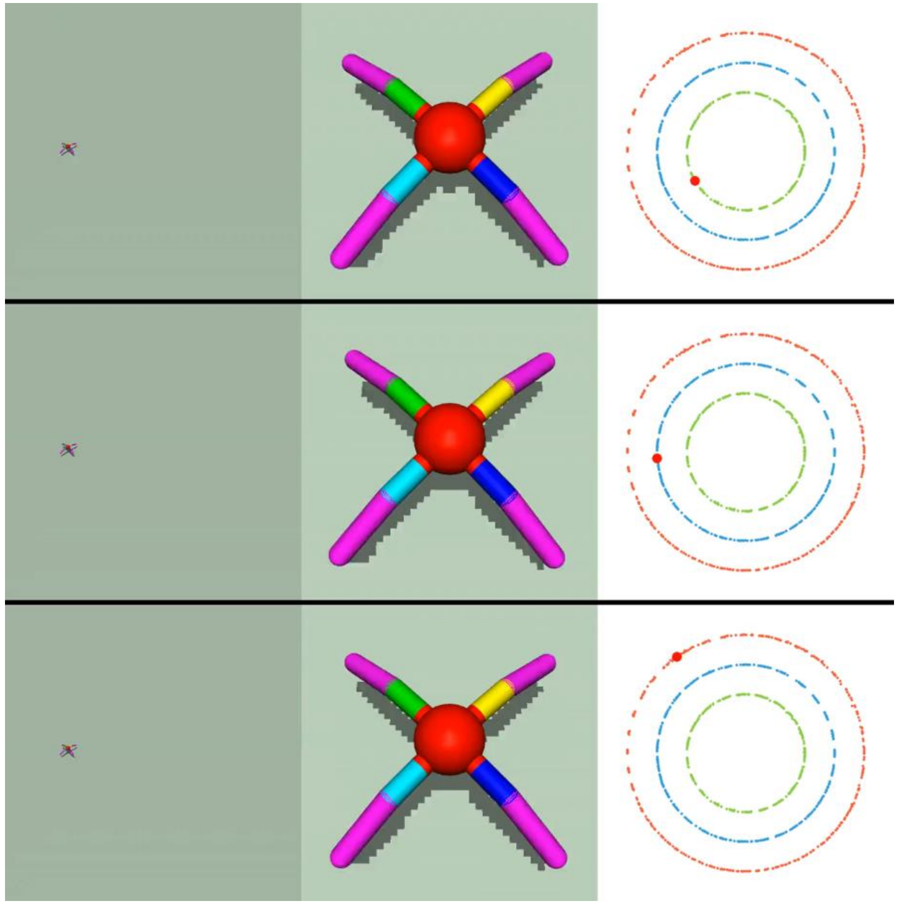


* See our paper for detailed update rule 🙂

Results : Latent Visualization



State-based environment



Pixel-based environment

* See our project page for more experimental demos 🙂

Results : PSD with METRA (Park et al., 2023)

Objective of METRA

$$\mathcal{J}_{\text{METRA},\phi_m} = \mathbb{E}_{(s,s',z)\sim\mathcal{D}}\left[(\phi_m(s') - \phi_m(s))^\top z + \lambda_m \cdot \min\left(\epsilon, 1 - \|\phi_m(s') - \phi_m(s)\|_2^2\right)\right]$$
$$\mathcal{J}_{\text{METRA},\lambda_m} = -\lambda_m \cdot \mathbb{E}_{(s,s',z)\sim\mathcal{D}}\left[\min\left(\epsilon, 1 - \|\phi_m(s') - \phi_m(s)\|_2^2\right)\right],$$

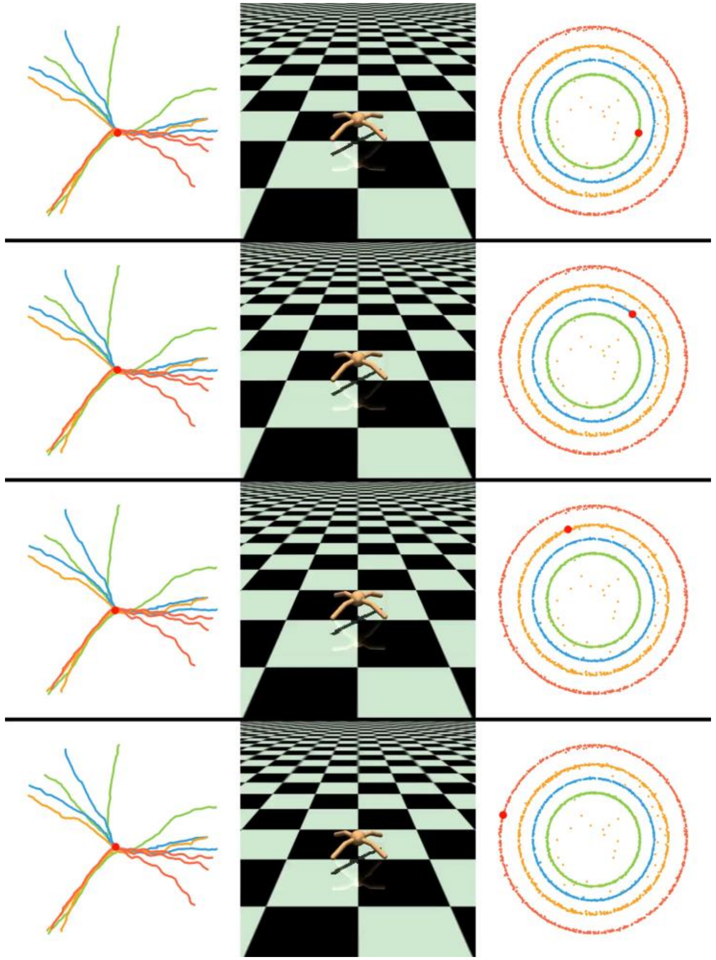
Intrinsic Reward of PSD with METRA (with mutual conditioning)

$$\phi_L(s) \longrightarrow \phi_L(s, z), \quad \phi_m(s) \longrightarrow \phi_m(s, L)$$

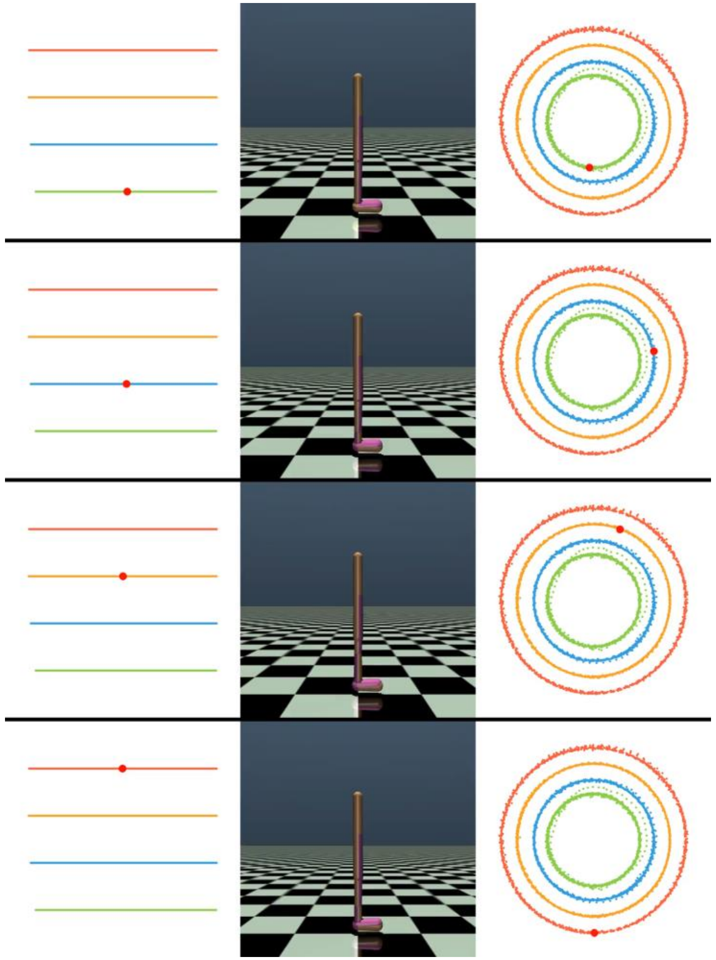
$$\pi(a\,|\,s, z, L) \leftarrow \arg\max_{\pi}\ \mathbb{E}_{p(\tau,z,L)}\Big[\sum_{t=0}^{T-1} \underbrace{(\phi_m(s_{t+1}) - \phi_m(s_t))^\top z}_{r_{\text{METRA}}} + \underbrace{\exp(-\kappa\,\Delta(L)^2)}_{r_{\text{PSD}}}\Big]$$

* See our paper for details 🙂

Results : PSD with METRA (Park et al., 2023)



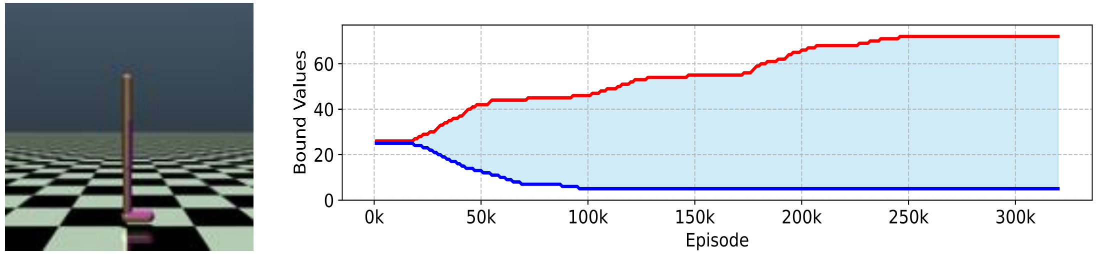
Ant



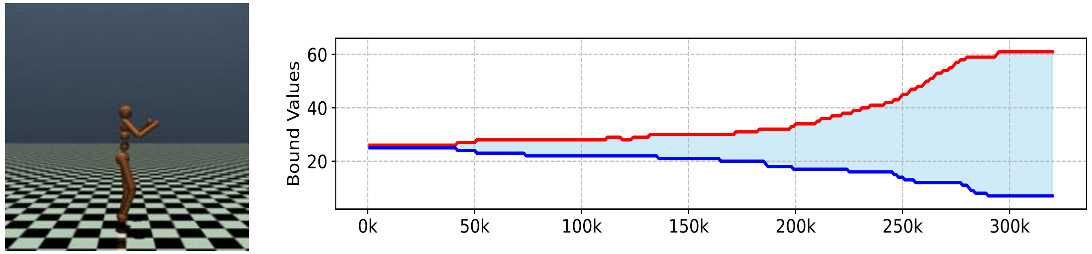
Walker2d

* See our project page for more experimental demos 🙂

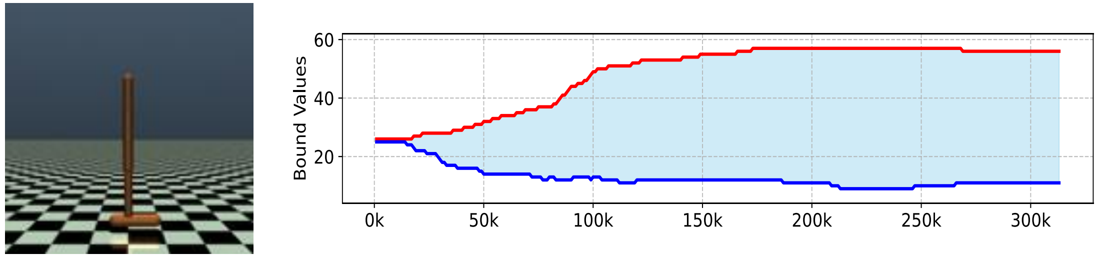
Results : Evolution of the sampling bounds during training



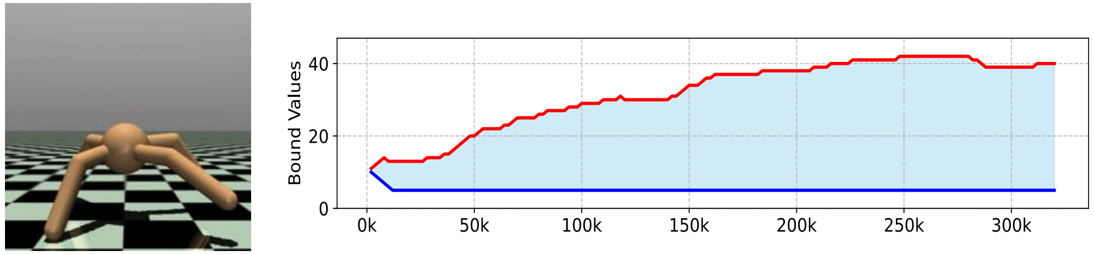
Walker2D



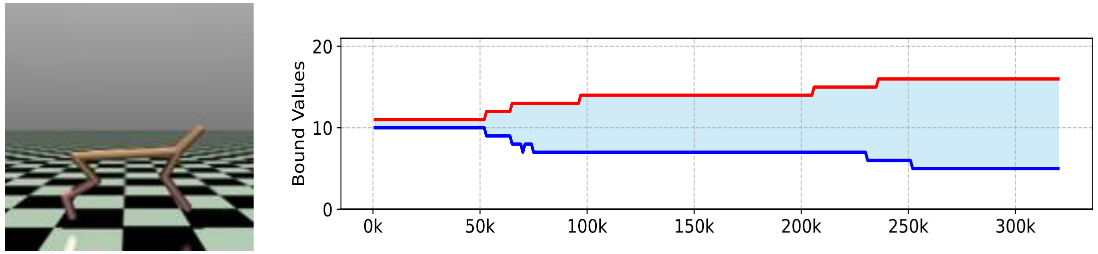
Humanoid



Hopper



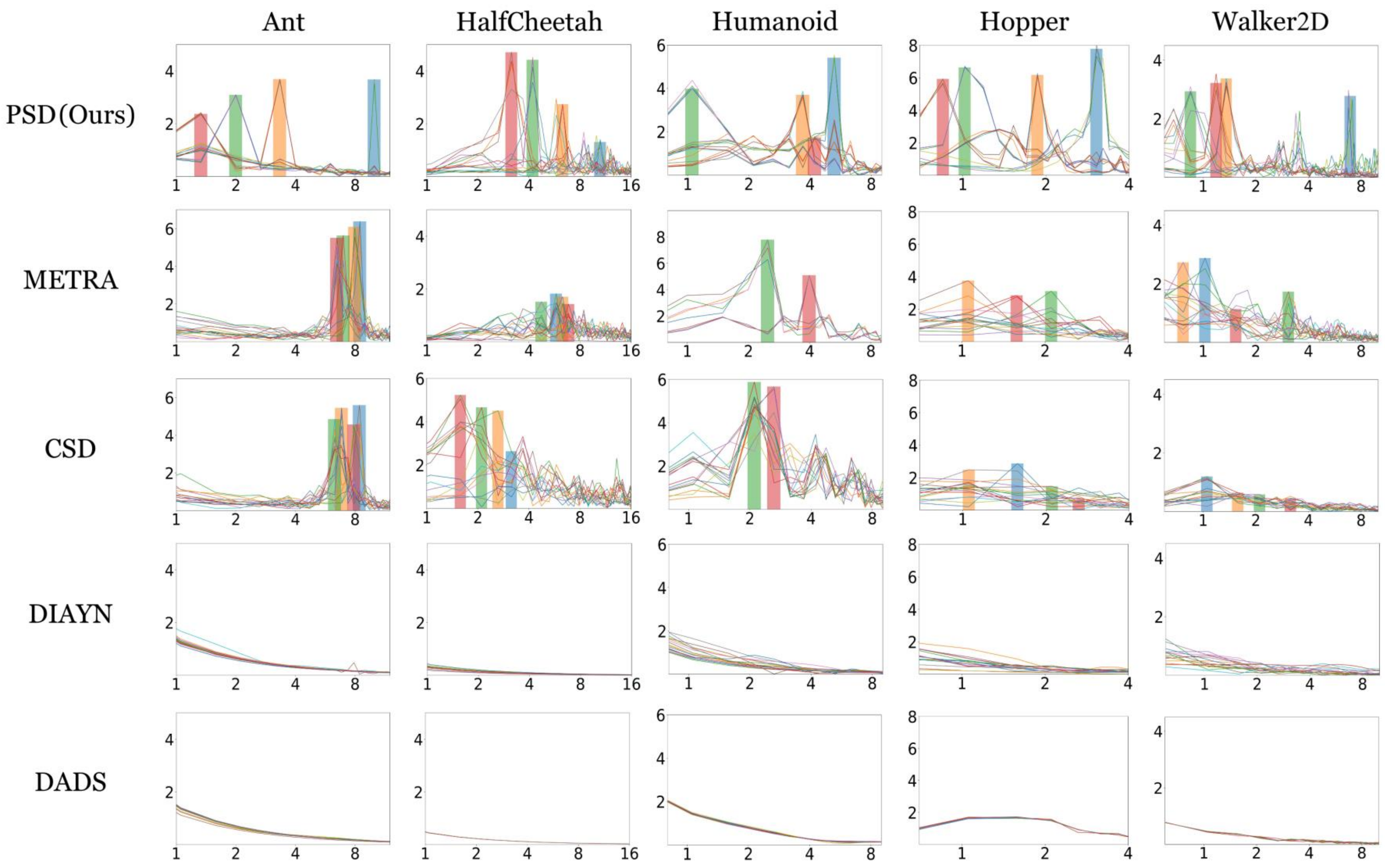
Ant



HalfCheetah

Trained Period Lower Bound Upper Bound

Results : Skill trajectories in the frequency domain



Results : Downstream task performance

Table 1: **Comparison of downstream task performance.** We evaluate PSD against existing skill discovery methods. High-level policies are trained using PPO with the skill policies kept frozen. All reported values are average returns over 10 seeds.

Downstream task	DIAYN	DADS	CSD	METRA	**PSD (Ours)**
HalfCheetah-hurdle	0.6±0.5	0.9±0.3	0.8±0.6	1.9±0.8	**3.8**±2.0
Walker2D-hurdle	2.6±0.5	1.9±0.3	4.1±1.3	3.1±0.5	**5.4**±1.4
HalfCheetah-friction	13.2±3.4	12.4±2.9	12.5±3.8	30.1±13.1	**43.4**±19.1
Walker2D-friction	4.6±1.2	1.6±0.1	5.3±0.3	5.2±1.6	**8.7**±1.7

Conclusion

We introduce **Periodic Skill Discovery (PSD)**, a framework for unsupervised skill discovery that captures the periodic nature of behaviors by embedding states into a circular latent space.

PSD provides a scalable and principled framework for discovering temporally structured behaviors in RL.

Project Page (Demos)