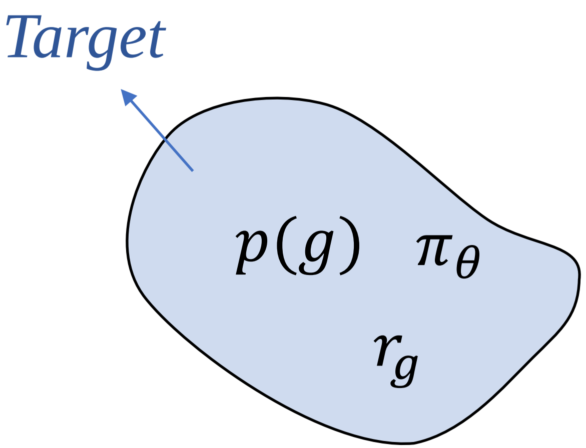
Unsupervised Domain Adaptation with Dynamics-Aware Rewards in Reinforcement Learning

Presenter: Jinxin Liu (liujinxin@westlake.edu.cn)

Co-Authors: Hao Shen, Donglin Wang, Yachen Kang, Qiangxing Tian

Unsupervised Reinforcement Learning (RL)

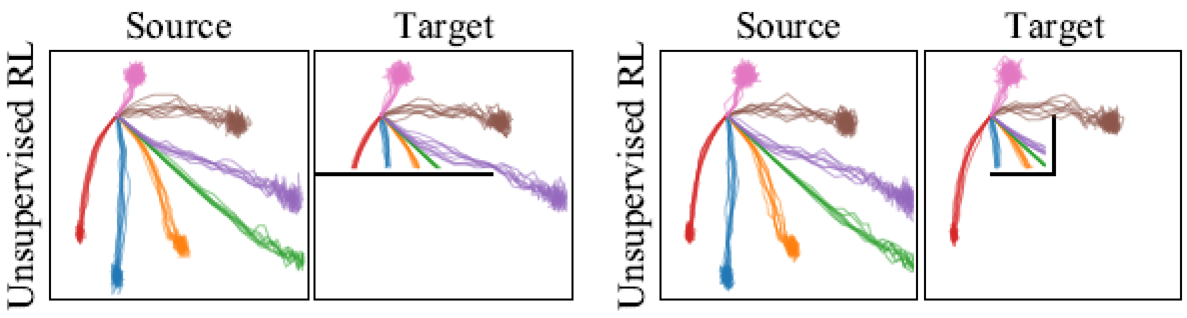
The standard unsupervised RL:
learning skills for the ***target*** environment.



- Representing goals:
  - Learning $p(g)$ in *target* environment.
  - Learning r_g in *target* environment.
- Learning π_θ in *target* environment.

1. Time-consuming and potentially expensive.

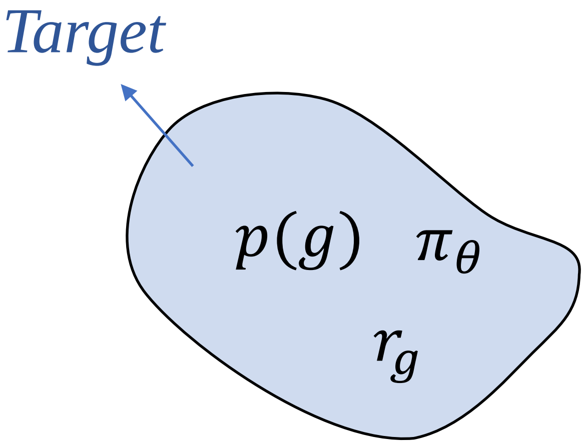
2. Transfer?
   Assuming a source environment.
   Direct transfer.



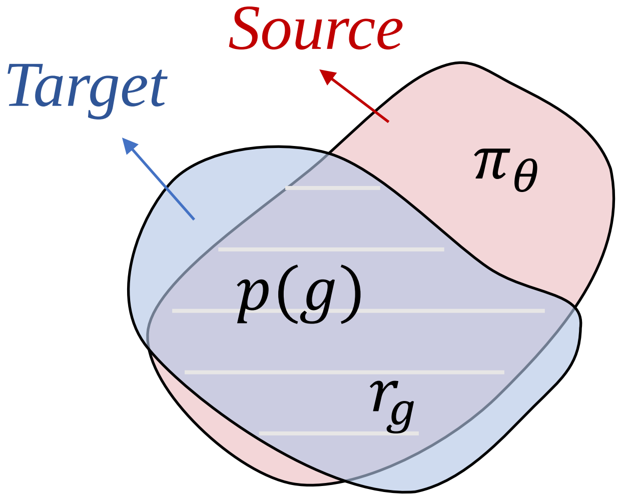
Domain Adaptation in Unsupervised RL

The standard unsupervised RL:
learning skills for the ***target*** environment.

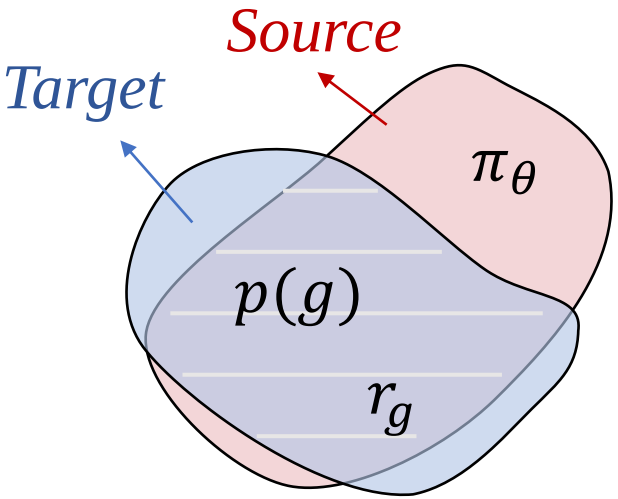
- Representing goals:
  - Learning $p(g)$ in *target* environment.
  - Learning r_g in *target* environment.
- Learning π_θ in *target* environment.



- Representing goals:
  - Learning $p(g)$ in source and *target*.
  - Learning r_g in source and *target*.
- Learning π_θ in source env.



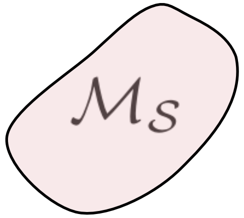
Domain Adaptation in Unsupervised RL



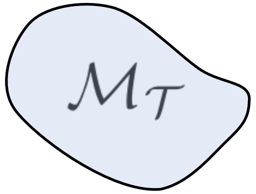
Source environment $\mathcal{M}_\mathcal{S}$, with transition dynamics $\mathcal{P}_\mathcal{S}$

Target environment $\mathcal{M}_\mathcal{T}$, with transition dynamics $\mathcal{P}_\mathcal{T}$

(with same initial state distribution, same state/action spaces)



Free



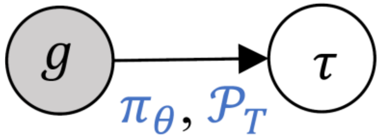
Limited

Assumption

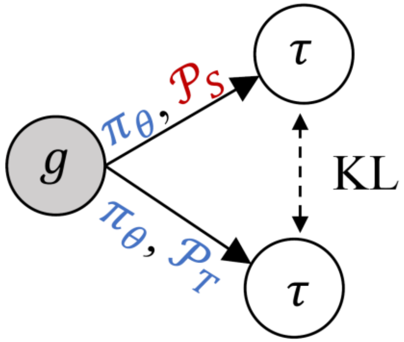
1. There is no transition that is possible in the target environment but impossible in the source environment: $\mathcal{P}_\mathcal{T}(s_{t+1}|s_t, a_t) > 0 \implies \mathcal{P}_\mathcal{S}(s_{t+1}|s_t, a_t) > 0$

2. The difference between environments in their dynamics negligibly affects the goal distribution.

Domain Adaptation in Unsupervised RL



(a)



(b)

$$\mathcal{I}_{\mathcal{P}_{\mathcal{T}},\pi_\theta}(g;\tau)$$

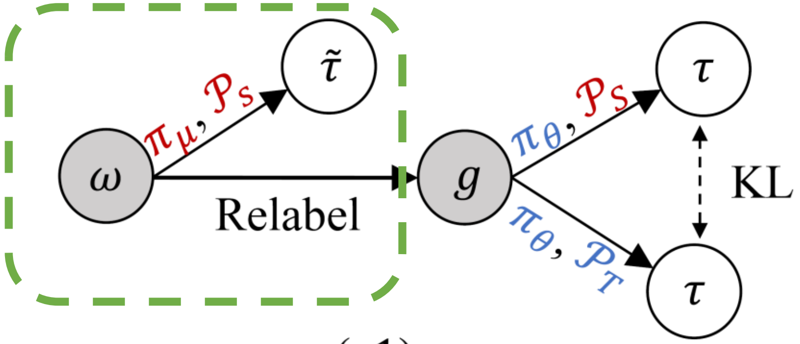
$$\mathcal{I}_{\mathcal{P}_{\mathcal{S}},\pi_\theta}(g;\tau) - \beta D_{\mathrm{KL}}\left(p_{\mathcal{P}_{\mathcal{S}},\pi_\theta}(g,\tau) \| p_{\mathcal{P}_{\mathcal{T}},\pi_\theta}(g,\tau)\right)$$

Time-consuming and potentially expensive.

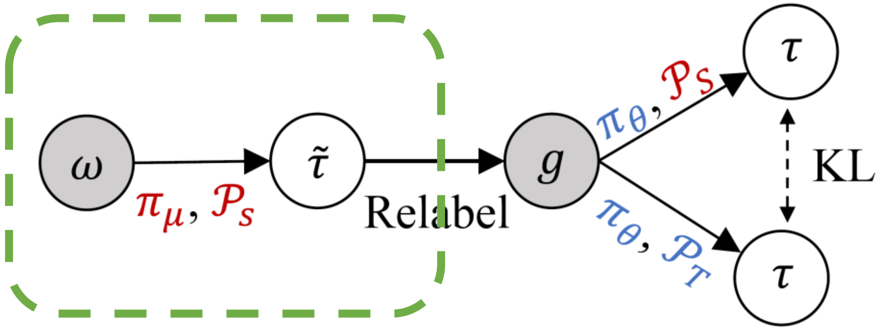
Assuming a source environment.

KL term penalizes producing a trajectory that cannot be generated in the target environment.

Domain Adaptation in Unsupervised RL



$(c1)$



$(c2)$

$$\mathcal{I}_{\mathcal{P}_\mathcal{S}, \pi_\mu}(\omega; \tilde{\tau}) + \mathcal{I}_{\mathcal{P}_\mathcal{S}, \pi_\theta}(g; \tau) - \beta D_{\mathrm{KL}}\left[p_{\mathcal{P}_\mathcal{S}, \pi_\theta}(g, \tau) \| p_{\mathcal{P}_\mathcal{T}, \pi_\theta}(g, \tau)\right]$$

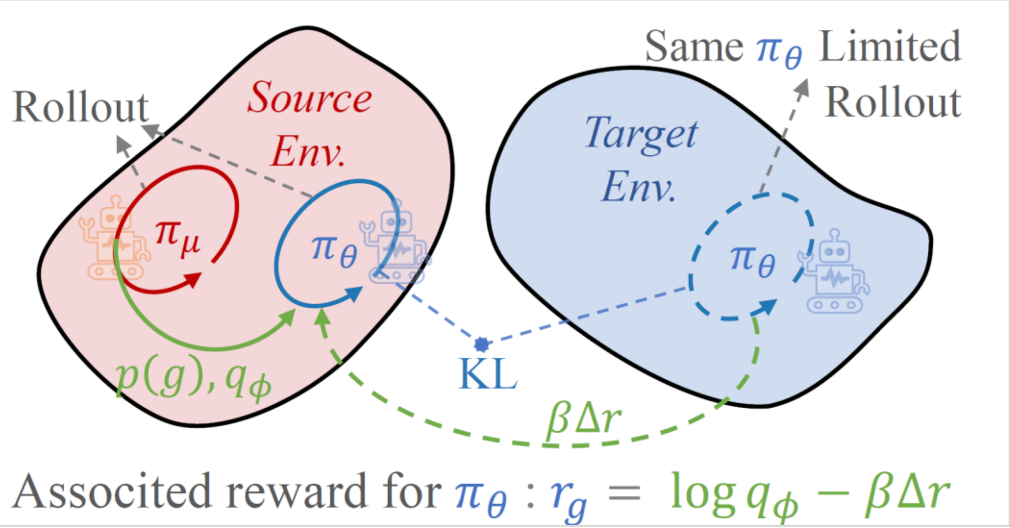
Probing policy π_μ : generating the goal distribution p(g) and acquiring the (partial) reward function r_g

Domain Adaptation in Unsupervised RL

$$\mathcal{I}_{\mathcal{P}_{\mathcal{S}},\pi_\mu}(\omega;\tilde{\tau}) + \mathcal{I}_{\mathcal{P}_{\mathcal{S}},\pi_\theta}(g;\tau) - \beta D_{\mathrm{KL}}\left[p_{\mathcal{P}_{\mathcal{S}},\pi_\theta}(g,\tau) \| p_{\mathcal{P}_{\mathcal{T}},\pi_\theta}(g,\tau)\right]$$

lower bound

$$2\mathcal{H}(\omega) + \mathbb{E}_{p_{\text{joint}}}\left[\log q_\phi(\omega|\tilde{s}_{t+1}) + \log q_\phi(\omega|s_{t+1})\right] - \mathbb{E}_{\mathcal{P}_{\mathcal{S}},\pi_\theta}\left[\beta \Delta r(s_t, a_t, s_{t+1})\right].$$



states $\tilde{s}_{t+1}$ and s_{t+1} are induced by the probing policy π_μ and the policy π_θ

p_{joint} denotes the joint distribution of ω, states $\tilde{s}_{t+1}$ and s_{t+1}

$$\Delta r(s_t, a_t, s_{t+1}) \triangleq \log \mathcal{P}_{\mathcal{S}}(s_{t+1}|s_t, a_t) - \log \mathcal{P}_{\mathcal{T}}(s_{t+1}|s_t, a_t)$$

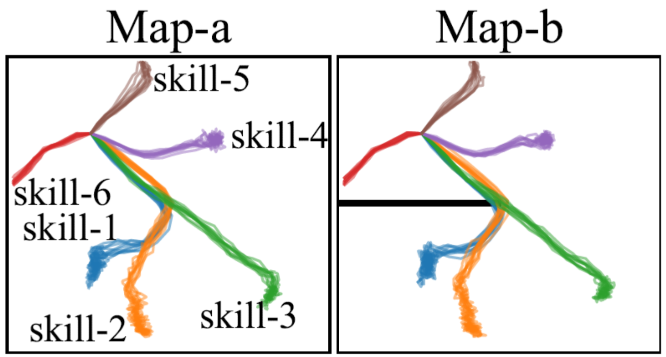
(State-action and state-action-next-state classifiers according Bayes' rule)

Connections to Prior Work

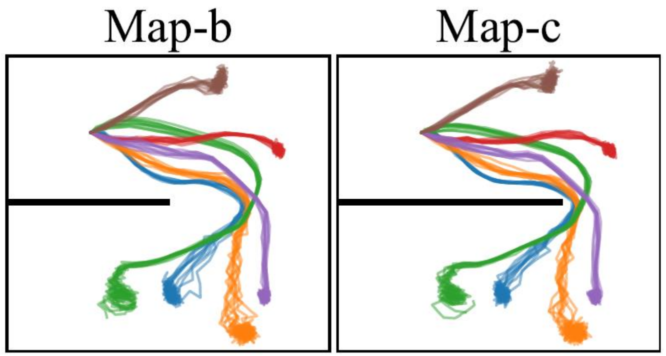
Unsupervised RL	Supervised RL
DIAYN; DADS; SMiRL; GPIM[1]	Off-Dynamics RL (DARC)
Cannot produce skills tailored to a new environment with dynamics shifts.	1. DARC requires prior reward function. maximize $-D_{\mathrm{KL}}(p_{\mathcal{P}_\mathcal{S},\pi_\theta}(\tau)\Vert p^*_{\mathcal{P}_\mathcal{T}}(\tau))$ 2. Our DARS is a *decoupled* objective. maximizing $-D_{\mathrm{KL}}(p_{\mathcal{P}_\mathcal{S},\pi_\theta}(\tau)\Vert p^*_{\mathcal{P}_\mathcal{S}}(\tau)) - \beta D_{\mathrm{KL}}(p_{\mathcal{P}_\mathcal{S},\pi_\theta}(\tau)\Vert p_{\mathcal{P}_\mathcal{T},\pi_\theta}(\tau))$

[1] Liu, Jinxin, et al. "Learn Goal-Conditioned Policy with Intrinsic Motivation for Deep Reinforcement Learning."

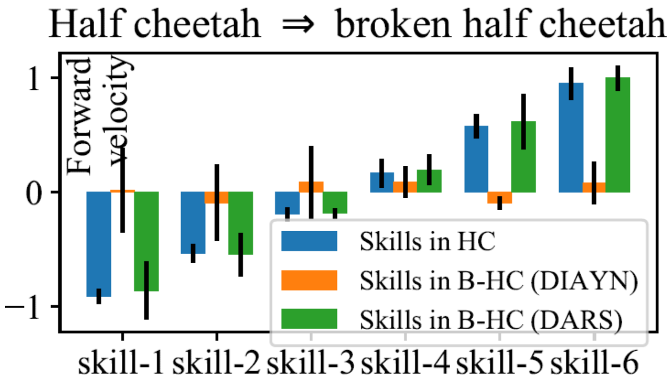
Experiments: Emergent Behaviors with DARS



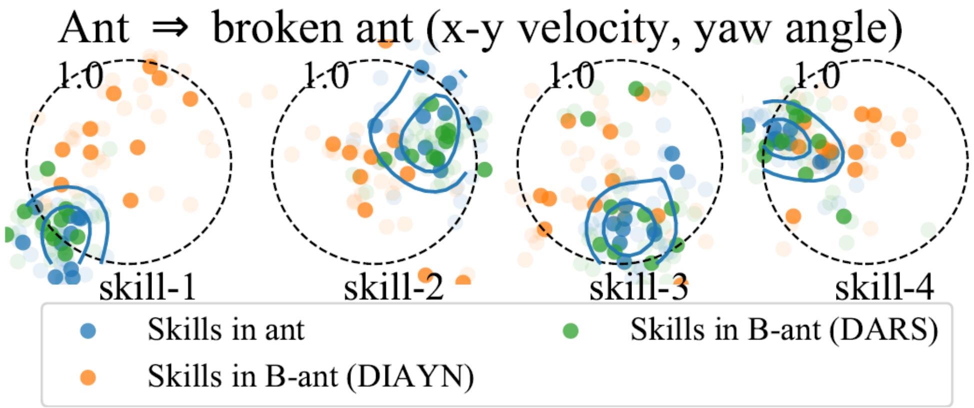
(a) *(Map-a, Map-b)*



(b) *(Map-b, Map-c)*

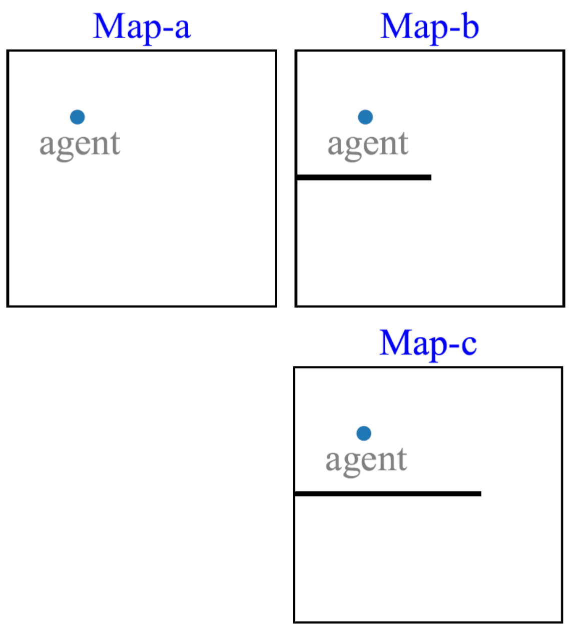


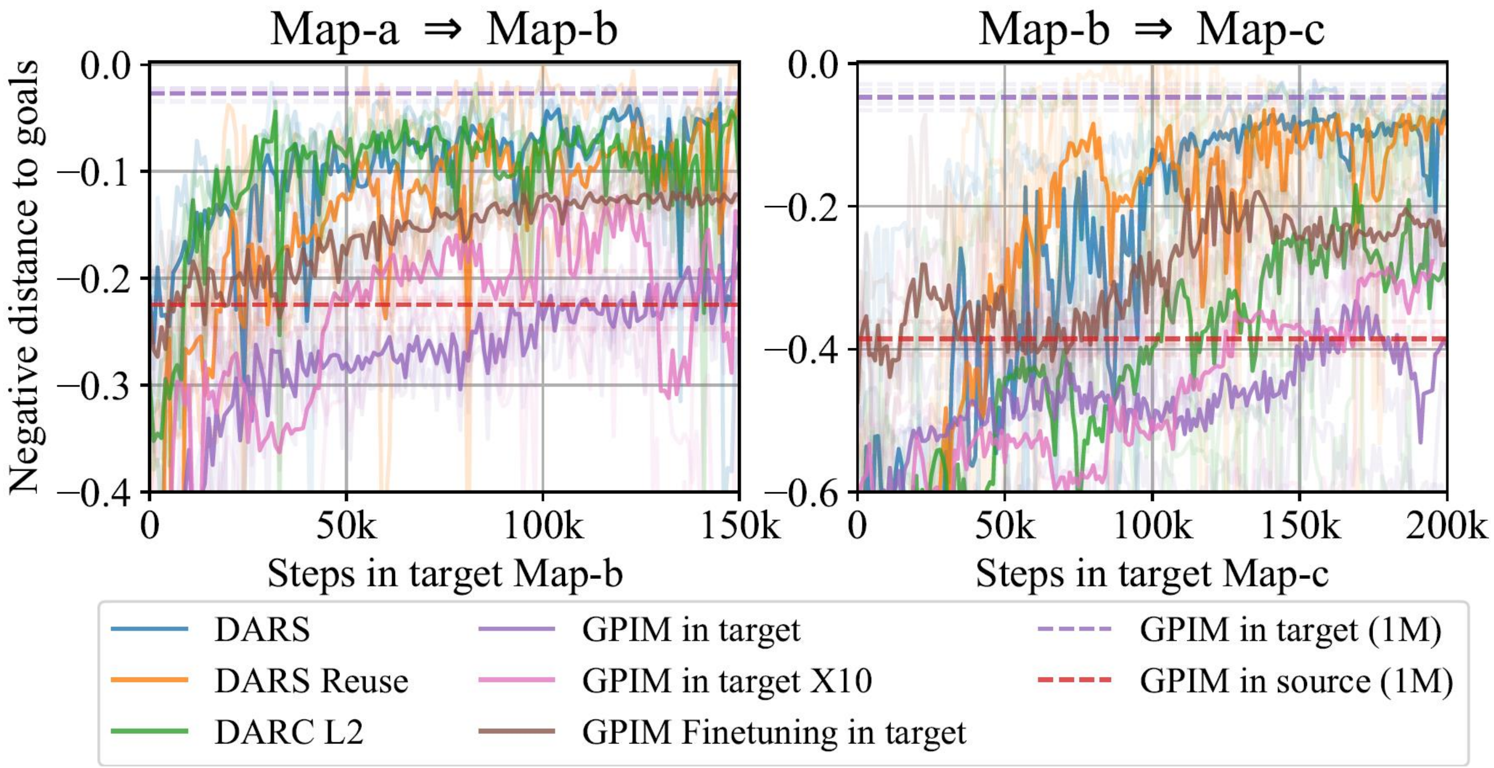
(c) *(HC, B-HC)*



(d) *(ant, B-ant)*

Experiments: Comparison with Baselines

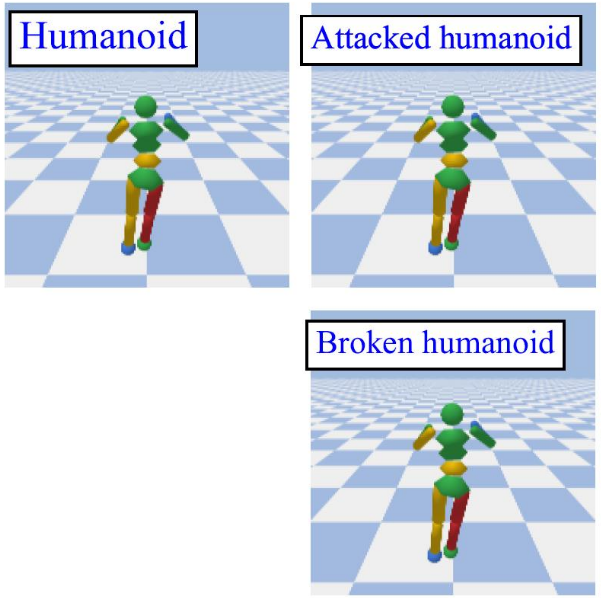


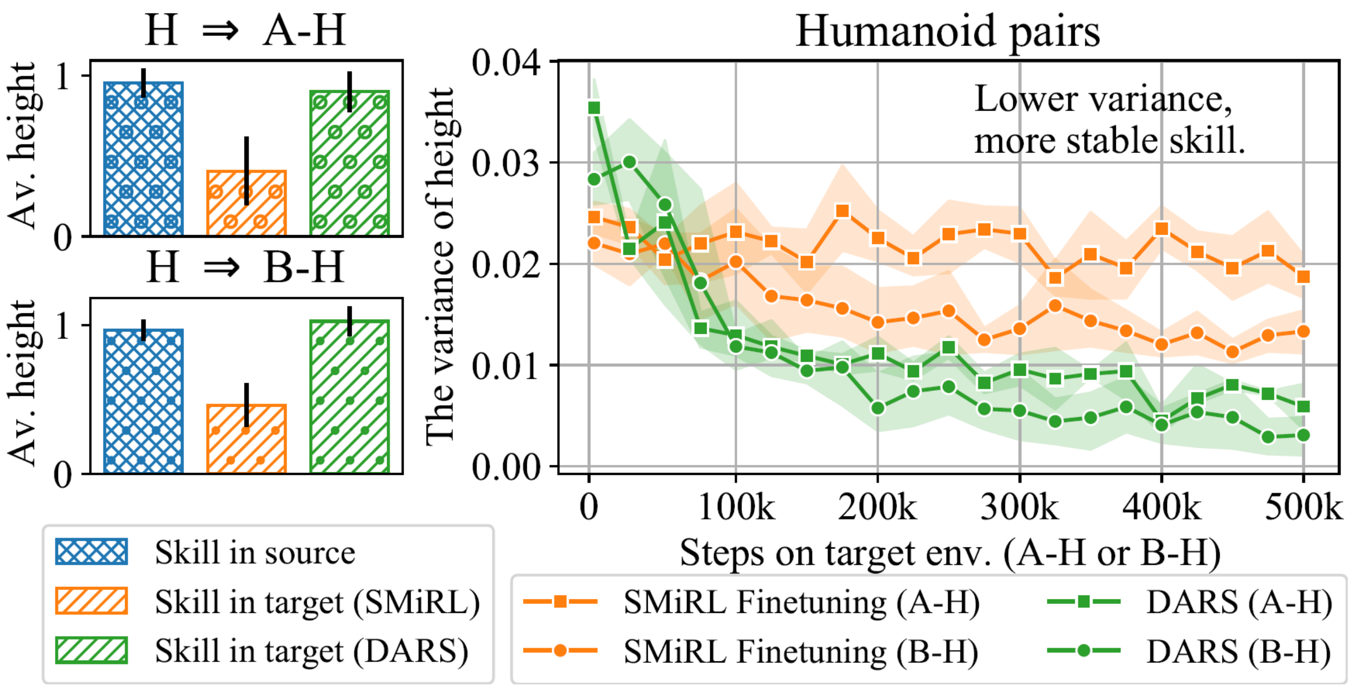


Stable environments

Our unsupervised DARS reaches comparable performance to (supervised) DARC L2.

Experiments: Comparison with Baselines



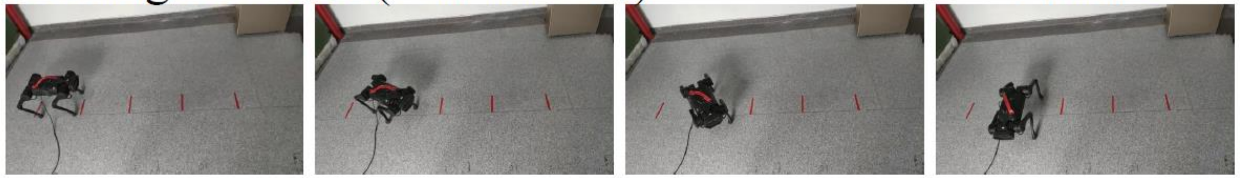


Unstable environments

With the same amount of rollout steps, we can find that DARS can learn a more stable skill for the target than SMiRL Finetuning.

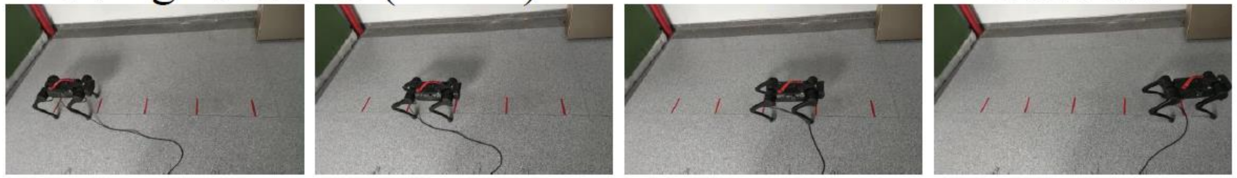
Experiments: Sim2real Transfer on Quadruped Robot

Moving forward. (*Full-in-real*) Failure ✗



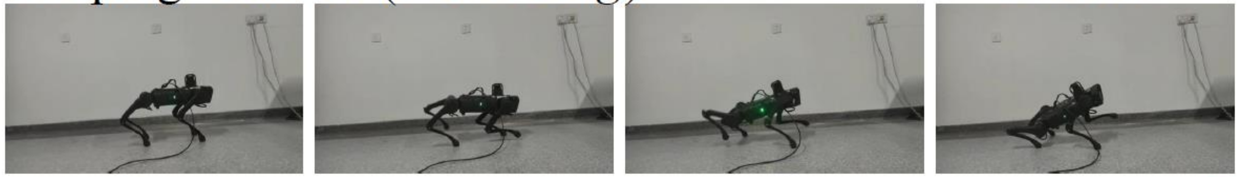
Moving forward. (*Finetuning*) Failure ✗

Moving forward. (*DARS*) Success ✓

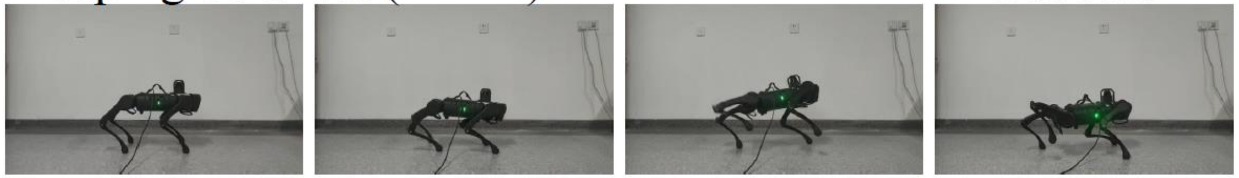


Keeping balance. (*Full-in-real*) Failure ✗

Keeping balance. (*Finetuning*) Failure ✗



Keeping balance. (*DARS*) Success ✓



1 Stable environments:

learn diverse skills (moving forward/backward)

2 Unstable environments:

keeping balance skill

	forward & backward	keeping balance
Full-in-real	> 6 h	> 6 h
Finetuning	> 6 h	4 h
DARS	3 h	1 h

Conclusion

1. we propose DARS to acquire adaptive skills for a target environment by training mostly in a source environment especially in the presence of dynamics shifts.

2. We show that our method obtains a near-optimal policy for target, as long as a mild assumption is met.

3. Experiments on a range of tasks confirm the effectiveness of our approach.