Towards an Information Theoretic Framework of Context-Based Offline Meta-Reinforcement Learning

NeurIPS 2024 Spotlight

Lanqing Li[1,2*], Hai Zhang[3*] , Xinyu Zhang[4] , Shatong Zhu[3] , Yang Yu[2] ,
Junqiao Zhao[3] , Pheng-Ann Heng[2]

1 Zhejiang Lab
2 The Chinese University of Hong Kong
3 Tongji University
4 Stony Brook University

Content

- Background
- Method
- Experiments

Background

Why Offline Meta-RL (OMRL)?

Offline RL		Meta-RL	
Safety	✓	Safety	✗
Sample Efficiency	✓	Sample Efficiency	✗
Adaptation	✗	Adaptation	✓
Generalization	✗	Generalization	✓

Background

- ## Context-Based Offline Meta-RL (COMRL)

Context-based OMRL (COMRL) seeks an optimal universal policy conditioning on a task representation $\boldsymbol{z^i}$ for any task/MDP M^i:

$$\pi(\boldsymbol{a}|\boldsymbol{s}, \boldsymbol{z}^i) = \arg\max_{\pi} \sum_{t=0}^{H-1} \gamma^t \mathbb{E}_{\boldsymbol{s}_t \sim \mu_\pi^t(\boldsymbol{s}), \boldsymbol{a}_t \sim \pi}[R^i(\boldsymbol{s}_t, \boldsymbol{a}_t)],\ \forall M^i$$

- ## Task Representation Learning in COMRL

Definition 1 *Given an input context variable $\boldsymbol{X} \in \mathcal{X}$ and its associated task/MDP random variable $\boldsymbol{M} \in \mathcal{M}$, task representation learning in COMRL aims to find a sufficient statistics $\boldsymbol{Z}$ of $\boldsymbol{X}$ with respect to $\boldsymbol{M}$.*

$$\boldsymbol{M} \longrightarrow \boldsymbol{X} \longrightarrow \boldsymbol{Z}$$

Markov Chain

Pre-existing Milestones

- **FOCAL**[1]

$$\mathcal{L}_{\text{FOCAL}} = \min_{\boldsymbol{\phi}} \mathbb{E}_{i,j} \left\{ \mathbb{1}\{i=j\} ||\boldsymbol{z}^i - \boldsymbol{z}^j||_2^2 + \mathbb{1}\{i \neq j\} \frac{\beta}{||\boldsymbol{z}^i - \boldsymbol{z}^j||_2^n + \epsilon} \right\}$$

- **CORRO**[2]

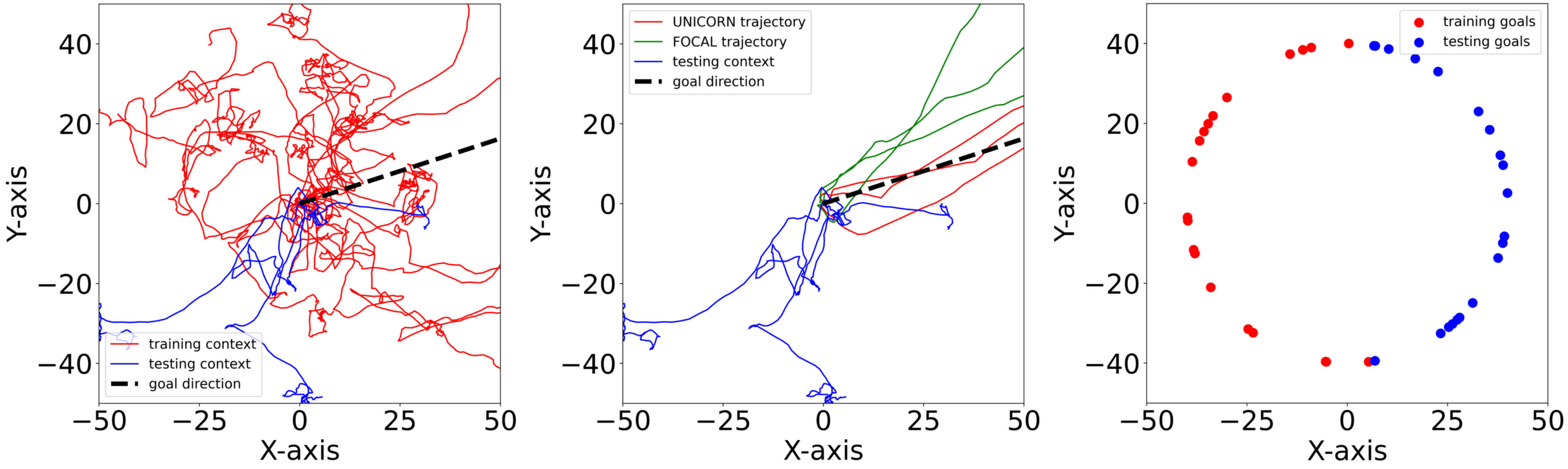
$$\mathcal{L}_{\text{CORRO}} = \min_{\boldsymbol{\phi}} \mathbb{E}_{\boldsymbol{x},\boldsymbol{z}} \left[-\log \left(\frac{h(\boldsymbol{x}, \boldsymbol{z})}{\sum_{M^* \in \mathcal{M}} h(\boldsymbol{x}^*, \boldsymbol{z})} \right) \right]$$

- **CSRO**[3]

$$\mathcal{L}_{\text{CSRO}} = \min_{\boldsymbol{\phi}} \{ \mathcal{L}_{\text{FOCAL}} + \lambda \mathbb{E}_i \left[\log q_{\boldsymbol{\phi}}(\boldsymbol{z}_i | \boldsymbol{s}_i, \boldsymbol{a}_i) - \mathbb{E}_j \left[\log q_{\boldsymbol{\phi}}(\boldsymbol{z}_j | \boldsymbol{s}_i, \boldsymbol{a}_i) \right] \right] \}$$

1. *Lanqing Li, Rui Yang, and Dijun Luo. Focal: Efficient fully-offline meta-reinforcement learning via distance metric learning and behavior regularization. ICLR 2021.*
2. *Haoqi Yuan and Zongqing Lu. Robust task representations for offline meta-reinforcement learning via contrastive learning. ICML 2022.*
3. *Yunkai Gao, et al. Context shift reduction for offline meta-reinforcement learning. NeurIPS 2023.*

Challenges

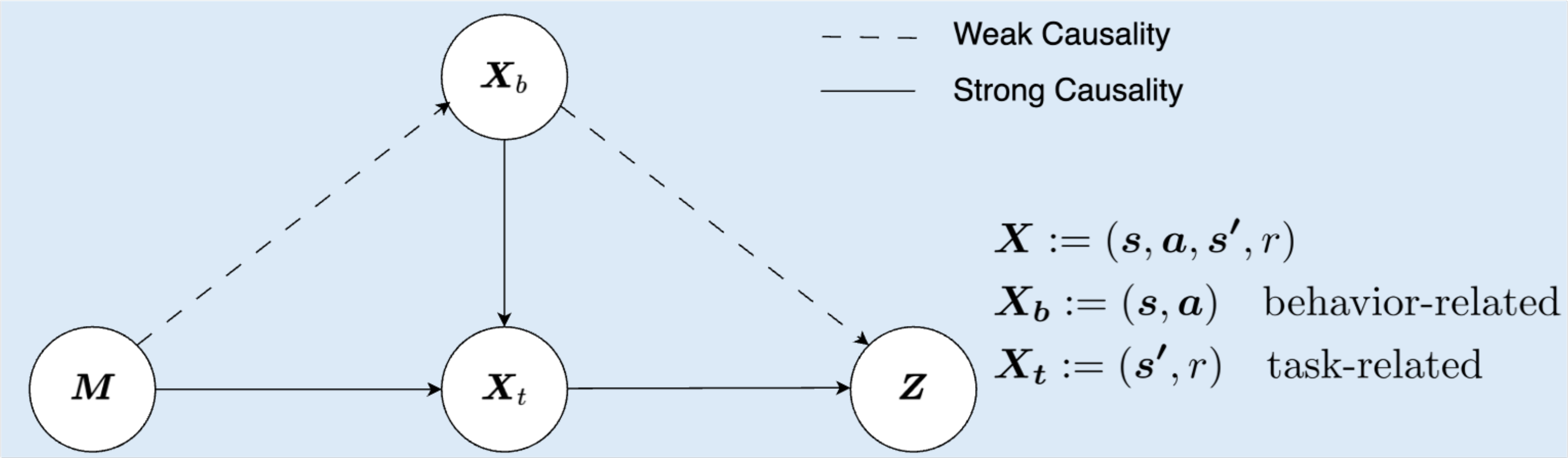


Context shift of COMRL. Since the offline training data are **static**, the agent could encounter severe context shift in state-action distribution (**left**) or task distribution (**right**) at test time.

Content

- Background
- Method
- Experiments

Decomposition of Input Data by Causality



$$I(\boldsymbol{Z}; \boldsymbol{X}) = \underbrace{I(\boldsymbol{Z}; \boldsymbol{X_t} | \boldsymbol{X_b})}_{\text{primary causality}} + \underbrace{I(\boldsymbol{Z}; \boldsymbol{X_b})}_{\text{lesser causality}}$$

The Central Theorem – An Information Theoretic Perspective

Theorem 1 (Central Theorem). *Let $\equiv$ denote equality up to a constant, then*

$$\underbrace{I(\boldsymbol{Z};\boldsymbol{X}_t|\boldsymbol{X}_b)}_{\text{primary causality}} \leq I(\boldsymbol{Z};\boldsymbol{M}) \leq I(\boldsymbol{Z};\boldsymbol{X}_t|\boldsymbol{X}_b)+I(\boldsymbol{Z};\boldsymbol{X}_b)= \underbrace{I(\boldsymbol{Z};\boldsymbol{X})}_{\text{primary + lesser causality}}$$

holds up to a constant, where

1. $\mathcal{L}_{\text{FOCAL}} \equiv -I(\boldsymbol{Z};\boldsymbol{X})$.
2. $\mathcal{L}_{\text{CORRO}} \equiv -I(\boldsymbol{Z};\boldsymbol{X}_t|\boldsymbol{X}_b)$.
3. $\mathcal{L}_{\text{CSRO}} \geq -\left((1-\lambda)I(\boldsymbol{Z};\boldsymbol{X})+\lambda I(\boldsymbol{Z};\boldsymbol{X}_t|\boldsymbol{X}_b)\right)$.

Take-away Message

$I(Z;M)$ operates as a unified learning objective and is robust to context shift, by trading off the primary and lesser causalities of COMRL.

Method

The Central Theorem offers ample implementation choices for $I(\boldsymbol{Z};\boldsymbol{M})$. This paper investigates 2 examples:

- **Supervised UNICORN**

$$
\begin{aligned}
\mathcal{L}_{\text{UNICORN-SUP}} &= I(\boldsymbol{Z};\boldsymbol{M}) \\
&\approx \underbrace{-\mathbb{E}_{\boldsymbol{x},\boldsymbol{z}\sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\sum_{j=1}^{n_M} \mathbb{1}(M^i = M)\log p_{\boldsymbol{\theta}}(M^i|\boldsymbol{z})\right]}_{\text{cross-entropy (predictive)}}
\end{aligned}
$$

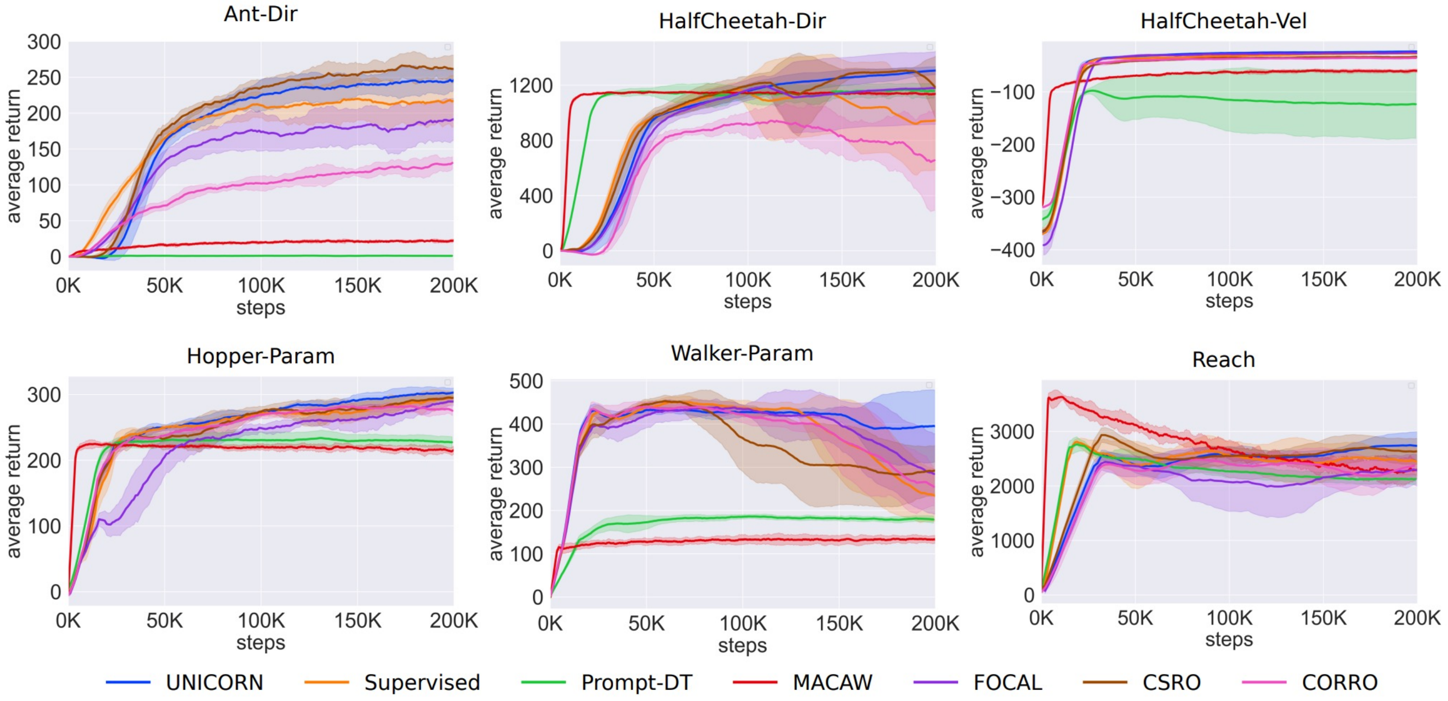
- **Self-Supervised UNICORN**

$$
\begin{aligned}
\mathcal{L}_{\text{UNICORN-SS}} &= \alpha I(\boldsymbol{Z};\boldsymbol{X}) + (1-\alpha) I(\boldsymbol{Z};\boldsymbol{X}_t|\boldsymbol{X}_b) \\
&\approx \underbrace{-\mathbb{E}_{\boldsymbol{x_t},\boldsymbol{x_b},\boldsymbol{z}\sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x_t},\boldsymbol{x_b})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x_t}|\boldsymbol{z},\boldsymbol{x_b})\right]}_{\text{reconstruction (generative)}} + \underbrace{\frac{\alpha}{1-\alpha}\mathcal{L}_{\text{FOCAL}}}_{\text{contrastive}}
\end{aligned}
$$

Content

- Background
- Method
- Experiments

Baseline Comparisons with IID/OOD Context Shift



Higher IID Performance

Higher Behavior-OOD Generalization Performance

Table 2: **Average testing returns of UNICORN against baselines on datasets collected by IID and OOD behavior policies.** Each result is averaged by 6 random seeds. The best is **bolded** and the second best is underlined.

Algorithm	HalfCheetah-Dir		HalfCheetah-Vel		Ant-Dir		Hopper-Param		Walker-Param		Reach	
	IID	OOD	IID	OOD	IID	OOD	IID	OOD	IID	OOD	IID	OOD
UNICORN-SS	**1307±26**	**1296±24**	**-22±1**	-94±5	267±14	236±18	**316±6**	**304±11**	**419±44**	**407±46**	**2775±241**	2604±183
UNICORN-SUP	1296±20	1130±76	-25±3	**-91±5**	250±4	**239±16**	312±4	302±12	322±28	312±39	2681±111	2641±140
CSRO	1180±228	458±253	-28±1	-102±5	**276±19**	233±12	310±6	301±10	310±58	279±65	2720±235	**2801±182**
CORRO	704±450	245±146	-37±3	-112±2	148±13	120±12	283±8	272±13	277±38	213±48	2468±175	2322±327
FOCAL	1186±272	861±253	-22±1	-97±2	217±29	173±24	302±4	297±13	308±98	286±91	2424±256	2316±303
Supervised	962±356	782±429	-24±1	-104±1	238±39	202±38	306±10	294±8	256±60	210±28	2489±248	2283±205
MACAW	1155±10	450±6	-56±2	-188±1	26±3	0±0	218±6	205±2	141±9	130±5	2431±157	1728±79
Prompt-DT	1176±40	-25±9	-118±66	-249±21	1±0	0±0	234±5	202±5	185±9	156±17	2165±85	1896±111

On Datasets of Varying Qualities

Table 3: **UNICORN vs. baselines on Ant-Dir datasets of various qualities.** Each result is averaged by 6 random seeds. The best is **bolded** and the second best is underlined.

Algorithm	Random		Medium		Expert	
	IID	OOD	IID	OOD	IID	OOD
UNICORN-SS	**81±18**	**62±6**	**220±23**	**243±10**	**279±10**	**262±13**
UNICORN-SUP	<u>75±15</u>	<u>60±5</u>	140±11	126±32	247±15	229±19
CSRO	2±3	0±1	166±10	<u>198±17</u>	<u>252±39</u>	202±45
CORRO	1±1	0±0	8±5	-7±2	-4±10	-14±9
FOCAL	67±26	44±10	<u>171±84</u>	187±86	229±42	<u>246±20</u>
Supervised	65±6	47±12	149±50	110±80	249±33	215±60
MACAW	3±1	0±0	28±2	1±1	88±43	1±1
Prompt-DT	1±0	0±0	2±4	0±1	78±15	1±2

Unanimous SoTA Performance on Random, Medium and Expert Data

Model-Agnostic (MLP → Decision Transformer[1-3])

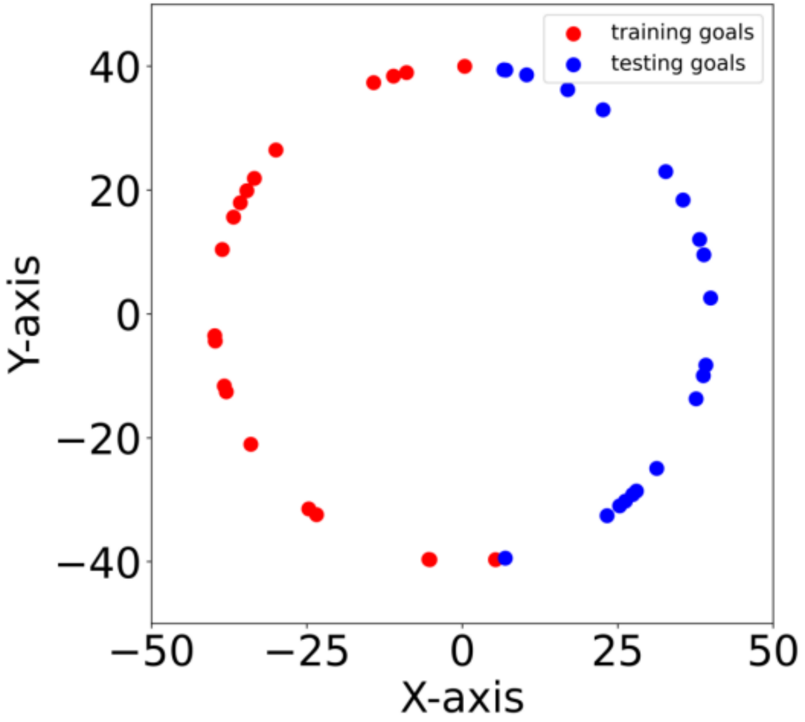
Table 4: **DT implementation of COMRL on HalfCheetah-Dir and Hopper-Param.** Each result is averaged by 6 random seeds.

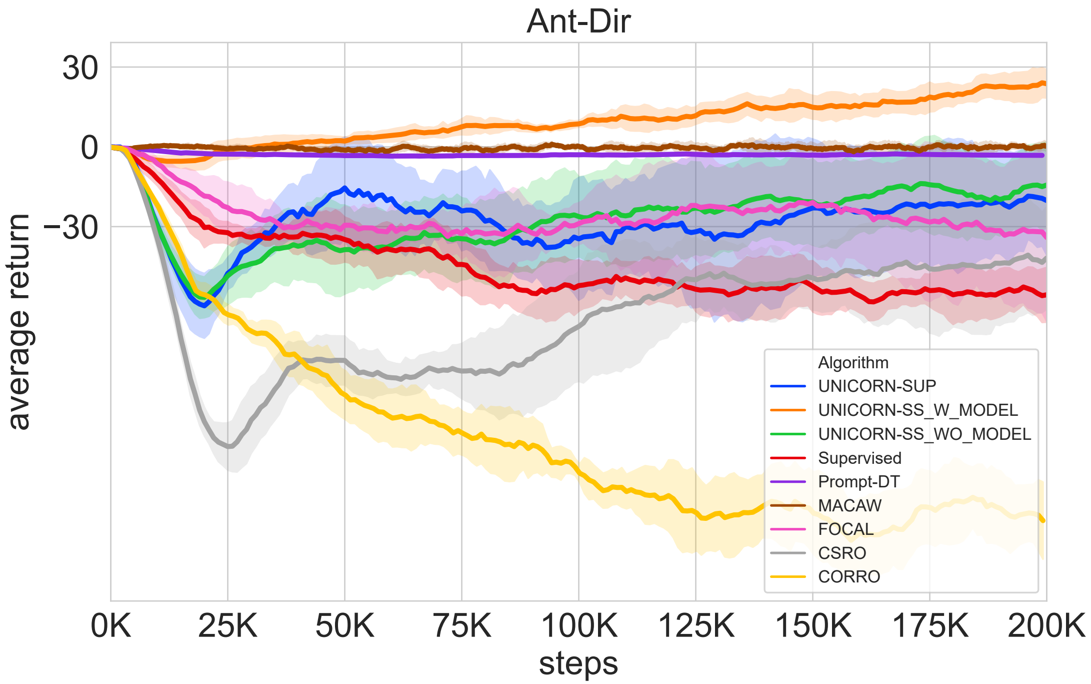
Algorithm	HalfCheetah-Dir		Hopper-Param	
	IID	OOD	IID	OOD
UNICORN-SS	1307±26	1296±24	316±6	304±11
UNICORN-SS-DT	1233±10	1186±43	304±4	291±4
UNICORN-SUP-DT	1227±21	1065±57	308±6	297±2
FOCAL-DT	1209±33	652±36	293±4	284±5
Prompt-DT	1177±40	-25±9	234±5	203±5

UNICORN is plug-and-play and transferrable across varying architectures

1. Chen, Lili, et al. "Decision transformer: Reinforcement learning via sequence modeling." NeurIPS 2021.
2. Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. NeurIPS 2021.
3. Xu, Mengdi, et al. "Prompting decision transformer for few-shot policy generalization." ICML 2022.

More Challenging Task-OOD Tests
—— Meta-Model-Enabled Model-Based RL





Meta-Model enables task-OOD (domain) generalization

Thank you for listening!

For more technical details, please refer to our paper:

ArXiv **Code** **OpenReview**

Poster Session

Date: Dec 11

Time: 4:30 p.m. — 7:30 p.m.

Place: West Ballroom A-D #6307