Offline Diversity Maximization Under Imitation Constraints

Marin Vlastelica

Jin Cheng

Georg Martius

Pavel Kolev

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

VolkswagenStiftung

ETH zürich

Motivation

Diverse

- Robust solutions
- Multiple options

[DIAYN, DADS, DOMINO]

online setting

Offline

- Use large datasets
- Safe learning

[AWAC, BC, CRR, IQL, CQL]

single expert, not diverse

Imitation

- No reward engineering
- Human demonstrations

[GAIL, SMODICE]

single expert, not diverse

Motivation

Diverse

- Robust solutions
- Multiple options

[DIAYN, DADS, DOMINO]

online setting

Offline

- Use large datasets
- Safe learning

[AWAC, BC, CRR, IQL, CQL]

single expert, not diverse

Imitation

- No reward engineering
- Human demonstrations

[GAIL, SMODICE]

single expert, not diverse

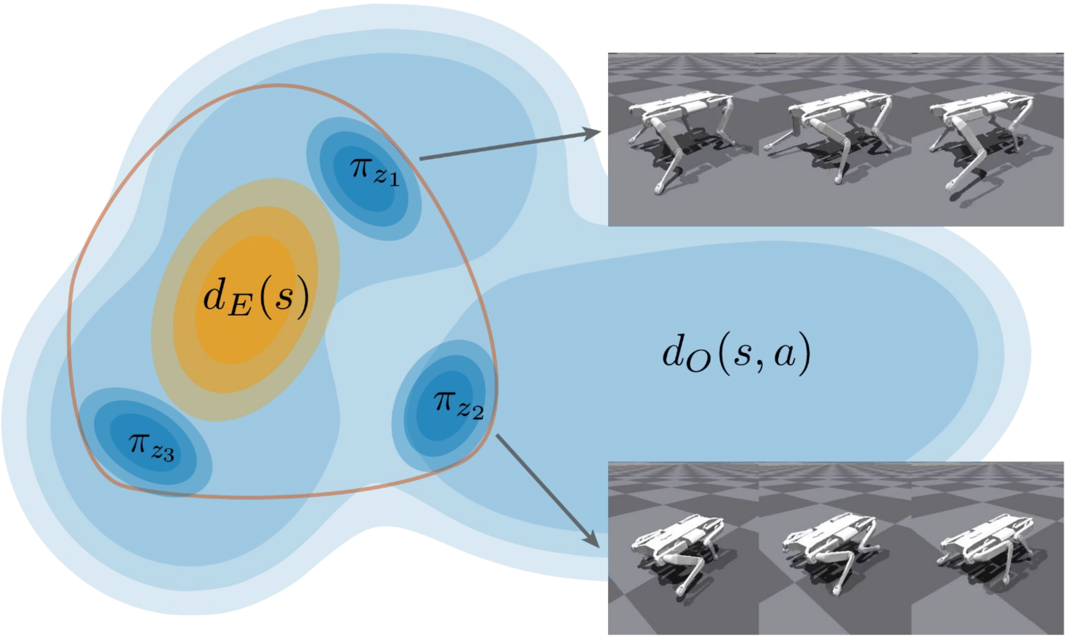
Propose: principled algorithm for **Diverse** **Offline** **Imitation** (DOI) learning

Problem Formulation

$$\max_{\{d_z(S)\}_{z \in Z}} \mathcal{I}(S; Z)$$

subject to

$$\mathrm{D_{KL}}(d_z(S) || d_E(S)) \leq \varepsilon \quad \forall z$$



Input:

state-action behavior dataset $\mathcal{D}_O \sim d_O(s, a)$

state-only expert dataset $\mathcal{D}_E \sim d_E(s)$

Relaxed Problem Formulation

$$\max_{\{d_z(S)\}_{z \in Z}} \mathcal{I}(S; Z)$$

Mutual Information: Variational Lower Bound

$$\mathcal{I}(S; Z) \geq \sum_z \mathbb{E}_{d_z(s)} \left[\frac{\log\left(|Z| q(z|s)\right)}{|Z|} \right]$$

$q(z|s)$ train a **skill-discriminator**

Relaxed Problem Formulation

$$\max_{\{d_z(S)\}_{z\in Z}} \quad \mathcal{I}(S;Z)$$

Mutual Information: Variational Lower Bound

$$\mathcal{I}(S;Z) \quad \geq \quad \sum_z \mathbb{E}_{d_z(s)}\left[\frac{\log\left(|Z|q(z|s)\right)}{|Z|}\right]$$

$q(z|s)$ train a **skill-discriminator**

subject to

$$\mathrm{D_{KL}}(d_z(S)||d_E(S)) \leq \varepsilon \quad \forall z$$

SMODICE expert **(offline)**

$$d_{\tilde{E}}(S,A) \approx \arg\min_{d(s,a)} \mathrm{D_{KL}}(d(S)||d_E(S))$$

subject to

$$\mathrm{D_{KL}}(d_z(S,A)||d_{\tilde{E}}(S,A)) \leq \varepsilon \quad \forall z$$

[SMODICE] Y. Ma, A. Shen, D. Jayaraman, O. Bastani "Versatile Offline Imitation from Observations and Examples via Regularized State-Occupancy Matching", ICML 2022

Algorithmic Approach (Lagrange)

$$\max_{\substack{d_z(s,a)\\ q(z|s)}} \min_{\lambda \geq 0} \sum_z \mathbb{E}_{d_z(s)}\left[\frac{\log\left(|Z| q(z|s)\right)}{|Z|}\right] + \boxed{\sum_z \lambda_z \left[\epsilon - \mathrm{D_{KL}}\left(d_z(S,A) || d_{\widetilde{E}}(S,A)\right)\right]}$$

Diversity Imitation

Algorithmic Approach (Fenchel)

$$\max_{\substack{d_z(s,a)\\ q(z|s)}} \min_{\lambda \geq 0} \sum_z \mathbb{E}_{d_z(s)} \left[\frac{\log(|Z| q(z|s))}{|Z|} \right] + \sum_z \lambda_z \left[\epsilon - \mathrm{D_{KL}} \left(d_z(S,A) || d_{\widetilde{E}}(S,A) \right) \right]$$

$$\max_{\substack{d_z(s,a)\\ q(z|s)}} \min_{\lambda > 0} \sum_z \lambda_z \left\{ \epsilon + \boxed{\mathbb{E}_{d_z(s,a)} \left[R_z^{\lambda}(s,a) \right] - \mathrm{D_{KL}} \left(d_z(S,A) || d_O(S,A) \right)} \right\}$$

Regularized RL Problem

DICE (offline)

$$\eta_z(s,a) = \frac{d_z(s,a)}{d_O(s,a)}$$

[DICE] O. Nachum, B. Dai, “Reinforcement learning via fenchel-rockafellar duality”, arXiv 2020

Algorithmic Approach

$$\max_{\substack{d_z(s,a)\\ q(z|s)}} \min_{\lambda \geq 0} \sum_z \mathbb{E}_{d_z(s)}\left[\frac{\log\left(|Z| q(z|s)\right)}{|Z|}\right] + \sum_z \lambda_z \left[\epsilon - \mathrm{D_{KL}}\left(d_z(S,A) || d_{\widetilde{E}}(S,A)\right)\right]$$

$$\max_{\substack{d_z(s,a)\\ q(z|s)}} \min_{\lambda > 0} \sum_z \lambda_z \left\{\epsilon + \mathbb{E}_{d_z(s,a)}\left[R_z^{\lambda}(s,a)\right] - \mathrm{D_{KL}}\left(d_z(S,A) || d_O(S,A)\right)\right\}$$

$$R_z^{\lambda}(s,a) := \underbrace{\frac{1}{\lambda_z}}_{\text{Constraint Violation}} \underbrace{\frac{\log\left(q(z|s)|Z|\right)}{|Z|}}_{\text{Skill Diversity}} + \underbrace{\log \eta_{\widetilde{E}}(s,a)}_{\text{Expert Imitation}}$$

DICE **(offline)**

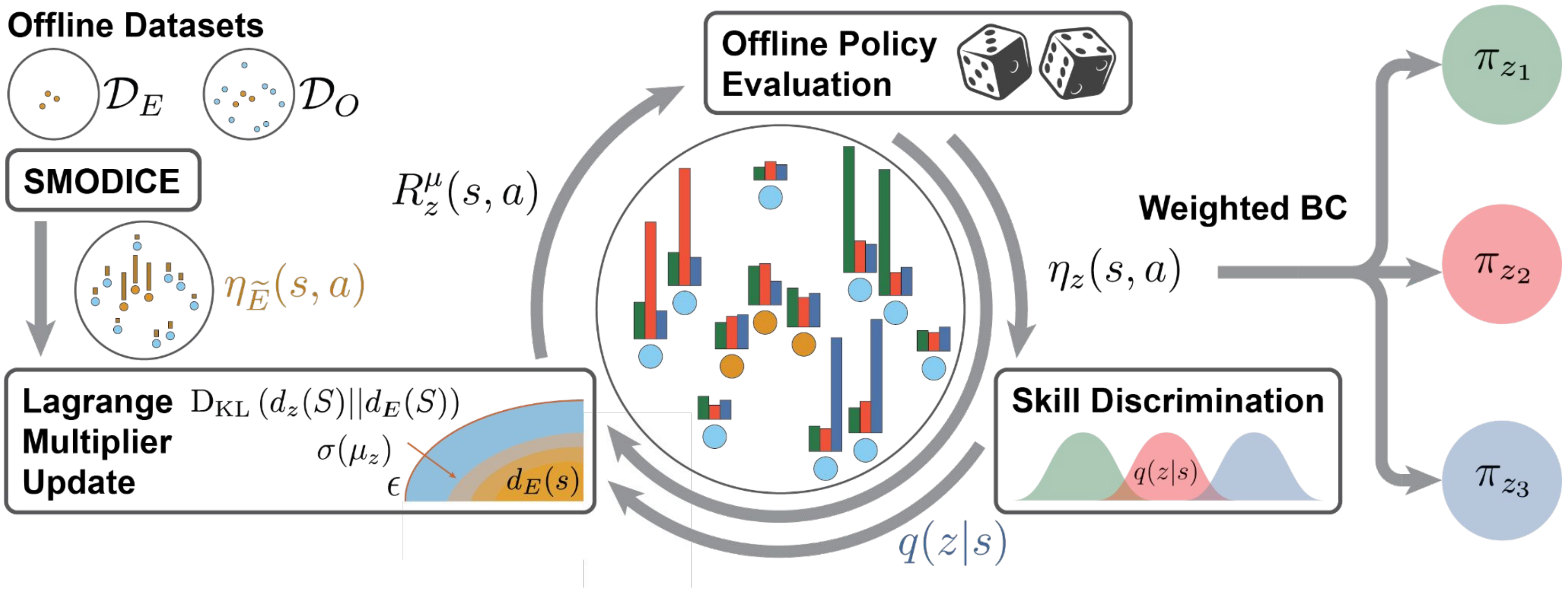
$$\eta_z(s,a) = \frac{d_z(s,a)}{d_O(s,a)}$$

SMODICE expert **(offline)**

$$\eta_{\widetilde{E}}(s,a) = \frac{d_{\widetilde{E}}(s,a)}{d_O(s,a)}$$

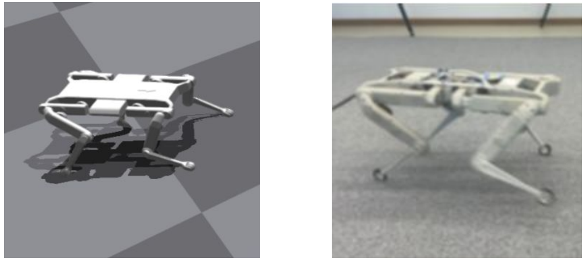
[DICE] O. Nachum, B. Dai, "Reinforcement learning via fenchel-rockafellar duality", arXiv 2020
[SMODICE] Y. Ma, A. Shen, D. Jayaraman, O. Bastani, "Versatile Offline Imitation from Observations and Examples via Regularized State-Occupancy Matching", ICML 2022

Alternating Optimization Scheme



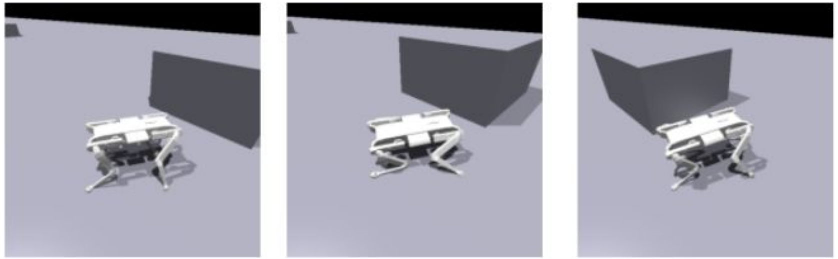
Experiments

I. Locomotion Task (Sim & Real)

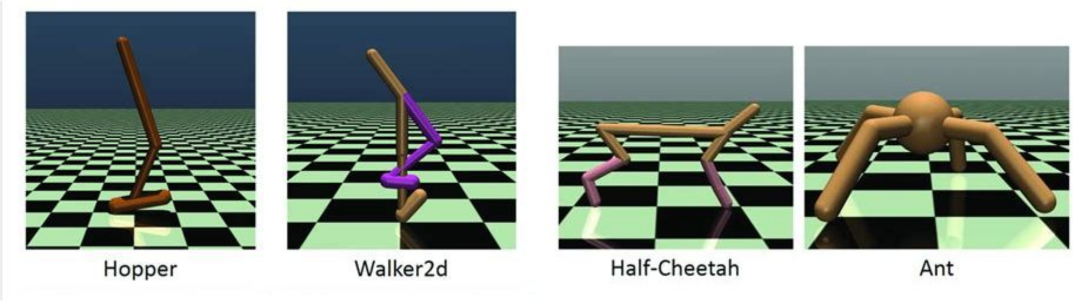


Solo12

II. Obstacle Navigation Task (Sim)

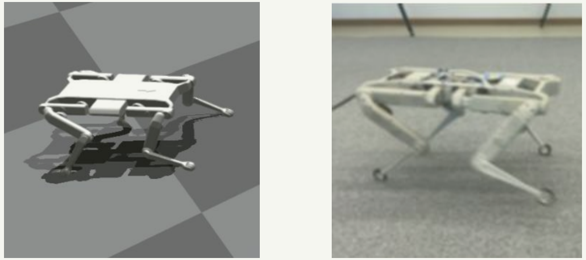


III. D4RL Envs (Sim)



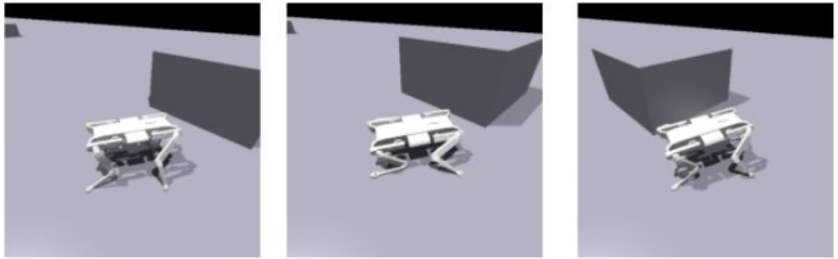
Experiments

I. Locomotion Task (Sim & Real)

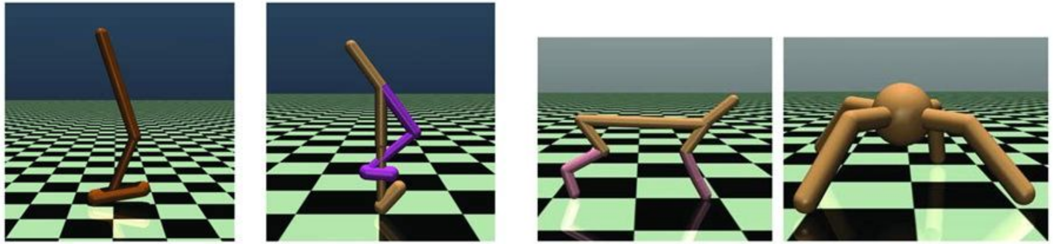


Solo12

II. Obstacle Navigation Task (Sim)



III. D4RL Envs (Sim)



Hopper Walker2d Half-Cheetah Ant

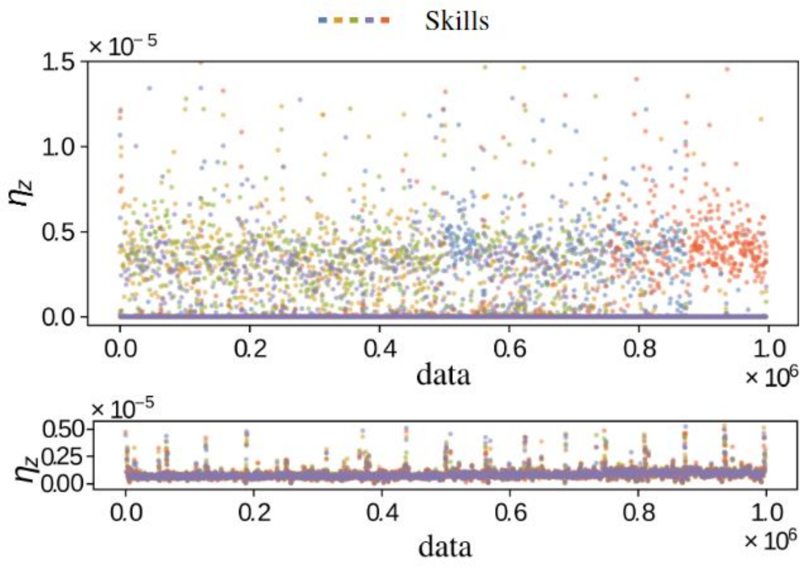
I. Locomotion Task (Sim)

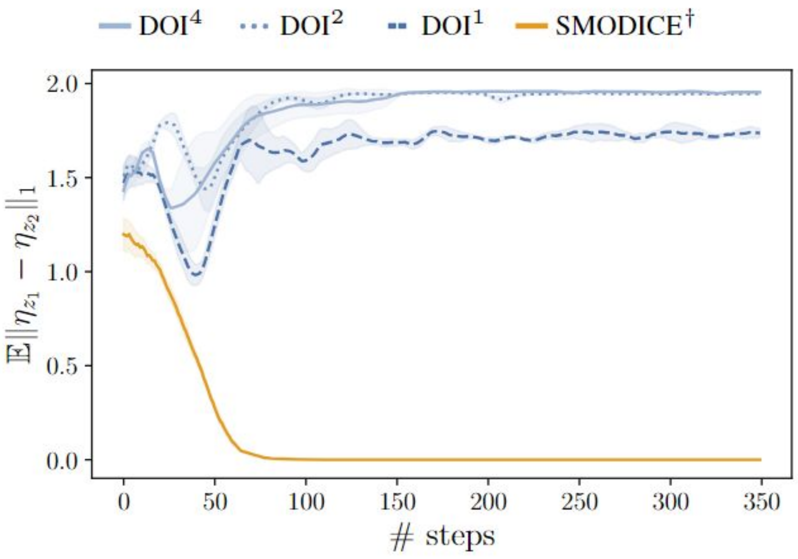
(Expected **Importance Ratios**)

Offline Evaluation

1) DOI skills well-separate data

2) Constraint level ε controls ratio distance



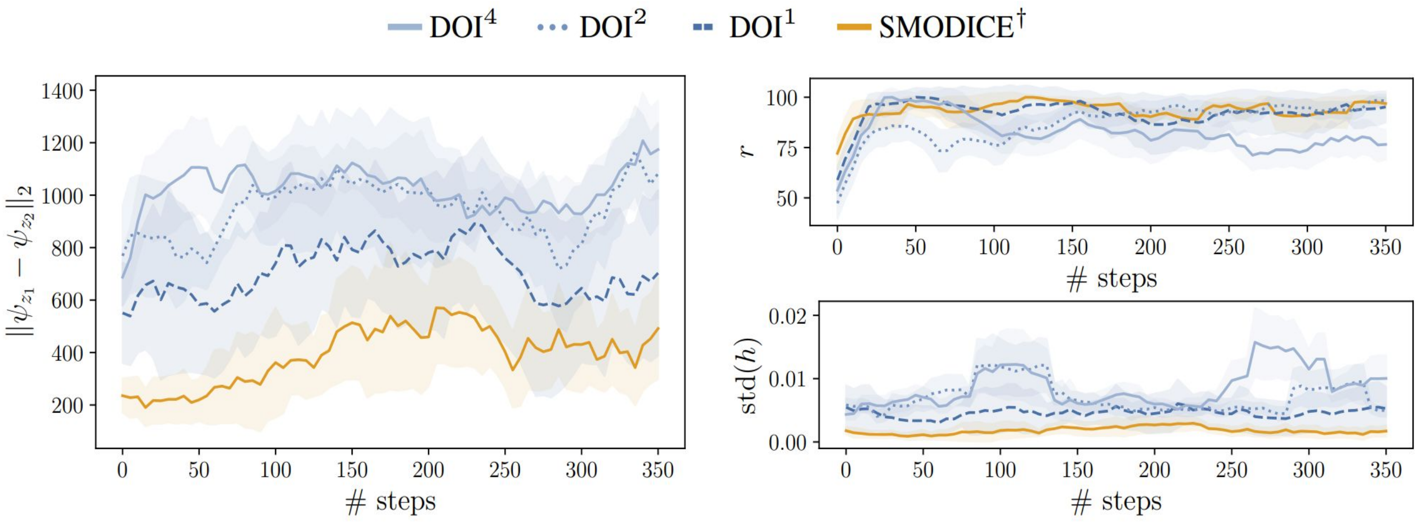


I. Locomotion Task (Sim)

(Expected **Successor Features**)

Online (Monte Carlo) Evaluation

3) Relaxed constraints yield increased diversity, albeit at the expense of performance loss.



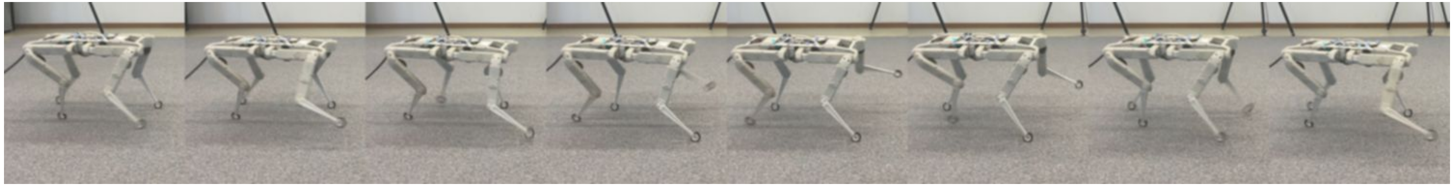
(a)

(b)

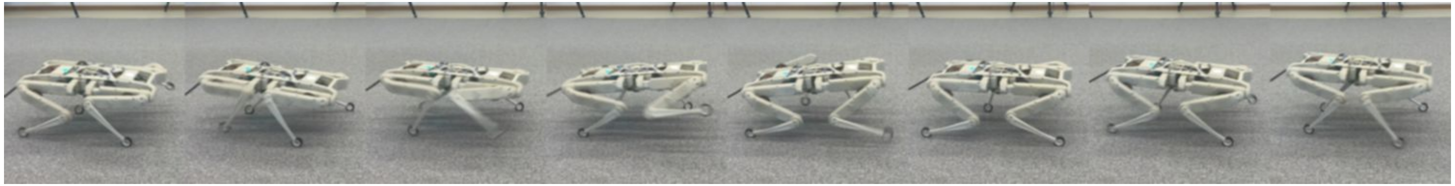
I. Locomotion Task (Real)

4) DOI skills trained in **SIM** (*with domain randomization)* are successfully deployed in the **Real System**

Trot – High



Wave – Low



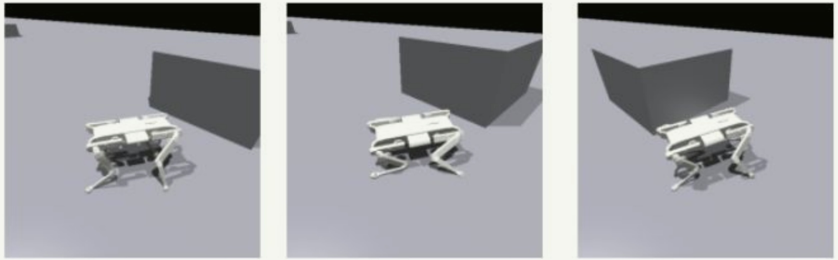
Trot – Middle

Experiments

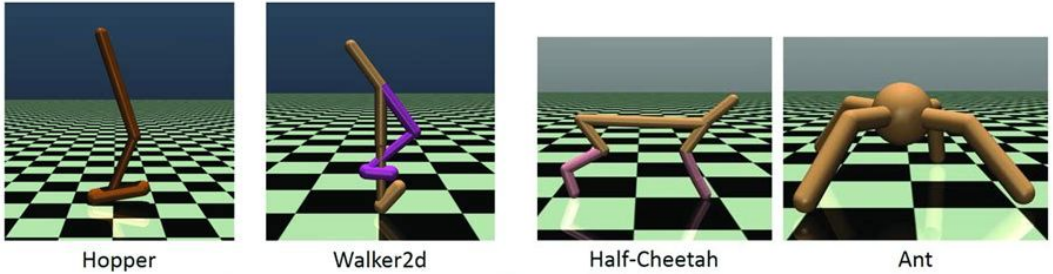
I. Locomotion Task (Sim & Real)

Solo12

II. Obstacle Navigation Task (Sim)



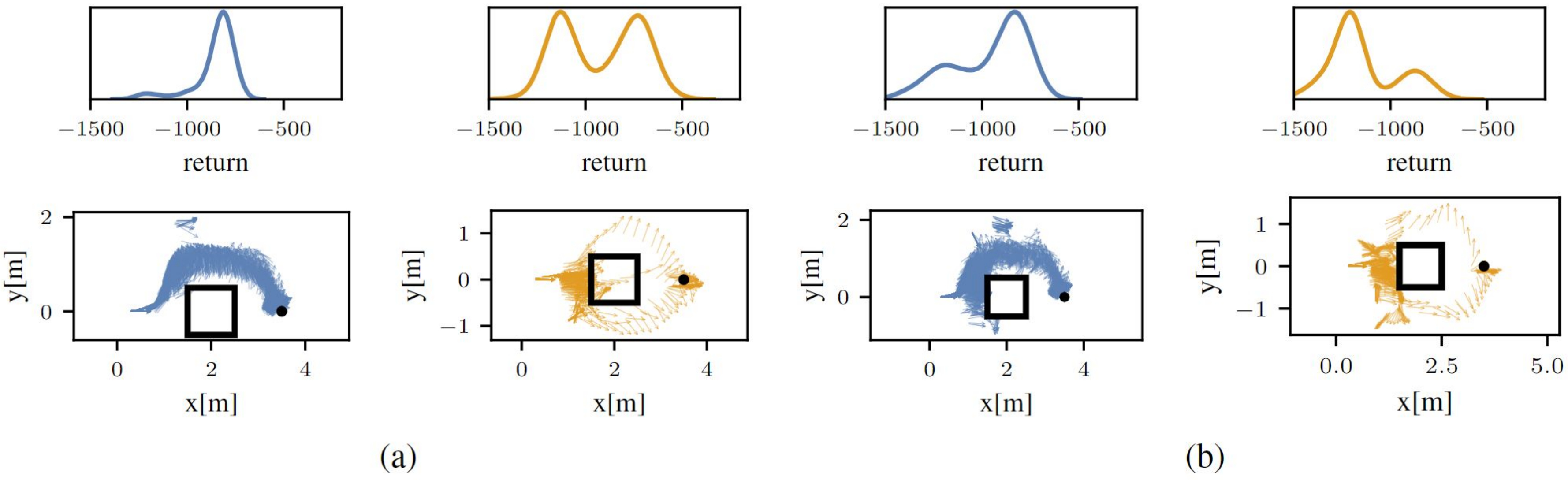
III. D4RL Envs (Sim)



II. Obstacle Navigation Task (Sim)

Online (Monte Carlo) Evaluation

5) SMODICE expert struggles with out-of-distribution (higher) box heights, while a robust DOI skill successfully navigates by detouring (to the left side).

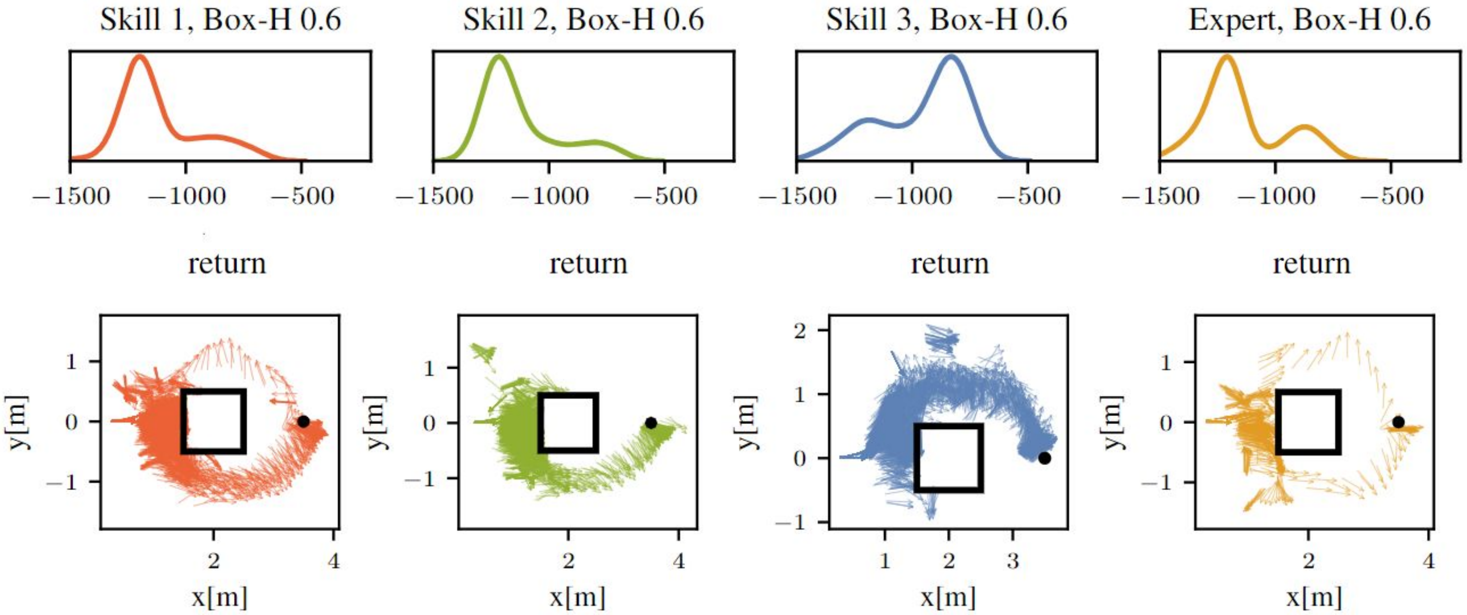


Box Height 0.3 m

Box Height 0.6 m

II. Obstacle Navigation Task (Sim)

Limitation: **6)** Not all learned DOI skills are robust. Selection is required.



Conclusion

Project Website

Principled algorithm (**DOI**)

Offline Diversity maximization under **Imitation** constraints

Experiments

Show **DOI**'s effectiveness on:

- Solo12 tasks (Locomotion & Obstacle Navigation)
- Standard D4RL environments

Limitation

Agent's performance is sensitive to relaxing the imitation constraints