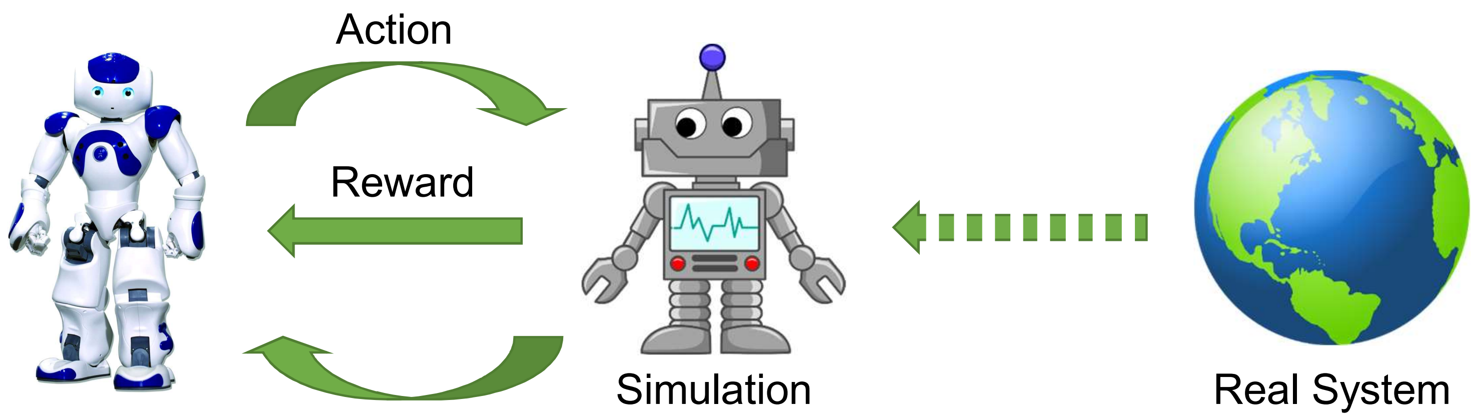
When to Trust Your Simulator: Dynamics-Aware Hybrid Offline-and-Online Reinforcement Learning

Haoyi Niu, Shubham Sharma, Yiwen Qiu, Ming Li, Guyue Zhou, Jianming Hu, Xianyuan Zhan

Real-World Challenges for RL



Visual Gap

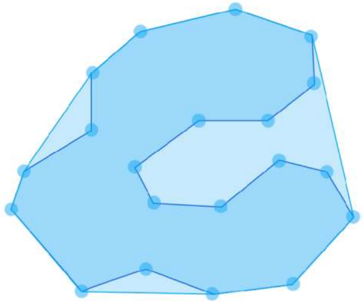
Simulation

Real World

Domain Randomization
Robust and transferable state-encoder
proper model setups, etc.

Dynamics Gap

Convex Surface

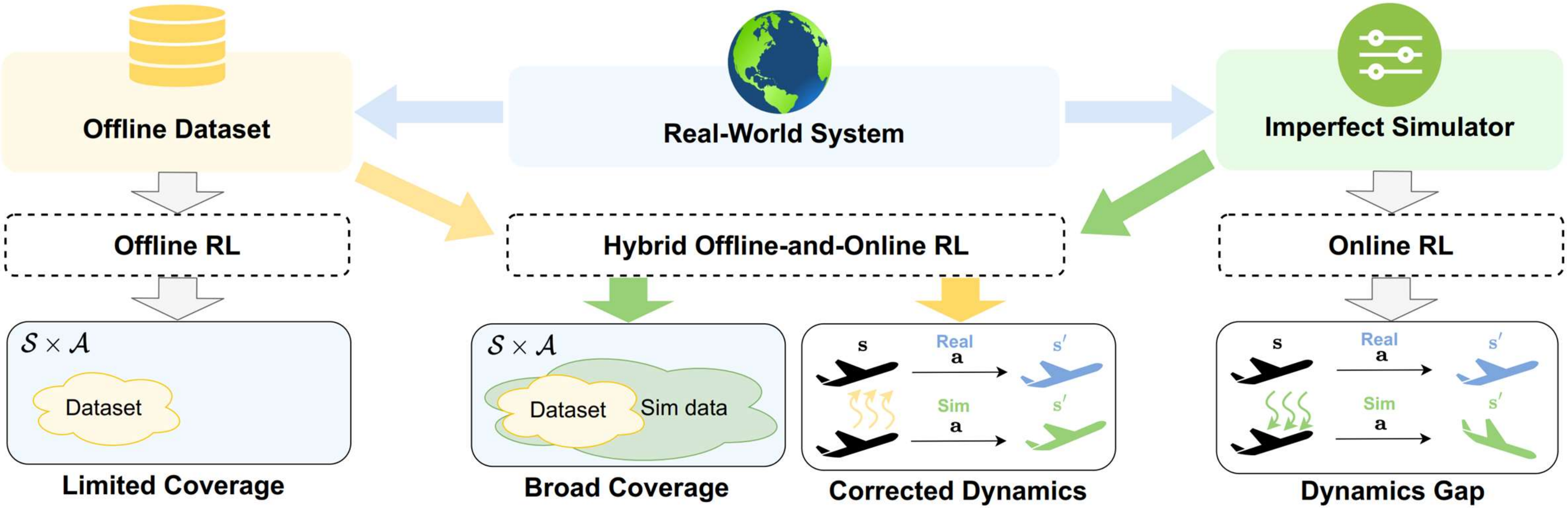


Rigid Body

causes a systematic bias & widely exist

Not easy to address

Hybrid Offline-and-Online RL (H2O)



Hybrid Offline-and-Online RL (H2O)

清華大学 智能产业研究院
Institute for AI Industry Research, Tsinghua University

NEURAL INFORMATION PROCESSING SYSTEMS

Dynamics-Aware Policy Evaluation:

$$\min_{Q} \beta \left(\log \sum_{\mathbf{s},\mathbf{a}} \omega(\mathbf{s},\mathbf{a}) \exp\left(Q(\mathbf{s},\mathbf{a})\right) - \mathbb{E}_{\mathbf{s},\mathbf{a}\sim\mathcal{D}}\left[Q(\mathbf{s},\mathbf{a})\right] \right) + \widetilde{\mathcal{E}}\left(Q, \hat{\mathcal{B}}^{\pi}\hat{Q}\right)$$

Minimize the dynamics-gap weighted soft-maximum of Q values:
Push down Q values on high dynamics-gap samples

Maximize Q values on data:
Pull up Q values on real offline data samples

$$\omega(\mathbf{s},\mathbf{a}) = u(\mathbf{s},\mathbf{a}) / \textstyle\sum_{\tilde{\mathbf{s}},\tilde{\mathbf{a}}} u(\tilde{\mathbf{s}},\tilde{\mathbf{a}})$$

$$\begin{aligned} u(\mathbf{s},\mathbf{a}) &:= D_{KL}\left(P_{\widehat{\mathcal{M}}} \| P_{\mathcal{M}}\right) \\ &\approx \sum_{\mathbf{s}'_i \sim P_{\widehat{\mathcal{M}}}(\mathbf{s}'_i|\mathbf{s},\mathbf{a})}^{N} \log \frac{P_{\widehat{\mathcal{M}}}(\mathbf{s}'_i|\mathbf{s},\mathbf{a})}{P_{\mathcal{M}}(\mathbf{s}'_i|\mathbf{s},\mathbf{a})} \\ &= \sum_{\mathbf{s}'_i \sim P_{\widehat{\mathcal{M}}}(\mathbf{s}'_i|\mathbf{s},\mathbf{a})}^{N} \log \left[\frac{1 - p(\mathrm{real}|\mathbf{s},\mathbf{a},\mathbf{s}')}{p(\mathrm{real}|\mathbf{s},\mathbf{a},\mathbf{s}')} \Big/ \frac{1 - p(\mathrm{real}|\mathbf{s},\mathbf{a})}{p(\mathrm{real}|\mathbf{s},\mathbf{a})} \right] \\ &= \sum_{\mathbf{s}'_i \sim P_{\widehat{\mathcal{M}}}(\mathbf{s}'_i|\mathbf{s},\mathbf{a})}^{N} \log \boxed{\left[\frac{1 - D_{sas}(\mathbf{s},\mathbf{a},\mathbf{s}')}{D_{sas}(\mathbf{s},\mathbf{a},\mathbf{s}')} \Big/ \frac{1 - D_{sa}(\mathbf{s},\mathbf{a})}{D_{sa}(\mathbf{s},\mathbf{a})} \right]} \end{aligned}$$

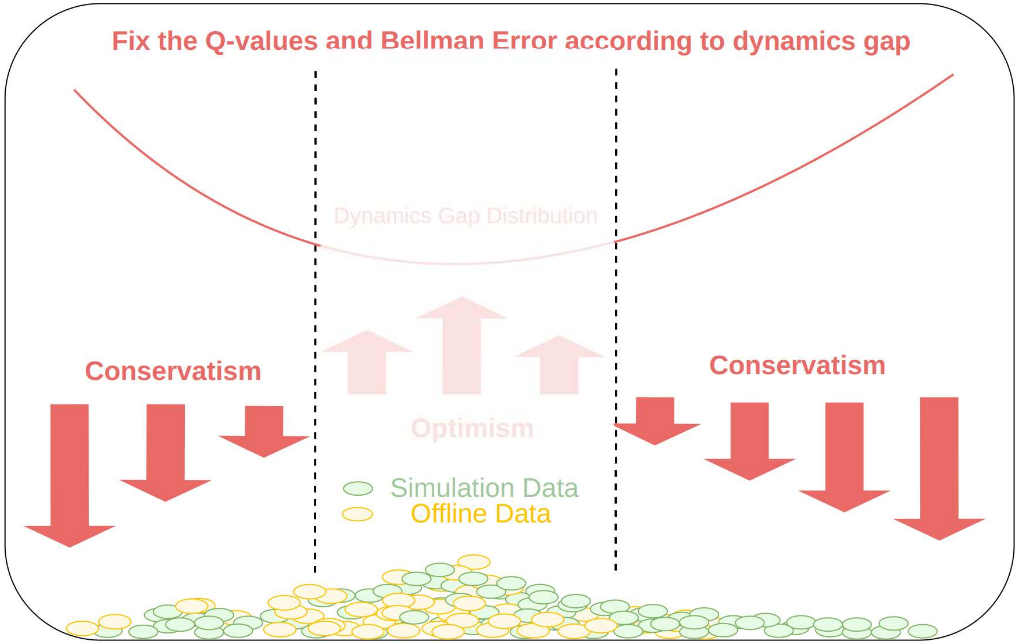
$$\widetilde{\mathcal{E}}\left(Q, \hat{\mathcal{B}}^{\pi}\hat{Q}\right) = \frac{1}{2}\mathbb{E}_{\mathbf{s},\mathbf{a},\mathbf{s}'\sim\mathcal{D}}\left[\left(Q - \hat{\mathcal{B}}^{\pi}\hat{Q}\right)(\mathbf{s},\mathbf{a})\right]^2 + \frac{1}{2}\mathbb{E}_{\mathbf{s},\mathbf{a},\mathbf{s}'\sim B}\left[\frac{P_{\mathcal{M}}(\mathbf{s}'|\mathbf{s},\mathbf{a})}{P_{\widehat{\mathcal{M}}}(\mathbf{s}'|\mathbf{s},\mathbf{a})}\left(Q - \hat{\mathcal{B}}^{\pi}\hat{Q}\right)(\mathbf{s},\mathbf{a})\right]^2$$

Learn on both offline data and online simulated samples

Fix Bellman error due to dynamics gap:
Use dynamics ratio as an importance sampling weight

Theoretical Interpretation

$$\min_{Q} \beta \left(\log \sum_{\mathbf{s},\mathbf{a}} \omega(\mathbf{s},\mathbf{a}) \exp\left(Q(\mathbf{s},\mathbf{a})\right) - \mathbb{E}_{\mathbf{s},\mathbf{a}\sim\mathcal{D}}\left[Q(\mathbf{s},\mathbf{a})\right] \right) + \widetilde{\mathcal{E}}\left(Q, \hat{\mathcal{B}}^{\pi}\hat{Q}\right)$$



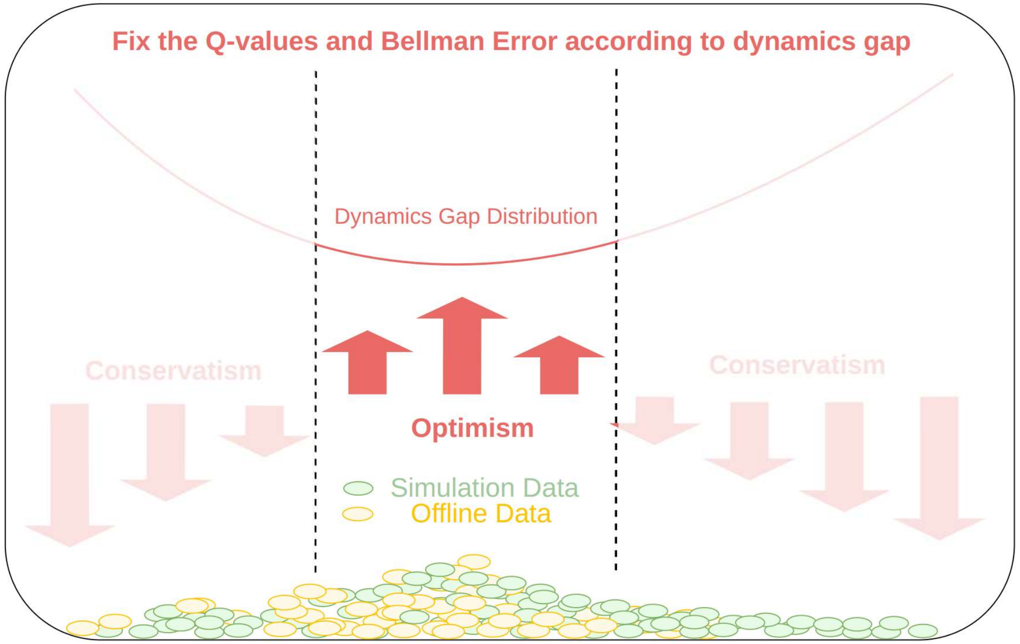
Adaptation but not Conservatism

Can be interpreted as adding an adaptive adjustment on Q-values:

$$\hat{Q}^{k+1}(\mathbf{s},\mathbf{a}) = (\hat{\mathcal{B}}^{\pi}\hat{Q}^{k})(\mathbf{s},\mathbf{a}) - \beta \left[\frac{\omega(\mathbf{s},\mathbf{a}) - d_{M}^{\pi_{\mathcal{D}}}(\mathbf{s},\mathbf{a})}{d_{M}^{\pi_{\mathcal{D}}}(\mathbf{s},\mathbf{a}) + d_{\widehat{M}}^{\pi}(\mathbf{s},\mathbf{a})} \right]$$

Theoretical Interpretation

$$\min_{Q} \beta \left(\log \sum_{\mathbf{s},\mathbf{a}} \omega(\mathbf{s},\mathbf{a}) \exp\left(Q(\mathbf{s},\mathbf{a})\right) - \mathbb{E}_{\mathbf{s},\mathbf{a}\sim\mathcal{D}}\left[Q(\mathbf{s},\mathbf{a})\right] \right) + \widetilde{\mathcal{E}}\left(Q, \hat{\mathcal{B}}^{\pi}\hat{Q}\right)$$

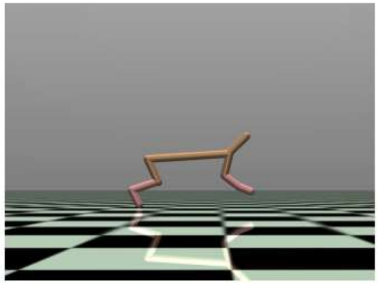
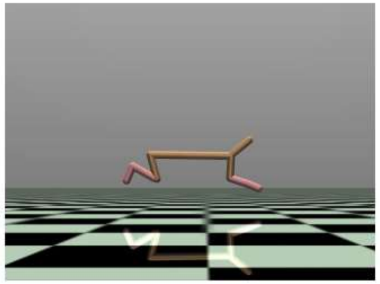
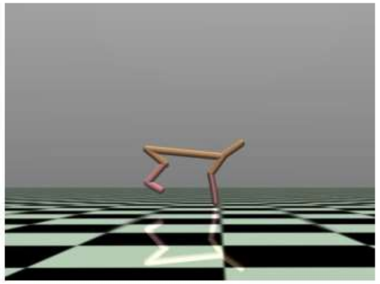
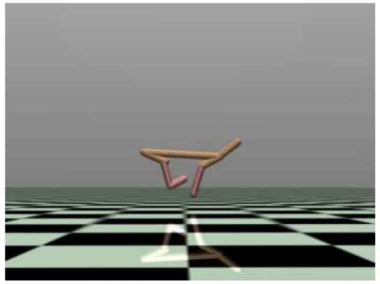


Adaptation but not Conservatism

Can be interpreted as adding an adaptive adjustment on Q-values:

$$\hat{Q}^{k+1}(\mathbf{s},\mathbf{a}) = (\hat{\mathcal{B}}^{\pi}\hat{Q}^{k})(\mathbf{s},\mathbf{a}) - \beta \left[\frac{\omega(\mathbf{s},\mathbf{a}) - d_{M}^{\pi_{\mathcal{D}}}(\mathbf{s},\mathbf{a})}{d_{M}^{\pi_{\mathcal{D}}}(\mathbf{s},\mathbf{a}) + d_{\widehat{M}}^{\pi}(\mathbf{s},\mathbf{a})} \right]$$

Simulation Experiments

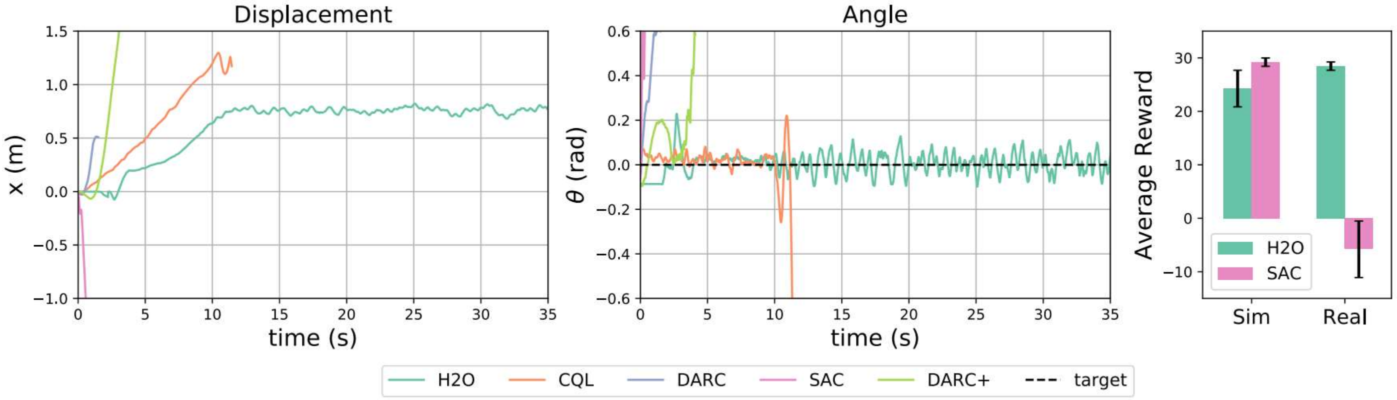
Dataset	Unreal Dynamics	SAC	CQL	DARC	DARC+	H2O
Medium	Gravity	4513±513	6066±73	5011±456	5706±440	**7085±416**
	Friction	2684±2646	6066±73	6113±104	6047±112	**6848±445**
	Joint Noise	4137±805	6066±73	5484±171	5314±520	**7212±236**
Medium Replay	Gravity	4513±513	5774±214	5105±460	4958±540	**6813±289**
	Friction	2684±2646	5774±214	5503±263	5288±100	**5928±896**
	Joint Noise	4137±805	5774±214	5137±225	5230±209	**6747±427**
Medium Expert	Gravity	4513±513	3748±892	**4759±353**	72±109	**4707±779**
	Friction	2684±2646	3748±892	**9038±1480**	7989±3999	6745±562
	Joint Noise	4137±805	3748±892	**5288±104**	733±767	**5280±1329**

MuJoCO HalfCheetah Environment

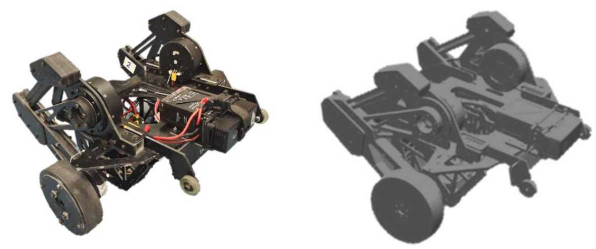
- **Offline dataset**: D4RL
- **Trained in simulation with unreal dynamics**: Gravity x2, Friction x0.3, Joint Noise N(0,1)
- **Evaluated in simulation with original dynamics:** Gravity x1, Friction x1, Joint Noise: none
- **Baseline**: SAC (online sim), CQL (offline real), DARC (online cross-domain, DARC+ (online+offline)

Real-World Validation

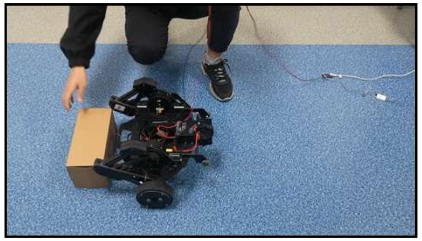
Standing Still



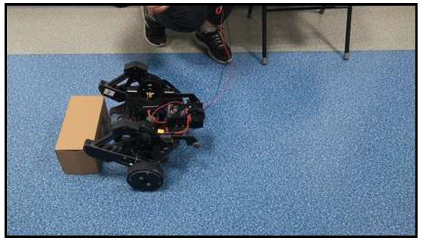
Wheel-Legged Robot



SAC

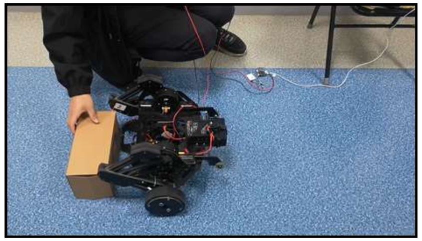


DARC

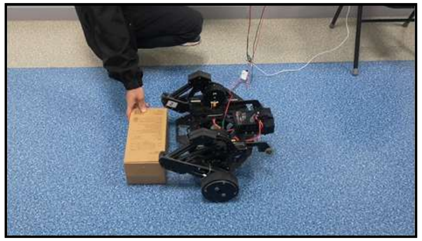


DARC+

CQL



H2O

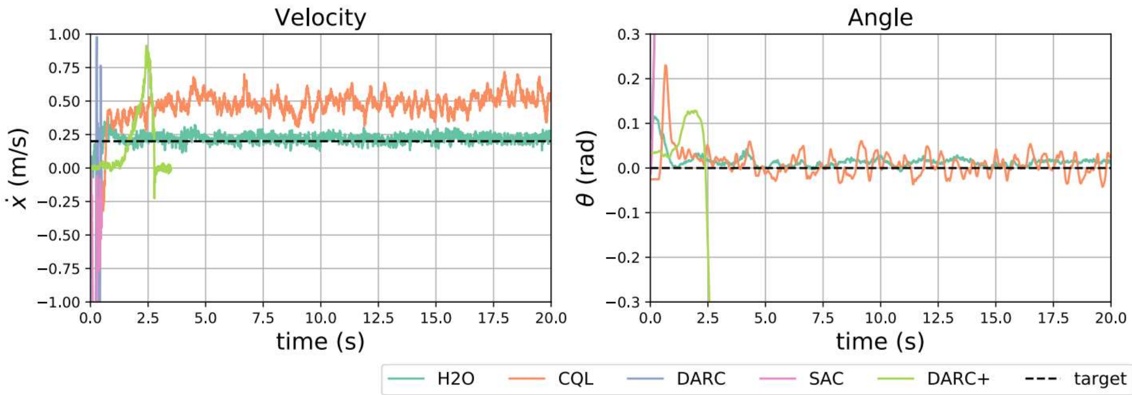


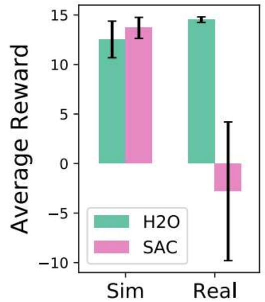
Real-World Validation

Institute for AI Industry Research, Tsinghua University

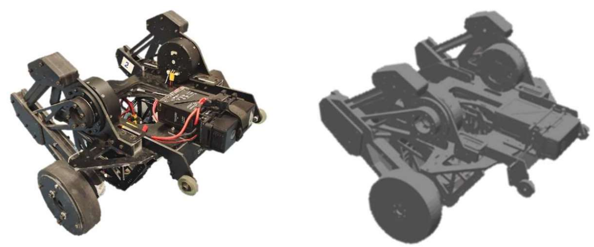
NEURAL INFORMATION PROCESSING SYSTEMS

Moving Straight

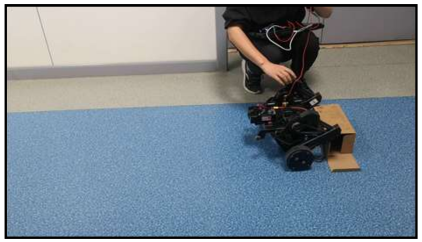
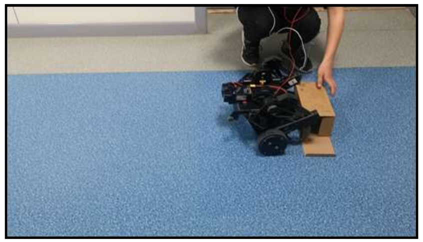




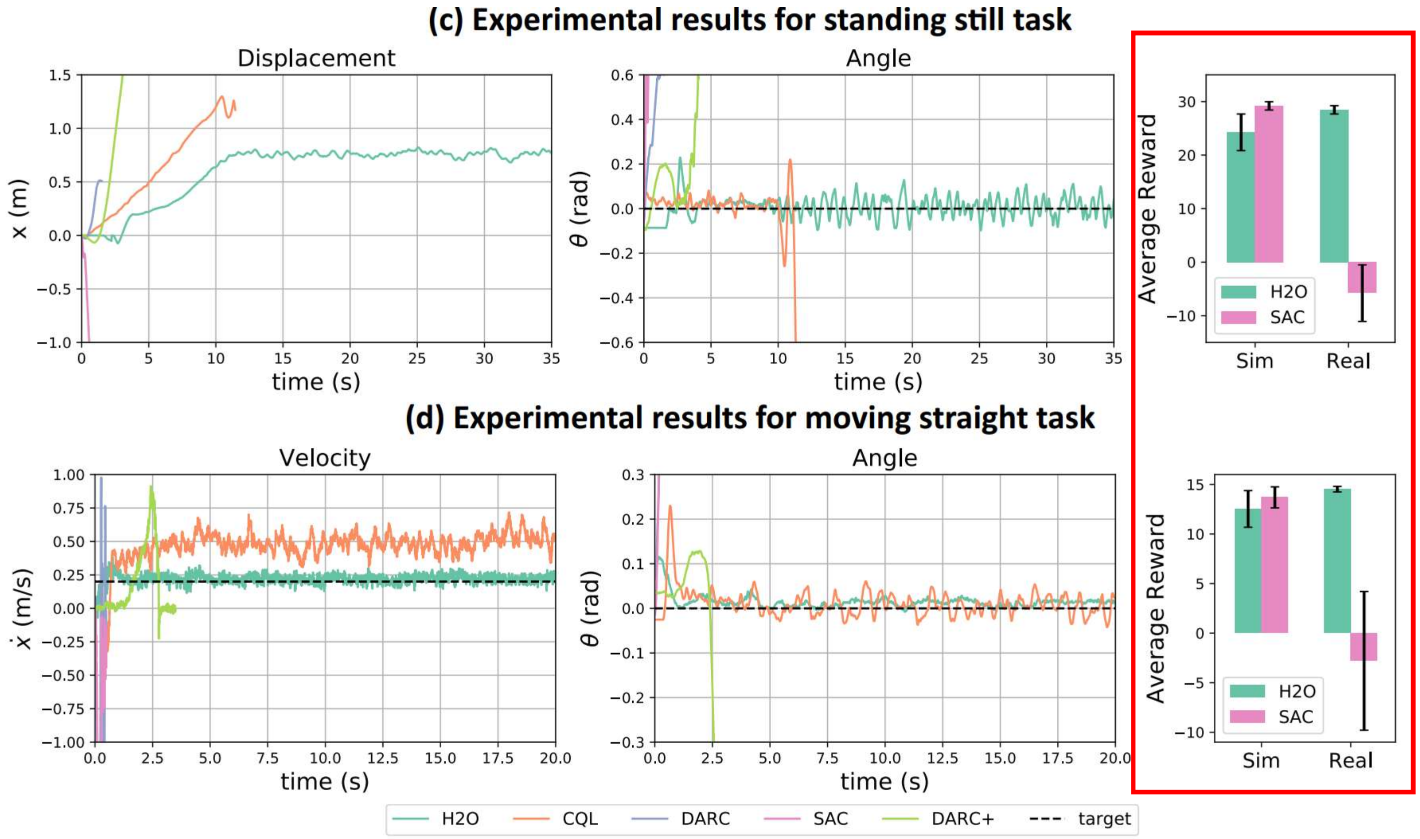
Wheel-Legged Robot



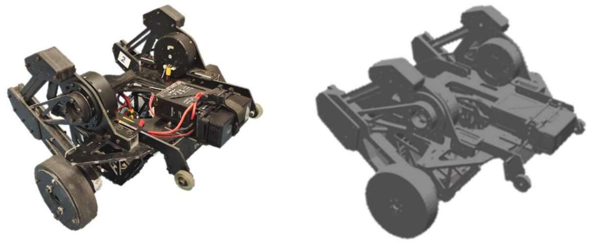
SAC **DARC** **DARC+** **CQL** **H2O**

Real-World Validation



Wheel-legged Robot



Rethinking: simulation-based evaluation can be very misleading, due to the dynamics gap.

When to Trust Your Simulator: Dynamics-Aware Hybrid Offline-and-Online Reinforcement Learning

Haoyi Niu, Shubham Sharma, Yiwen Qiu, Ming Li, Guyue Zhou, Jianming Hu, Xianyuan Zhan

Thanks!

https://github.com/t6-thu/H2O

https://www.youtube.com/watch?v=WRyEB6WEGc4

t6.da.thu@gmail.com