CATAL: Causally Disentangled Task Representation Learning for Offline Meta Reinforcement Learning

Shan Cong[1,2], **Chao Yu**[1,2*], **Xiangyuan Lan**[1*]

[1]Sun Yat-sen University
[2]Pengcheng Laboratory
congsh5@mail2.sysu.edu.cn, yuchao3@mail.sysu.edu.cn, lanxy@pcl.ac.cn

Abstract

Context-based Offline Meta Reinforcement Learning (COMRL) has shown promising results in improving the cross-task generalization ability of meta-policies. However, current methods often lead to entangled task representations, in which each latent dimension is influenced by multiple causal factors that govern variations in environment dynamics and reward mechanisms. This entanglement can degrade generalization performance, particularly when multiple causal factors vary simultaneously across tasks. To address this limitation, we propose **CA**usally disentangled **TA**sk representation **L**earning (**CATAL**) method for COMRL that aims to improve the generalization ability of the meta-policy, where each latent dimension in the task representations aligns to a single causal factor. Theoretically, we show that under mild conditions, the task representations learned by CATAL are causally disentangled. Empirically, extensive results on multi-task MuJoCo benchmarks show that CATAL consistently outperforms existing COMRL baselines in both in-distribution and out-of-distribution generalization.

Introduction

Meta-reinforcement learning (Meta-RL) (Beck et al. 2023; Nagabandi et al. 2018; Duan et al. 2016) aims to develop meta-policies capable of rapid adaptation to new tasks by leveraging prior experience from multiple related tasks. This paradigm has shown strong potential in domains such as robotics, healthcare, and recommendation systems. However, the reliance on frequent online interactions limits the practicality of Meta-RL methods in real-world scenarios where environment access is expensive, constrained, or potentially unsafe. To mitigate this limitation, offline meta-RL (OMRL) has emerged as a promising alternative that enables meta-policies to perform task adaptation solely on pre-collected offline datasets (Mitchell et al. 2021; Zhao et al. 2022). A prominent line of research in OMRL is context-based OMRL (COMRL) (Gao et al. 2024; Li et al. 2024), which leverages a context encoder to infer task representations from pre-collected trajectories and conditions the meta-policy on these representations for task adaptation.

Although COMRL demonstrates significant potential in task generalization, task representations learned by existing methods often suffer from causal entanglement(see Fig. 1). Specifically, the context encoder often associates a single task representation dimension with multiple causal factors, which govern variations in both environment dynamics and reward mechanisms. As a result, the learned representation does not clearly reveal the specific configuration of each individual causal factor. For example, consider a series of robot navigation tasks where the robot must navigate to different target locations under varying ground conditions. The differences between tasks arise from two causal factors: ground friction and target location. Ground friction directly influences the environmental dynamics, while the target location determines the reward mechanism of the task. In a causally entangled representation, a single latent dimension may simultaneously encode information from both causal factors, making it difficult to determine whether changes in the task representation are due to adjustments in friction or shifts in the target location. This ambiguity in the representation prevents the meta-policy from accurately identifying the precise causal configuration of each task, thereby hindering its ability to adapt effectively to new task settings.

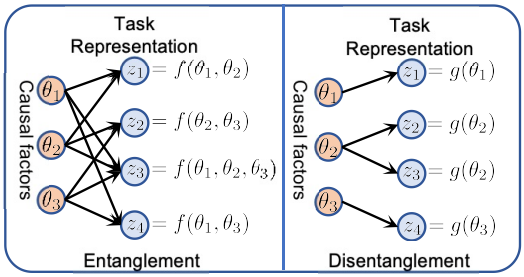

Figure 1: Causal entanglement in task representations mixes multiple causal factors within a single dimension, whereas causal disentanglement isolates each factor into its own independent dimension.

To tackle these challenges, we propose **CA**usally disentangled **TA**sk representation **L**earning (**CATAL**), a novel COMRL method that trains a context encoder to extract causally disentangled task representations, where each latent dimension exclusively corresponds to a single causal factor. To this end, CATAL introduces three key components: (1) an *independence constraint* to ensure non-overlapping in-

*Corresponding authors.

formation across dimensions; (2) a *consistency constraint* to enforce invariance of dimensions corresponding to unchanged causal factors, guided by a lightweight *soft intervention indicator* that estimates causal factor changes across tasks; and (3) a *reconstruction objective* with a *remapping mechanism* to encourage the decoder to reconstruct transitions using task representations where dimensions corresponding to unchanged causal factors are replaced with their cross-task averages. By integrating these components, CATAL infer causally disentangled task representations from pre-collected trajectories, which significantly enhance the generalization ability of the meta-policy.

Theoretically, we prove that CATAL recovers causally disentangled task representations under mild conditions. Empirically, CATAL consistently outperforms competitive COMRL approaches on multi-task MuJoCo benchmarks and demonstrates strong generalizability across both in-distribution and out-of-distribution tasks.

Related Works

OMRL. OMRL methods can be broadly categorized into gradient-based and context-based paradigms. The gradient-based OMRL (Mitchell et al. 2021) adapts to novel tasks by updating the meta-policy via gradient descent on a task-specific loss. In contrast, the COMRL (Dorfman, Shenfeld, and Tamar 2021; Pong et al. 2022) avoids gradient updates by leveraging task representation learning. In COMRL, a prevalent line of work (Li, Yang, and Luo 2020; Yuan and Lu 2022) employs contrastive objectives to pull together representations inferred from trajectories of the same task while pushing apart those derived from different tasks in the representation space. An alternative approach (Zhou et al. 2024) introduces generative models to ensure that the learned representations capture the underlying features of environment dynamics and reward mechanisms. Despite their effectiveness, their performance often degrades significantly when test tasks differ from training tasks along multiple task factors, revealing limitations in their generalization ability. To address this issue, we explicitly align each dimension (or group of dimensions) of the task representation with task factors so that the meta-policy can accurately understand task identity and thereby enhance its generalization capability across tasks.

Causal Inference for Policy Generalization. To improve policy generalizability to unseen tasks, recent work increasingly leverages causal inference to identify and exploit causal factors from experience, forming a promising direction in RL. Existing approaches can be broadly categorized into active and passive causal inference. Active approaches aim to explicitly uncover causal factors by performing interventions—either by executing actions and analyzing the resulting feedback (Sontakke et al. 2021; Cai et al. 2024), or by modifying causal variables in a controlled simulation environment to observe their effects (Li et al. 2020). Passive approaches (Dunion et al. 2024; Yuan, Lu, and Liu 2024) infer causal factors directly from dynamically evolving experience without explicit interventions, typically relying on a structural causal model (SCM) to uncover causal relationships. In contrast, our approach operates entirely in an offline setting, where causal inference is performed using static experience data collected from multiple tasks, without any further interaction with the environment. In addition, the most closely related work is CausalCOMRL (Zhang et al. 2025), which is designed for offline settings. However, CausalCOMRL relies on annotated task factors as supervision, while our approach does not require such annotations—we only know that tasks differ, without any knowledge of the specific nature of these differences.

Preliminaries

In this section, we formalize the offline policy generalization problem under the framework of Factored Markov Decision Processes (FMDPs). This framework provides a unified formulation for multi-task offline datasets and policy generalization objectives. Building on this foundation, we introduce the COMRL paradigm, which enables policy generalization to unseen tasks using only offline data.

Factored Markov Decision Process (FMDP). A Factored Markov Decision Process (FMDP) is defined as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \Theta, \mathcal{T}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, $\gamma \in (0, 1)$ is the discount factor, and $\Theta = \{\Theta_i\}_{i=1}^N$ represents the space of task causal factors. The transition dynamics $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \Theta \to \Delta(\mathcal{S})$ and reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \Theta \to \mathbb{R}$ are parameterized by task-specific factors $\theta = \{\theta_i\}_{i=1}^N \in \Theta$. Different tasks correspond to different instantiations of θ, inducing variations in the underlying environment dynamics and rewards.

Offline Policy Generalization Problem. We study offline policy generalization in the FMDP setting. Given a multi-task offline dataset $\mathcal{D} = \{\mathcal{D}_\theta \mid \theta \sim p_{\text{train}}(\theta)\}$, where each task is specified by a set of task factors $\theta = \{\theta_i\}_{i=1}^N \in \Theta$, and each task dataset $\mathcal{D}_\theta = \{(s_t, a_t, r_t, s_{t+1})\}_{t=0}^{T_\theta - 1}$ is collected by a behavior policy without further environment interaction, the goal is to learn a task-conditioned policy $\pi(a \mid s, \theta)$ that generalizes beyond the training task distribution. For a test task with factors $\theta_{\text{test}} \sim p_{\text{test}}(\theta)$, the policy is evaluated by its expected discounted return:

$$\hat{\pi} = \arg\max_{\pi} \mathbb{E}_{s_0 \sim \rho_{\theta_{\text{test}}},\, a_t \sim \pi(\cdot \mid s_t, \theta_{\text{test}})} \left[\sum_{t=0}^{\infty} \gamma^t r_t \right]. \quad (1)$$

COMRL Paradigm. The COMRL paradigm addresses offline policy generalization by learning task representations from offline datasets and conditioning a meta-policy on these representations. Given a context window $c_t = \{(s_{t-k}, a_{t-k}, r_{t-k}, s_{t-k+1})\}_{k=1}^K$, a context encoder f infers a latent task representation $z_t = f(c_t)$, which captures task-relevant factors. The meta-policy and value functions are then conditioned on z_t: $\pi(a \mid s, z_t)$, $V(s, z_t)$, and $Q(s, a, z_t)$. These functions are optimized using offline reinforcement learning objectives to maximize the expected return across tasks. At test time, the learned meta-policy adapts to a new task by inferring its task representation z_t from offline context, enabling generalization without additional interaction with the environment.

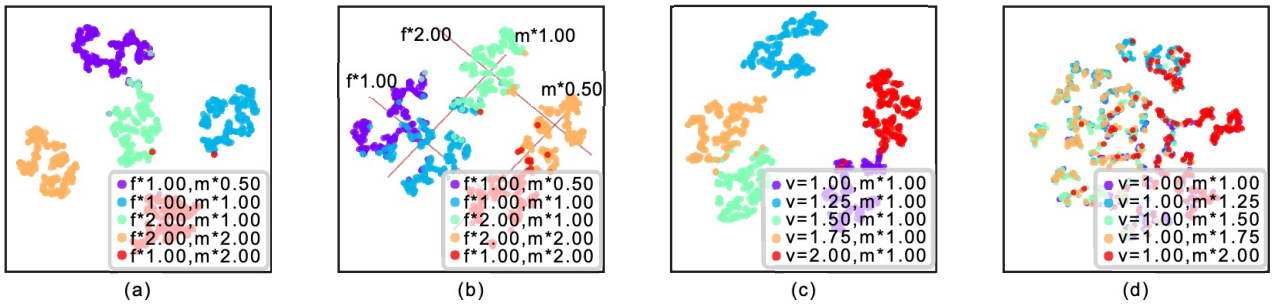

Figure 2: Task representation visualizations of CORRO and CATAL on the HalfCheetah benchmark. (a–b) Task representations learned by CORRO and CATAL under joint variations of body mass and ground friction. (a) CORRO produces entangled representations that do not separate the two causal factors. (b) CATAL successfully separates the underlying causal factors, forming a structured representations aligned with independent variations in mass and friction. (c–d) Task representations learned by CORRO under single-factor variations. (c) When only the target velocity changes, CORRO captures clear distinctions between tasks. (d) When only body mass changes, the representations exhibit substantial overlap, indicating limited sensitivity to this factor.

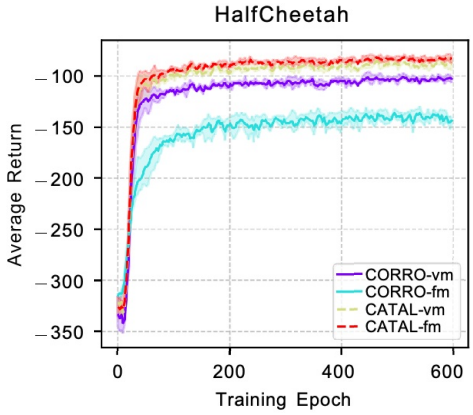

Figure 3: Performance comparison under multi-factor task variations on the HalfCheetah benchmark. CATAL achieves higher performance than CORRO.

Methods

Causal Entanglement Problem in COMRL

A fundamental challenge in COMRL is the causal entanglement of learned task representations. Ideally, each representation dimension should correspond to an independent causal factor (e.g., target velocity or object mass) to support systematic generalization. However, existing methods such as CORRO often learn entangled embeddings, where multiple causal factors are mixed within individual dimensions. As shown in Fig. 2(a–b), although CORRO can distinguish tasks, the variations in its embedding space are not aligned with ground-truth causal changes, resulting in representations that are discriminative but not causally interpretable. This entanglement limits generalization to complex task compositions. Moreover, Fig. 2(c–d) shows that CORRO captures variations in target velocity but fails to encode body mass, leading to overlapping representations and degraded performance when multiple factors vary jointly (Fig. 3). Such causal entanglement is not unique to CORRO but is common across COMRL methods (Li, Yang, and Luo 2020; Gao et al. 2024; Rakelly et al. 2019), highlighting a fundamental limitation in disentangling task-level causal factors from offline context data.

Causally Disentangled Task Representation Learning

We introduce CATAL to address task representation entanglement in COMRL. CATAL employs a context encoder together with three key components: an independence constraint that reduces statistical dependence across latent dimensions, a consistency constraint that aligns each dimension with its corresponding causal factor, and a reconstruction objective with remapping mechanism that enhances representation completeness. This framework enables the learned task representations to accurately reflect the underlying causal factors.

Context Encoder. We define the context encoder as $q_{\phi_z}(z \mid x, y)$, which takes transition tuples $\langle x, y\rangle$ as input, where $x = (s, a)$ denotes a state–action pair and $y = (s', r)$ denotes the corresponding next state and reward. The encoder maps each tuple to a Gaussian distribution over the latent task representation z, parameterized by a mean $\mu_z(x, y)$ and a covariance matrix $\Sigma_z(x, y)$. This formulation captures task uncertainty and enables context-aware sampling of task representations.

Independence Constraint. To enforce causal disentanglement in the latent task representation, we introduce an *independence constraint* based on the Hilbert–Schmidt Independence Criterion (HSIC) (Gretton et al. 2005), which explicitly reduces statistical dependence among latent dimensions. For each transition tuple (x_k, y_k) sampled from the dataset D, the context encoder $q_{\phi_z}(z \mid x_k, y_k)$ produces a latent sample z_k. Given a batch of such samples $\{z_k\}_{k=1}^{n}$, the empirical HSIC between dimensions (z^i, z^j) can be ef-

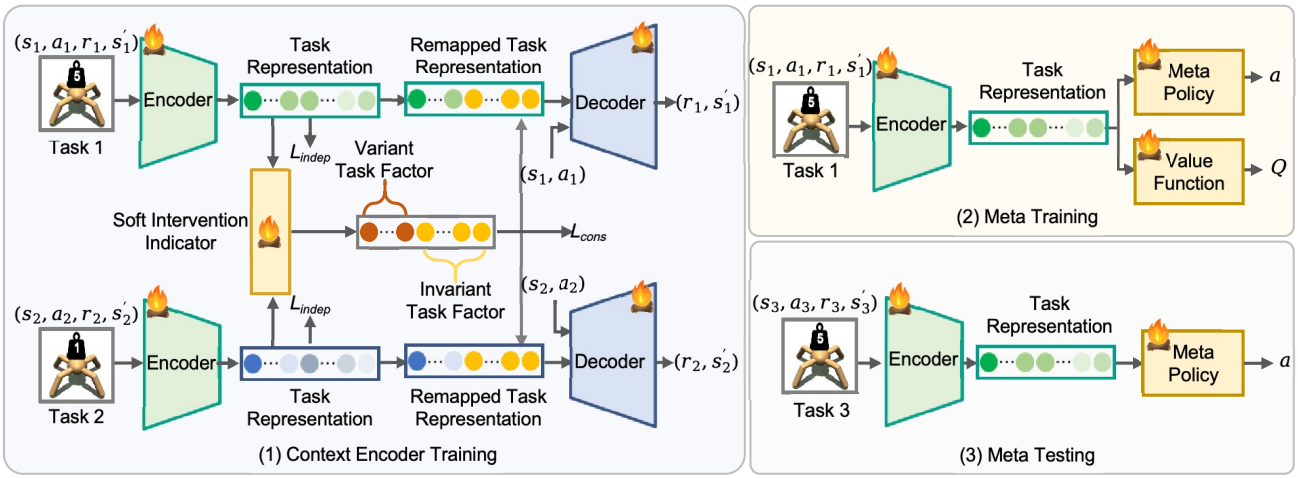

Figure 4: The CATAL framework operates in three sequential stages:(1) Context Encoder Training. The context encoder embeds paired transitions into compact task representations, regularized by an independence constraint to separate causal factors. An intervention indicator estimates the likelihood that each dimension corresponds to an intervened factor, followed by a consistency constraint to stabilize predictions. The decoder then reconstructs transition–reward pairs from state–action pairs and the projected representations.(2) Meta-Training. The learned task representations condition the meta-policy, which is optimized offline for cross-task generalization. (3) Meta-Testing. The trained meta-policy is evaluated on unseen tasks to assess generalization performance.

ficiently computed as:

$$\mathrm{HSIC}(z^i, z^j) = \frac{1}{(n-1)^2}\mathrm{tr}(KHLH), \tag{2}$$

where K and L are Gram matrices for z^i and z^j, respectively, and $H = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the centering matrix. The independence loss is defined as the expected HSIC over all dimension pairs:

$$\mathcal{L}_{\text{indep}} = \mathbb{E}\left[\frac{1}{d(d-1)}\sum_{i\neq j}\mathrm{HSIC}(z^i, z^j)\right]. \tag{3}$$

Minimizing $\mathcal{L}_{\text{indep}}$ explicitly suppresses inter-dimensional dependencies, thereby promoting independent and causally separated latent factors.

Consistency Constraint. To align each latent dimension with its corresponding causal factor, we introduce a *consistency constraint*: if a causal factor remains unchanged across tasks, the associated latent dimension should remain invariant. For two transition tuples (x_1, y_1) and (x_2, y_2) sampled from the dataset D, the context encoder $q_{\phi_z}(z \mid x, y)$ produces latent representations $z_1 \sim q_{\phi_z}(z \mid x_1, y_1)$ and $z_2 \sim q_{\phi_z}(z \mid x_2, y_2)$. We employ a binary intervention indicator $\mathcal{I}(x_1, x_2) \in \{0,1\}^d$, where $\mathcal{I}^i = 0$ indicates that the i-th causal factor remains unchanged, and $\mathcal{I}^i = 1$ indicates that it has changed. The consistency loss is defined as:

$$\mathcal{L}_{\text{cons}} = \mathbb{E}\left[\frac{1}{d}\sum_{i=1}^{d}\left\|(1-\mathcal{I})\cdot(z_1 - z_2)\right\|_2^2\right]. \tag{4}$$

Minimizing $\mathcal{L}_{\text{cons}}$ penalizes variability in latent dimensions corresponding to unchanged causal factors, thereby enforcing that each latent dimension consistently encodes a specific causal factor.

Soft Intervention Indicator. In realistic offline settings, explicit intervention labels $\mathcal{I}^i \in \{0,1\}$ are unavailable. To address this limitation, we introduce a heuristically driven *soft intervention indicator*. The key intuition is that dimension-wise variations in the latent representation provide reliable signals of whether a causal factor has been intervened upon. Specifically, we obtain latent means μ_z^i from the context encoder and compute the dimension-wise difference between two tasks as:

$$\delta_i = \mu_z^i(x_1, y_1) - \mu_z^i(x_2, y_2). \tag{5}$$

We then construct a parameterized quadratic mapping:

$$\hat{\mathcal{I}}^i(x_1, y_1, x_2, y_2) = \sigma\left(\frac{1}{C}\left(a + b\,|\delta_i| + c\,|\delta_i|^2\right)\right), \tag{6}$$

where learnable parameters a, b, c control the sensitivity to latent differences, C is a normalization constant for training stability, and $\sigma(\cdot)$ is the sigmoid function. The soft intervention indicator is incorporated into the consistency constraint:

$$\mathcal{L}_{\text{cons}} = \mathbb{E}\Big[\frac{1}{d}\sum_{i=1}^{d}(1-\hat{\mathcal{I}}^i)\cdot\|z_1^i - z_2^i\|_2^2\Big], \tag{7}$$

The soft gating weight $\hat{\mathcal{I}}^i$ enables adaptive adjustment of the constraint strength: when $\hat{\mathcal{I}}^i \to 1$ (large latent differences), the model infers a potential intervention on the i-th factor and down-weights the consistency constraint; when $\hat{\mathcal{I}}^i \to 0$ (small latent differences), the constraint is strengthened to maintain representation stability. This heuristic soft intervention mechanism enables effective causal disentanglement in a fully unsupervised setting, addressing the challenge of learning causally disentangled representations without intervention labels.

Reconstruction Objective with Remapping Mechanism. Although the independence and consistency constraints encourage disentanglement, they do not guarantee that each latent dimension fully captures its underlying causal factor. To ensure representation completeness, we introduce a *remapping-based reconstruction objective* that constrains reconstruction to rely primarily on task-varying latent dimensions, thereby reinforcing the alignment between latent features and causal factors. Given two latent representations z_1 and z_2, we compute their mean $\bar{z}$ and construct a remapped embedding using the soft intervention indicator $\hat{\mathcal{I}}$:

$$\tilde{z}_1 = \hat{\mathcal{I}} \odot z_1 + (1 - \hat{\mathcal{I}}) \odot \bar{z}. \tag{8}$$

The decoder $q_{\phi_y}(y_1 \mid x_1, \tilde{z}_1)$ reconstructs the next state and reward, yielding the reconstruction loss:

$$\mathcal{L}_{\text{recon}} = -\mathbb{E}\left[\log q_{\phi_y}(y_1 \mid x_1, \tilde{z}_1)\right]. \tag{9}$$

This objective encourages each latent dimension to encode the causal factor and enhances the completeness of the learned representation.

Overall Objective. We jointly train the context encoder, decoder, and soft intervention indicator under a unified objective, which aims to minimize a combination of the independence constraint $\mathcal{L}_{\text{indep}}$ (eq. 3), the consistency constraint $\mathcal{L}_{\text{cons}}$ (eq. 7), and the reconstruction loss $\mathcal{L}_{\text{recon}}$ (eq. 9). Formally, the joint loss is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{indep}}\mathcal{L}_{\text{indep}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} + \lambda_{\text{recon}}\mathcal{L}_{\text{recon}}, \tag{10}$$

where $\lambda_{\{\cdot\}}$ are non-negative hyperparameters that control the relative importance of each component.

Theoretical Analysis and Practical Improvements

To theoretically analyze the feasibility of CATAL in learning causally disentangled task representations, we begin by formalizing the causal generative model of paired transition-reward tuples.

Definition 2 (Causal Generative Model). Given state–action tuples $x \sim p(x)$ and $\tilde{x} \sim p(\tilde{x})$, the causal generative model of the transition–reward tuples y and $\tilde{y}$ is defined by the following generative process:

$$\begin{aligned} &x \sim p(x),\ \theta \sim p(\theta),\ y = g(x, \theta), \\ &\tilde{x} \sim p(\tilde{x}),\ I \sim p(I),\ \tilde{\theta} \sim p(\tilde{\theta} \mid \theta, I),\ \tilde{y} = g(\tilde{x}, \tilde{\theta}). \end{aligned} \tag{11}$$

where $p(\theta) = \prod_{i=1}^{n} p(\theta_i)$ denotes the joint distribution over causal factors, and $I \sim p(I)$ specifies the set of intervened factors. The conditional distribution $p(\tilde{\theta} \mid \theta, I)$ defines the remapping mechanism:

$$\tilde{\theta}_i = \begin{cases} \tilde{\theta}_i \sim p(\tilde{\theta}_i), & i \in I, \\ \theta_i, & i \notin I, \end{cases} \tag{12}$$

which selectively updates causal factors by replacing intervened components while keeping non-intervened components unchanged. The generative function g maps a state-action pair and its associated causal factors to the corresponding transition-reward outputs.

We next establish, under mild assumptions, that CATAL learns causally disentangled task representations, as stated in the theorem below.

Theorem 1 (Causal Disentanglement of Task Representations). Consider the causal generative process in Eq. (11). Assume that: (1) for the intervention set $I \subset [n]$, its complement $S = [n] \setminus I$ satisfies $p(S \cap S' = \{i\}) > 0$ for all $i \in [n]$ with $I, I' \sim p(I)$; (2) the generative function $g : \mathcal{X} \times \Theta \to \mathcal{Y}$ is continuously invertible in θ and g^{-1} is continuously differentiable; (3) the observational distribution is consistent with the generative factorization $p(x, y, \tilde{x}, \tilde{y}, I) = p(y \mid x, \theta)\, p(\tilde{\theta} \mid \theta, I)\, p(\tilde{y} \mid \tilde{x}, \tilde{\theta})\, p(x)\, p(\tilde{x})\, p(I)\, p(\theta)$. Then the aggregated posterior $q(z) = \iiint q(z \mid x, y)\, p(y \mid x, \theta)\, p(x)\, p(\theta)\, \mathrm{d}\theta\, \mathrm{d}x\, \mathrm{d}y$, where $q(z \mid x, y) \propto q(y \mid x, z)\, p(z)$, is *causally disentangled*, meaning that it corresponds to a dimension-wise reparameterization of the ground-truth prior $p(\theta)$, up to a permutation of indices.

Discussions The above theoretical analysis reveals that CATAL is unable to achieve alignment between latent dimensions and individual causal factors in two particular cases. First, if a causal factor i never appears in the intervention set I, its value remains constant across tasks, making identification impossible. Second, if a causal factor i is always included in I, its value changes in every task, preventing distinction from other varying factors. These limitations reveal a key challenge for CATAL: limited diversity in offline datasets inherently constrains the disentanglement and identifiability of causal factors.

Practical Improvements. As previously discussed, the limited diversity in offline datasets can result in the unidentifiability of certain causal factors. To address this issue, we decompose the task representation z into three distinct components: z_s, z_l and z_g, each processed independently to enhance the causal disentanglement of task representations. The first component z_s encodes shared causal factors consistent across tasks, with a per-dimension consistency constraint and remapping mechanism to enhance stability and reduce noise. The second z_l captures locally varying factors and adopts the consistency constraint and remapping machianism based on soft intervention indicator without modification. The third z_g represents globally varying factors and remains unconstrained, preserving its original posterior distribution. This structured decomposition reduces interference among causal factors and strengthens the disentanglement of task representations.

Offline Meta-Policy Optimization

To optimize policies under causally disentangled task representations, we propose a meta-policy learning method based on *representation filtering*, which removes shared causal components from task embeddings to enhance task-specific expressiveness. To mitigate overestimation in offline reinforcement learning, we adopt the BRAC framework by constraining the learned policy toward the behavior policy via a KL regularization. Under an actor–critic architecture, the

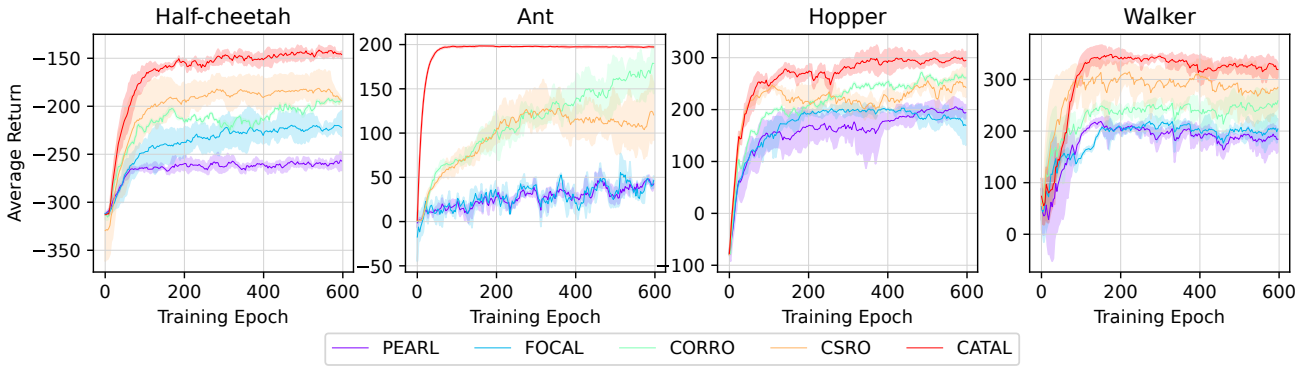

Figure 5: Comparison of average returns between CATAL and baseline methods under the recombination generalization setting.

Benchmark	Half-Cheetah		Ant		Hopper		Walker	
	ID	OOD	ID	OOD	ID	OOD	ID	OOD
Offline Pearl	$-261.0_{\pm 10.4}$	$-289.4_{\pm 3.2}$	$50.7_{\pm 10.9}$	$31.3_{\pm 20.5}$	$196.5_{\pm 26.4}$	$146.6_{\pm 18.2}$	$185.3_{\pm 13.2}$	$132.43_{\pm 14.0}$
FOCAL	$-226.1_{\pm 15.3}$	$-240.1_{\pm 13.4}$	$69.3_{\pm 20.7}$	$51.4_{\pm 32.4}$	$164.0_{\pm 17.2}$	$107.1_{\pm 19.4}$	$225.8_{\pm 27.7}$	$178.4_{\pm 21.2}$
CORRO	$-196.8_{\pm 1.2}$	$-257.8_{\pm 6.2}$	$177.1_{\pm 26.4}$	$136.4_{\pm 20.9}$	$261.8_{\pm 24.5}$	$212.8_{\pm 29.4}$	$270.1_{\pm 17.1}$	$214.4_{\pm 21.4}$
CSRO	$-197.8_{\pm 2.7}$	$-211.5_{\pm 7.4}$	$106.4_{\pm 29.5}$	$67.3_{\pm 16.4}$	$232.6_{\pm 19.6}$	$188.4_{\pm 14.7}$	$297.8_{\pm 32.2}$	$230.3_{\pm 21.3}$
CATAL	$\mathbf{-150.2}_{\pm 13.4}$	$\mathbf{-184.2}_{\pm 6.4}$	$\mathbf{197.0}_{\pm 4.8}$	$\mathbf{164.7}_{\pm 5.4}$	$\mathbf{289.9}_{\pm 22.7}$	$\mathbf{243.6}_{\pm 29.3}$	$\mathbf{318.3}_{\pm 12.6}$	$\mathbf{256.3}_{\pm 22.9}$

Table 1: Comparison of average returns between CATAL and baseline methods under the in-distribution (ID) and out-of-distribution (OOD) generalization settings.

optimization objectives are:

$$
\begin{aligned}
L_{\text{critic}} = \mathbb{E}_{\mathcal{D}}\Big[\big(Q_\omega(s,a,z_l,z_g) - r \\
- \gamma Q_\omega^{\text{target}}(s',a',z_l,z_g)\big)^2\Big], \\
L_{\text{actor}} = -\mathbb{E}_{\mathcal{D}}\Big[\ Q_\omega(s,a',z_l,z_g) \\
- \alpha\, D_{\text{KL}}(\pi_\xi \| \pi_\beta)\Big].
\end{aligned}
\tag{13}
$$

where $\mathcal{D}$ denotes the replay buffer, $a' \sim \pi_\xi$, and α controls the KL regularization toward π_β.

Experiments

In the experiments, we aim to evaluate: (1) the generalization ability of CATAL across diverse tasks involving variations in multiple task factors; (2) the quality of causally disentanglement in the task representations learned by CATAL; and (3) the contribution of causally disentanglement through ablation studies.

Experimental Settings

Benchmarks. Evaluations are conducted on multi-task MuJoCo benchmarks for offline meta-RL(Todorov, Erez, and Tassa 2012), with the following details: (1) Half-Cheetah, where the agent is required to achieve a specified target velocity; (2) Ant, which involves controlling the agent to move in various target directions; (3) Hopper, where the objective is to maximize the agent's movement speed; and (4) Walker, which focuses on controlling a bipedal agent to move forward as quickly as possible. Unlike prior work, we evaluate generalization across tasks with multiple varying causal factors by modifying environmental parameters (mass, damping, friction, inertia).

Offline Dataset Collections. For each benchmark, we sample 20 training and 20 testing tasks from the task distribution. For each task, an offline dataset is generated by training an SAC (Haarnoja et al. 2018) agent and randomly retaining 200 trajectories from its interactions.

Baseline Methods. We compare CATAL against: (1) **Offline PEARL**, which jointly optimizes the context encoder and Q-function; (2) **FOCAL** (Li, Yang, and Luo 2020), which uses metric distance learning to infer task representations; (3) **CORRO** (Yuan and Lu 2022), which employs InfoNCE (Oord, Li, and Vinyals 2018) for task representation learning; and (4) **CSRO** (Gao et al. 2024), which optimizes the context encoder via mutual information. All methods adopt the same offline RL algorithm BRAC (Wu, Tucker, and Nachum 2019) to train meta-policy, for fair comparison.

Experimental Results

Generalization Ability. To comprehensively evaluate the generalization capability of CATAL, we compare it with baseline methods under three evaluation settings, where "$*$" denotes a value expressed as a multiple of the default parameter setting: (1) **Recombination Generalization**, where unseen tasks are formed by recombining previously ob-

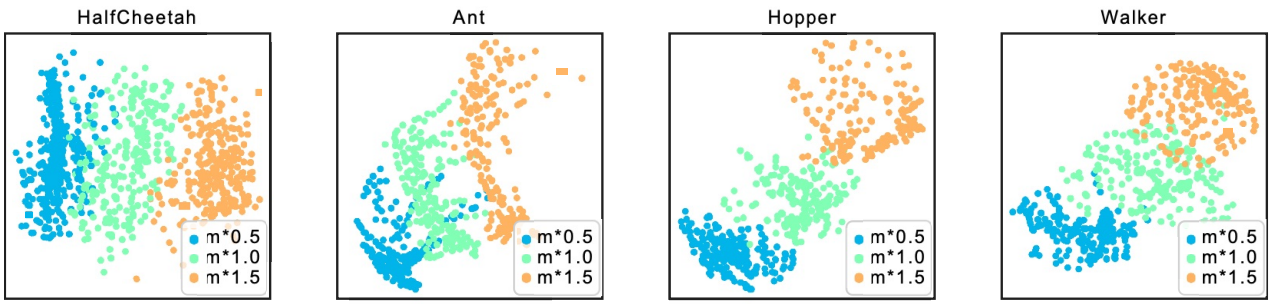

Figure 6: Visualisation of task representations via t-SNE.

Benchmark	Half-Cheetah		Ant		Hopper		Walker	
	ID	OOD	ID	OOD	ID	OOD	ID	OOD
CATAL-ICR	$-189.5_{\pm 8.3}$	$-208.7_{\pm 16.4}$	$167.4_{\pm 14.4}$	$123.3_{\pm 15.4}$	$243.6_{\pm 12.3}$	$161.2_{\pm 13.8}$	$214.8_{\pm 22.3}$	$150.3_{\pm 24.3}$
CATAL-C	$-201.3_{\pm 12.4}$	$-213.5_{\pm 15.3}$	$173.7_{\pm 21.5}$	$124.5_{\pm 18.4}$	$252.5_{\pm 18.2}$	$146.1_{\pm 21.7}$	$231.4_{\pm 24.7}$	$178.4_{\pm 19.2}$
CATAL-I	$-157.2_{\pm 14.7}$	$-251.5_{\pm 30.4}$	$174.2_{\pm 30.5}$	$159.3_{\pm 14.5}$	$196.6_{\pm 48.4}$	$159.9_{\pm 31.6}$	$242.8_{\pm 22.3}$	$164.2_{\pm 19.2}$
CATAL-I	$-171.4_{\pm 16.2}$	$-238.9_{\pm 27.8}$	$186.1_{\pm 32.7}$	$147.8_{\pm 13.4}$	$182.3_{\pm 45.6}$	$171.6_{\pm 28.3}$	$228.7_{\pm 19.8}$	$176.4_{\pm 21.1}$
CATAL	$\mathbf{-150.2}_{\pm 13.4}$	$\mathbf{-184.2}_{\pm 6.4}$	$\mathbf{197.0}_{\pm 4.8}$	$\mathbf{164.7}_{\pm 5.4}$	$\mathbf{289.9}_{\pm 22.7}$	$\mathbf{243.6}_{\pm 29.3}$	$\mathbf{318.3}_{\pm 12.6}$	$\mathbf{256.3}_{\pm 22.9}$

Table 2: Ablation study analyzing the effects of the independence constraint, consistency constraint, and remapping mechanism on the performance of CATAL under both ID and OOD evaluation settings.

served values of task factors (e.g., training on (mass $=$ $m * 1.0$, friction $= f * 0.3$) and (mass $= m * 1.2$, friction $=$ $f * 0.5$), testing on (mass $= m * 1.0$, friction $= f * 0.5$)); (2) **In-distribution Generalization**, where test tasks contain unseen values drawn from the same range as training (e.g., train on mass $\in [m * 1.0, m * 1.2]$, test on mass $= m * 1.15$); and (3) **Out-of-distribution Generalization**, where test tasks contain values outside the training range (e.g., train on mass $\in [m * 1.0, m * 1.2]$, test on mass $= m * 1.3$). Similar settings are applied to other task factors such as friction. The experimental results under these three settings are illustrated in Fig. 5 for Recombination Generalization, and in Table 1 for both In-distribution and Out-of-distribution Generalization. Across all settings, CATAL consistently outperforms the baselines, demonstrating strong and versatile generalization capability.

Disentanglement Quality. We assess the disentanglement quality of CATAL by visualizing task representations under controlled variations of the mass factor ($0.5\times$, $1.0\times$, and $1.5\times$) across four benchmark environments. As shown in Fig. 6, CATAL yields clearly separated representation clusters for different mass settings, with embedding magnitudes varying smoothly and proportionally to the underlying causal changes. This structured behavior suggests that the learned representations effectively isolate the mass factor while remaining largely invariant to other irrelevant dynamics. Moreover, similar and interpretable patterns are consistently observed under additional causal factors, such as friction and velocity, further demonstrating robust and systematic disentanglement across diverse causal dimensions.

Ablation Study. To evaluate the necessity of the independence constraint, consistency constraint, and remapping mechanism in CATAL, we conduct the following ablation studies: (1) **CATAL-ICR**: removes all three components—the independence constraint, consistency constraint, and remapping mechanism—retaining only the base framework for task representation inference; (2) **CATAL-I**: removes the independence constraint while keeping the consistency constraint and remapping mechanism; (3) **CATAL-C**: removes the consistency constraint while preserving the independence constraint and remapping mechanism; (4) **CATAL-R**: removes only the remapping mechanism while retaining both constraints. As shown in Table 2, the full CATAL model consistently outperforms all ablated variants across all evaluation metrics, demonstrating the essential roles of the structural constraints, intervention prediction module, and remapping mechanism within the overall framework.

Conclusions

In this work, we introduced CATAL, a novel framework that leverages causally disentangled task representations to improve the generalization capability of context-based OMRL methods. Evaluations across recombination, in-distribution, and out-of-distribution generalization settings demonstrate that CATAL is highly effective in addressing tasks involving multiple varying task factors. Nevertheless, as the assumptions regarding the intervention set may not always hold in real-world scenarios, future work will focus on developing more flexible approaches that require less prior knowledge.

Acknowledgments

We gratefully acknowledge the support from the Distinguished Young Scholars Project of the Natural Science Foundation of Guangdong Province (No. 2025B1515020060), the Basic and Applied Basic Research Program of the Guangzhou Science and Technology Plan (No. 2025A04J7141), the National Natural Science Foundation of China (Nos. 62402252 and 62536003), the Guangdong High-Level Talent Programme (No. 2024TQ08X283), and the Pengcheng Laboratory Project (Nos. PCL2025AS14 and PCL2025A02).

References

Beck, J.; Vuorio, R.; Liu, E. Z.; Xiong, Z.; Zintgraf, L.; Finn, C.; and Whiteson, S. 2023. A survey of meta-reinforcement learning. *arXiv preprint arXiv:2301.08028*.

Cai, R.; Huang, S.; Qiao, J.; Chen, W.; Zeng, Y.; Zhang, K.; Sun, F.; Yu, Y.; and Hao, Z. 2024. Learning by Doing: An Online Causal Reinforcement Learning Framework with Causal-Aware Policy. *arXiv preprint arXiv:2402.04869*.

Dorfman, R.; Shenfeld, I.; and Tamar, A. 2021. Offline Meta Reinforcement Learning–Identifiability Challenges and Effective Data Collection Strategies. *Advances in Neural Information Processing Systems*, 34: 4607–4618.

Duan, Y.; Schulman, J.; Chen, X.; Bartlett, P. L.; Sutskever, I.; and Abbeel, P. 2016. Rl2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*.

Dunion, M.; McInroe, T.; Luck, K. S.; Hanna, J.; and Albrecht, S. 2024. Conditional mutual information for disentangled representations in reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Gao, Y.; Zhang, R.; Guo, J.; Wu, F.; Yi, Q.; Peng, S.; Lan, S.; Chen, R.; Du, Z.; Hu, X.; et al. 2024. Context shift reduction for offline meta-reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Gretton, A.; Bousquet, O.; Smola, A.; and Schölkopf, B. 2005. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*, 63–77. Springer.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, 1861–1870. PMLR.

Li, J.; Vuong, Q.; Liu, S.; Liu, M.; Ciosek, K.; Christensen, H.; and Su, H. 2020. Multi-task batch reinforcement learning with metric learning. *Advances in Neural Information Processing Systems*, 33: 6197–6210.

Li, L.; Yang, R.; and Luo, D. 2020. Focal: Efficient fully-offline meta-reinforcement learning via distance metric learning and behavior regularization. *arXiv preprint arXiv:2010.01112*.

Li, L.; Zhang, H.; Zhang, X.; Zhu, S.; Yu, Y.; Zhao, J.; and Heng, P.-A. 2024. Towards an information theoretic framework of context-based offline meta-reinforcement learning. *Advances in Neural Information Processing Systems*, 37: 75642–75667.

Mitchell, E.; Rafailov, R.; Peng, X. B.; Levine, S.; and Finn, C. 2021. Offline meta-reinforcement learning with advantage weighting. In *International Conference on Machine Learning*, 7780–7791. PMLR.

Nagabandi, A.; Clavera, I.; Liu, S.; Fearing, R. S.; Abbeel, P.; Levine, S.; and Finn, C. 2018. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. *arXiv preprint arXiv:1803.11347*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pong, V. H.; Nair, A. V.; Smith, L. M.; Huang, C.; and Levine, S. 2022. Offline meta-reinforcement learning with online self-supervision. In *International Conference on Machine Learning*, 17811–17829. PMLR.

Rakelly, K.; Zhou, A.; Finn, C.; Levine, S.; and Quillen, D. 2019. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International conference on machine learning*, 5331–5340. PMLR.

Sontakke, S. A.; Mehrjou, A.; Itti, L.; and Schölkopf, B. 2021. Causal curiosity: Rl agents discovering self-supervised experiments for causal representation learning. In *International conference on machine learning*, 9848–9858. PMLR.

Todorov, E.; Erez, T.; and Tassa, Y. 2012. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5026–5033. IEEE.

Wu, Y.; Tucker, G.; and Nachum, O. 2019. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*.

Yuan, H.; and Lu, Z. 2022. Robust task representations for offline meta-reinforcement learning via contrastive learning. In *International Conference on Machine Learning*, 25747–25759. PMLR.

Yuan, L.; Lu, X.; and Liu, Y. 2024. Learning task-relevant representations via rewards and real actions for reinforcement learning. *Knowledge-Based Systems*, 294: 111788.

Zhang, Z.; Meng, W.; Sun, H.; and Pan, G. 2025. Causal-COMRL: Context-Based Offline Meta-Reinforcement Learning with Causal Representation. *arXiv preprint arXiv:2502.00983*.

Zhao, T. Z.; Luo, J.; Sushkov, O.; Pevceviciute, R.; Heess, N.; Scholz, J.; Schaal, S.; and Levine, S. 2022. Offline meta-reinforcement learning for industrial insertion. In *2022 international conference on robotics and automation (ICRA)*, 6386–6393. IEEE.

Zhou, R.; Gao, C.-X.; Zhang, Z.; and Yu, Y. 2024. Generalizable Task Representation Learning for Offline Meta-Reinforcement Learning with Data Limitations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 17132–17140.