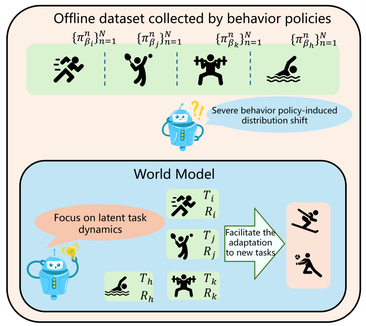
Behavior-Invariant Task Representation Learning with Transformer-based World Models for Offline Meta-Reinforcement Learning

Fuyuan Qian*[1], Menglong Zhang*[2], Song Wang[1], Quanying Liu[1]
[1]Southern University of Science and Technology, [2]The Hong Kong University of Science and Technology (Guangzhou)
mzhang943@connect.hkust-gz.edu.cn

ICML International Conference On Machine Learning

Motivation

- **Offline meta-RL** enables adaptation to unseen tasks using only static multi-task datasets.
- However, task inference suffers from **context distribution shift**: training contexts come from offline behavior policies, while test-time contexts are collected by the agent's online exploration policy.
- Meanwhile, **policy distribution shift** arises because offline data cover only limited state-action regions, causing value overestimation and unstable policy improvement.
- These challenges are amplified in **sparse-reward and out-of-distribution (OOD) settings**, where task-identifying signals are weak and offline trajectories provide incomplete evidence.



We propose **MetaSTAR**, a novel offline **Meta**-reinforcement learning framework with a **S**tochastic **T**ransformer-based world model for beh**A**vior-invariant task **R**epresentations

Contributions

- From an **information-theoretic perspective** on task representation, we propose MetaSTAR that leverages world models to encourage representations to capture transition-related and reward-related task dynamics while suppressing behavior-policy-specific information.
- We **theoretically validate** the effectiveness of world models in filtering the impact of behavior-policy-induced distribution shift during the meta-training stage, providing principled justification for behavior-invariant task representation learning.
- MetaSTAR **alleviates the inherent pattern dilemma** and achieves strong online adaptation performance across a wide range of sparse-reward and OOD tasks.

Information-Theoretic Task Representation

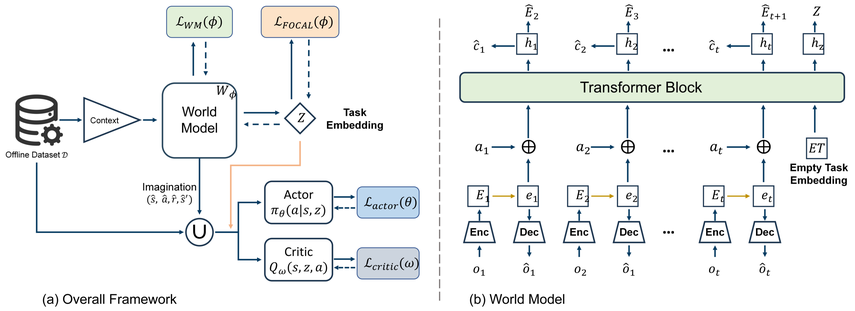
An ideal task representation Z should maximize the mutual information $I(Z;M)$ with the task variable M. Since the true task identity M is unobservable, we infer the task representation Z from the observed context X, decomposed into behavior-related components $X_b = (s,a)$ and task-related components $X_t = (r,s')$. This yields

$$I(Z;X) = \underbrace{I(Z;X_t \mid X_b)}_{\text{Primary Causality}} + \underbrace{I(Z;X_b)}_{\text{Lesser Causality}}.$$

$$I(Z;X_t \mid X_b) \le I(Z;M) \le I(Z;X_t \mid X_b) + I(Z;X_b).$$

Therefore, robustness to context shift can be promoted by **maximizing the primary causality $I(Z;X_t \mid X_b)$** while suppressing the lesser causality $I(Z;X_b)$.

Method



Latent Dynamics for Task Representation

- Observation $o_t = [s_t, r_{t-1}]$ (state and previous reward)
- VAE encoder maps observations to latent states e_t; decoder reconstructs $\hat{o}_t$.
- World model predicts future latent states $\hat{e}_{t+1}$ and continuation $\hat{c}_t$. The objectives:

$$\mathcal{L}_{\text{pred}}(\phi) \doteq -\log p_\phi(o_t \mid e_t) - \log p_\phi(c_t \mid h_t),$$
$$\mathcal{L}_{\text{dyn}}(\phi) \doteq \max\left(1, \mathrm{KL}\left[\mathrm{sg}\left(q_\phi(e_{t+1}|o_{t+1})\right) \,\|\, p_\phi(\hat{e}_{t+1}|h_t)\right]\right),$$
$$\mathcal{L}_{\text{rep}}(\phi) \doteq \max\left(1, \mathrm{KL}\left[q_\phi(e_{t+1}|o_{t+1}) \,\|\, \mathrm{sg}\left(p_\phi(\hat{e}_{t+1}|h_t)\right)\right]\right).$$
$$\mathcal{L}_{\text{WM}}(\phi) \doteq \mathbb{E}\left[\sum_{t=1}^{T}\left(\mathcal{L}_{\text{pred}}(\phi) + \mathcal{L}_{\text{dyn}}(\phi) + \beta_{\text{rep}}\mathcal{L}_{\text{rep}}(\phi)\right)\right].$$

- Learnable query token aggregates the latent history via self-attention to produce a global context embedding h_z; an head projects h_z to the task embedding z.
- FOCAL metric learning makes z discriminative and task-sufficient for the task variable M. It is defined as:

$$\mathcal{L}_{\text{FOCAL}}(\phi) \doteq \mathbb{E}_{i,j}\left[\mathbf{1}_{\{i=j\}}\left\|z^i - z^j\right\|_2^2 + \mathbf{1}_{\{i\neq j\}}\frac{1}{\|z^i - z^j\|_2^2 + \epsilon}\right].$$

Conservative Policy Learning with Contextual Imagination

- Contextual imagined rollouts $\hat{d}_\phi = (\hat{s},\hat{a},\hat{r},\hat{s}')$ augment real transitions d to form $d_f \doteq d \cup \hat{d}_\phi$.
- Conservative critic penalizes unsupported state-action values.

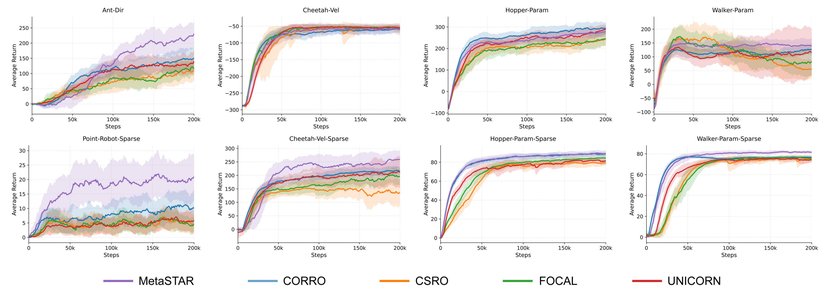
$$\mathcal{L}_{\text{critic}}(\omega) \doteq \beta\left(\mathbb{E}_{s,a\sim\rho}[Q_\omega(s,z,a)] - \mathbb{E}_{s,a\sim d}[Q_\omega(s,z,a)]\right) + \frac{1}{2}\mathbb{E}_{s,a\sim d_f}\left[\left(Q_\omega(s,z,a) - \widehat{\mathcal{B}}^\pi Q_\omega(s,z,a)\right)^2\right],$$

Theoretical Analysis of Behavior-Invariant Representation

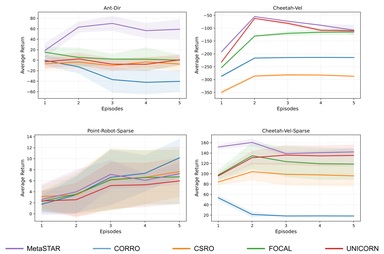
World model dynamics as primary causality: optimizing the latent dynamics of the world model $\mathcal{L}_{\text{dyn}}(\phi) \doteq \mathrm{KL}\left[\mathrm{sg}\left(q_\phi(e_{t+1}|o_{t+1})\right) \,\|\, p_\phi(\hat{e}_{t+1}|h_t)\right]$ can be viewed as maximizing the primary causality $I(Z;X_t \mid X_b)$.

Experiments

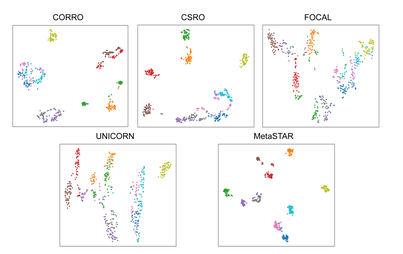
Online Testing Performance



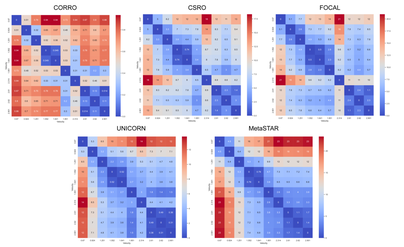
OOD Online Adaptation



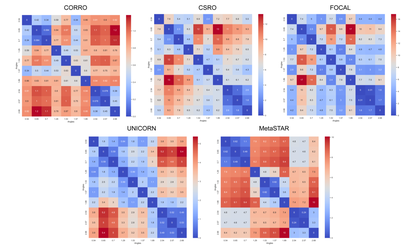
T-SNE Visualization (Hopper-Param)



Euclidean Distance (Cheetah-Vel-Sparse)



Euclidean Distance (Point-Robot-Sparse)



Ablation Study