Efficient Offline Meta-Reinforcement Learning via Robust Task Representations and Adaptive Policy Generation

Zhengwei Li[1,2], **Zhenyang Lin**[1,2], **Yurou Chen**[2] and **Zhiyong Liu**[1,2*]
[1]School of Artificial Intelligence, University of Chinese Academy of Science (UCAS), 100049, Beijing, China
[2]National Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China
{lizhengwei2022, linzhenyang2021, chenyurou2019, zhiyong.liu}@ia.ac.cn

Abstract

Zero-shot adaptation is crucial for agents facing new tasks. Offline Meta-Reinforcement Learning (OMRL), utilizing offline multi-task datasets to train policies, offers a way to attain this ability. Although most OMRL methods construct task representations via contrastive learning and merge them with states for policy input, these methods may have inherent problems. Specifically, integrating task representations with states for policy input limits learning efficiency, due to failing to leverage the similarities among tasks. Moreover, uniformly sampling an equal number of negative samples from different tasks in contrastive learning can hinder differentiation of more similar tasks, potentially diminishing task representation robustness. In this paper, we introduce an OMRL algorithm to tackle the aforementioned issues. We design a network structure for efficient learning by leveraging task similarity. It features shared lower layers for common feature extraction with a hypernetworks-driven upper layer, customized to process features per task's attributes. Furthermore, to achieve robust task representations for generating task-specific control policies, we utilize contrastive learning and introduce a novel method to construct negative sample pairs based on task similarity. Experimental results show that our method notably boosts learning efficiency and zero-shot adaptation in new tasks, surpassing previous methods across multiple challenging domains.

1 Introduction

The ability to adapt to new tasks and excel in the first episode, without pre-collecting samples and updating its network parameters, known as zero-shot adaptation [Shinzaki *et al.*, 2021; Ball *et al.*, 2021], is crucial for general-purpose agents operating in dynamic and open-ended environments. While reinforcement learning has seen significant achievements in areas such as gaming, autonomous driving, and robotics, these algorithms are typically tailored for a particular task. When confronted with new tasks, these algorithms often struggle to adapt using prior knowledge, usually requiring a new learning process. This requirement to learn from scratch is often impractical in real-world situations.

Meta-Reinforcement Learning (meta-RL) [Duan *et al.*, 2016; Wang *et al.*, 2016; Finn *et al.*, 2017; Beck *et al.*, 2023b] offers a solution to the challenge of enabling zero-shot adaptation in agents for new tasks. It encourages agents to develop meta-knowledge for tackling unseen tasks through training across a variety of tasks characterized by different dynamic parameters or reward functions. Consequently, when confronted with a new task, the agent is capable of accomplishing zero-shot adaptation. Context-based meta-RL methods [Rakelly *et al.*, 2019a; Zintgraf *et al.*, 2020], known for avoiding gradient updates and excelling in zero-shot adaptation, are preferred. These methods first infer a task's characteristics from context information (i.e., transitions) and then condition the policy on both task characterization and environmental state to determine the appropriate action.

Although meta-RL enhances adaptability to new tasks, it requires substantial environment interactions for training, which is often impractical in real-world scenarios due to the risk of real-time interactions, as well as the expense and time of data acquisition. Offline Reinforcement Learning (Offline RL) [Lange *et al.*, 2012; Kumar *et al.*, 2020; Fujimoto and Gu, 2021] offers an effective solution by using pre-collected datasets instead of real-time environmental interaction. Consequently, Offline Meta-Reinforcement Learning (OMRL) [Li *et al.*, 2021; Gao *et al.*, 2023] combines meta-RL's zero-shot adaptation for new tasks with the benefits of offline RL's training approach that avoids direct environmental interaction.

Most OMRL methods [Li *et al.*, 2021; Yuan and Lu, 2022; Gao *et al.*, 2023] focus on deriving task representations through contrastive learning and integrating them with environmental states for policy input. Specifically, [Li *et al.*, 2021] uses a negative-power distance metric loss to constrain the task embedding space. [Yuan and Lu, 2022] develops a framework for robust task representations against training and testing behavior policy distribution mismatches. Similarly, [Gao *et al.*, 2023] employs negative-power distance metric loss and introduces a loss minimizing policy influence

*Contact Author

in context. While these methods enhance adaptation to new tasks, they face certain challenges. The approach of integrating task representations with environmental states is inefficient because it fails to fully leverage the characteristics of task similarity for learning purposes. Moreover, uniformly sampling an equal number of negative samples from different tasks can impede distinguishing between more similar tasks, potentially diminishing the robustness of task representations and leading to decreased policy performance.

In this paper, we present a novel method incorporating **R**obust task **R**epresentations and adaptive **P**olicy **G**eneration for **O**MRL (**R2PGO**), aimed at enhancing learning efficiency in multi-task datasets and facilitating superior zero-shot adaptation in new tasks. Our approach utilizes a context-based meta-RL architecture, comprising a task inference module and a conditional policy module. The key insight of our method is that the zero-shot adaptation to new tasks can be enhanced by exploiting the characteristic of meta-RL, which involves learning across a multitude of similar tasks. Firstly, to fully leverage the characteristic of task similarity to develop a conditional policy module that facilitates efficient learning. Drawing inspiration from multi-task learning [Hessel *et al.*, 2019; Zhang and Yang, 2021], we design a novel network structure merging a shared foundational layer for common feature extraction across tasks with task-specific upper layers based on hypernetworks [Ha *et al.*, 2016], tailored to process features according to each task's attributes. Secondly, as the generation of effective task-specific control policies depends on accurate task representations, to enhance the robustness of the task inference module, we leverage contrastive learning grounded in the insight that task embeddings should be distinctly separated in the embedding space. We implement the InfoNCE contrastive loss [Oord *et al.*, 2018], proficient in handling large sets of negative samples, thereby enabling more efficient representation learning. Furthermore, we introduce a novel method in contrastive learning for constructing negative sample pairs based on task similarities, directing our task representation model to adeptly differentiate between more similar tasks, thus bolstering robustness.

Experimentally, we compare our proposed R2PGO with previous OMRL methods across various environments. These include the Point-Robot-Sparse navigation task and several challenging locomotion tasks, incorporating changes in reward or dynamic functions. Experimental results demonstrate that our proposed method substantially enhances both learning efficiency in offline multi-task datasets and zero-shot adaptation capabilities in new tasks, outperforming prior methods on a range of challenging domains. In summary, our contributions are highlighted as follows:

- We propose a novel network architecture that exploits task similarities to enhance learning efficiency in multi-task datasets and generates task-specific policies for enhanced zero-shot adaptation to new tasks.
- We introduce a method for constructing negative sample pairs in contrastive learning, aimed at better differentiating similar tasks to enhance the robustness of task representations.
- We conduct experiments in high-dimensional continuous environments to validate the effectiveness of our method compared to state-of-the-art methods. The experimental results demonstrate that our method surpasses previous methods in performance.

2 Related Work

Meta Reinforcement Learning Meta-RL aims to equip agents with the capability to rapidly adapt to new tasks by utilizing knowledge acquired from multi-task datasets. Meta-RL can be primarily classified into two categories: gradient-based methods [Finn *et al.*, 2017] and context-based methods [Rakelly *et al.*, 2019a; Zintgraf *et al.*, 2020]. The central aim of gradient-based meta-RL methods is to construct a model that can swiftly adapt to new tasks through a limited number of gradient updates. Consequently, this kind of approach does not facilitate zero-shot adaptation. Context-based meta-RL methods consist of a task inference module and a conditional policy module. The purpose of task inference lies in deducing the characteristics of a task from trajectory information, whereas the function of the conditional policy module is to guide the agent in selecting suitable actions based on the environmental state and task representation provided by the preceding module. In this paper, we adopt the architecture of the context-based meta-RL.

Offline Reinforcement Learning Offline RL, as opposed to traditional RL's learning paradigm, does not require an agent's real-time interaction with the environment for new data acquisition. Offline RL [Lange *et al.*, 2012; Kumar *et al.*, 2020; Fujimoto and Gu, 2021] employs pre-collected data generated by behavioral strategies or expert policies, reducing the risks of exploratory strategies and data collection costs. A key challenge in offline RL is the distribution shift between the learned policy and the behavioral policy. [Kumar *et al.*, 2020] addresses this by conservatively estimating Q-functions, assigning lower Q values to out-of-distribution actions. [Fujimoto *et al.*, 2019; Kumar *et al.*, 2019] tackle it by minimizing the divergence between the learned and behavioral policies. Our research follows the offline RL setting, concentrating on acquiring robust task representations and efficiently learning from offline multi-task datasets.

Offline Meta Reinforcement Learning OMRL utilizes pre-collected data for training models to quickly adapt to new tasks. Existing OMRL methods [Li *et al.*, 2020; Li *et al.*, 2021; Dorfman *et al.*, 2021; Yuan and Lu, 2022; Pong *et al.*, 2022; Wang *et al.*, 2023a; Gao *et al.*, 2023] fall into two categories based on their meta-testing approach: offline testing [Li *et al.*, 2020; Li *et al.*, 2021] and online testing [Dorfman *et al.*, 2021; Pong *et al.*, 2022; Gao *et al.*, 2023]. In the offline setting, agents adapt solely using offline data, limiting practicality in new tasks. Conversely, in the online setting, agents learn an exploration policy from offline data and apply it in online testing. Both approaches use context-based meta-RL architecture. Previous research [Li *et al.*, 2020; Li *et al.*, 2021; Yuan and Lu, 2022; Gao *et al.*, 2023] has focused on refining task inference modules for better robustness of task representations. However, we argue that effective context-based meta-RL architecture requires not just a robust task inference module but also a conditional policy module

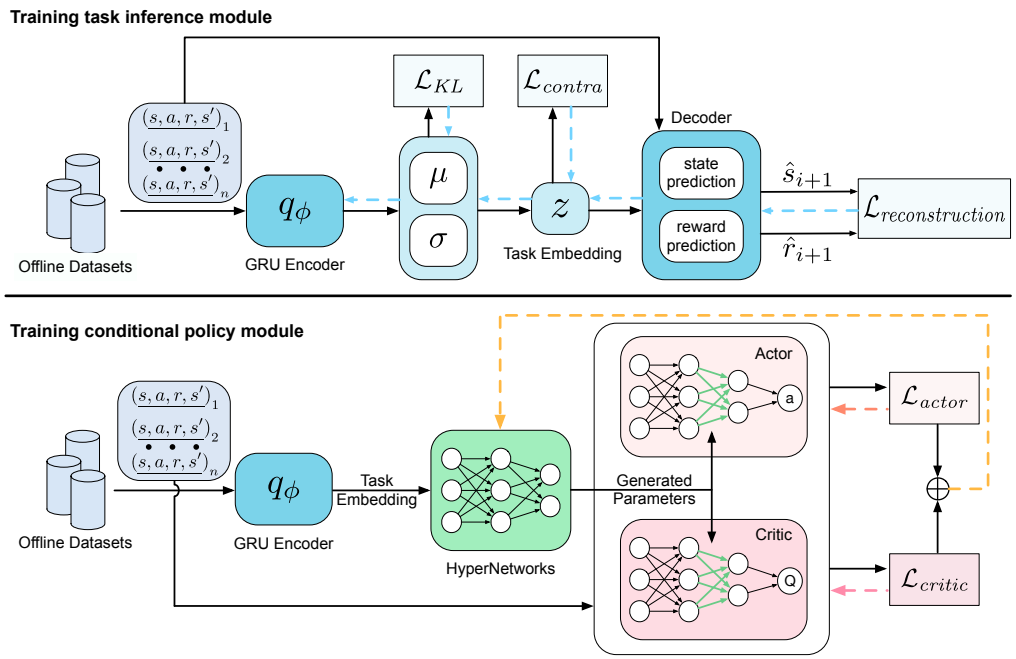

Figure 1: **Components and training procedure of R2PGO**. **Top:** During the training phase of the task inference module, the encoder extracts task embeddings from transition tuples sampled from offline datasets. Concurrently, the decoder predicts states and rewards, constituting the reconstruction loss. **Bottom:** In the training phase of the conditional policy module, the previously trained encoder module is employed to derive task embeddings from the transitions. Subsequently, hypernetworks are utilized to generate task-specific parameters in the penultimate layers of both the actor and critic networks, contingent on the input task embeddings.

that efficiently learns from large offline datasets of similar tasks. Therefore, our research aims to enhance both the task inference module's robustness and the learning efficiency of the conditional policy module.

Contrastive learning To develop robust task representations, we apply contrastive learning to improve the task inference procedure. Prior studies [Li *et al.*, 2020; Li *et al.*, 2021; Yuan and Lu, 2022; Gao *et al.*, 2023] have also used contrastive learning for this purpose. [Li *et al.*, 2021] introduced a loss function using negative-power distance metrics to constarin the task embedding space. However, its limited efficiency stems from focusing solely on binary group relationships (q_i, q_j) in loss computation, impeding learning from a wide range of samples. Similarly, [Gao *et al.*, 2023] employs negative-power distance metric loss and adds a loss to reduce policy influence within the context. Yet, the uniform sampling of equal negative samples from different tasks in these methods may weaken their ability to differentiate between more similar tasks, thereby reducing the robustness of task representation. Contrary to these approaches, our work implements an efficient contrastive loss and devises a method for constructing negative sample pairs based on task similarity, thereby enhancing the robustness of our task inference module.

Hypernetwork Hypernetworks [Ha *et al.*, 2016], networks that generate weights for other networks, have been previously utilized in meta-RL. Beck et al. [Beck *et al.*, 2023a] focused on solving the problem of initializing hypernetworks parameters, they used a recurrent network to encode the current task and generated the weights of the policy network based on the task encoding. Sarafian et al. [Sarafian *et al.*, 2021] also applied hypernetworks for policy network weight generation, but based on oracle task context. However, generating weights for entire policy networks can lead to convergence challenges and increased training complexity of hypernetworks. Beyond the scope of meta-RL, Mickael et al. [Beukman *et al.*, 2023] have implemented hypernetworks to tackle issues related to RL generalization. This approach introduces an adapter module, whose weights are generated by hypernetworks, informed by oracle task information. Nevertheless, methods reliant on the oracle task context may be impractical in unknown environments. Xu et al. [Xu *et al.*, 2023] use demonstration data to train a hypernetwork, then set the adapter module's parameters via the pretrained hypernetwork, followed by online rollouts for fine-tuning, which eases the overall network's fine-tuning process. Conversely, our approach inputs task representations encoded by a VAE into the hypernetwork, directly generating a task-specific control policy without fine-tuning. To our knowledge, prior research has largely overlooked refining the conditional policy module in context-based meta-RL architectures. In this paper, we leverage the properties of hypernetworks to enhance the learning efficiency of the conditional policy and improve adaptation performance.

3 Preliminaries

Problem Formulation In standard RL, solving a task by formulating as Markov Decision Process (MDP) [Bellman,

1966]. MDP is defined as a tuple $M = (S, A, P, \rho_0, R, \gamma)$ with state space S, action space A, transition function $P(s'|s, a)$, initial state distribution $\rho_0(s)$, reward function $R(s, a)$, and discount factor γ. The objective of RL is to maximize the expected cumulative rewards $J(\pi) = \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t R\left(s_t, a_t\right)\right]$ to get an optimal policy π.

In the context of offline meta-RL, we assume the existence of a prior distribution of tasks, denoted as $p(M)$. Each MDP is sampled from this distribution $p(M)$, representing a distinct task. These sampled MDPs share the same state space and action space but differ in their reward functions or transition functions. Consequently, the task distribution can be expressed as $p(R, P)$. Since the specific parameters of the different tasks are unknown, context-based OMRL samples a mini-batch of historical transitions from offline datasets. Subsequently, the task inference module transforms these transitions into task representations, on which both the policy and value function are conditioned. OMRL aims to maximize the expected cumulative rewards in the test tasks:

$$J(\pi_\theta) = \mathbb{E}_{M \sim p(M)}[J_M(\pi_\theta)] \tag{1}$$

Task Auto-Encoder In the task adaptation phase, an agent explores new tasks to gather context information, which aids in task inference and enables policy adjustment to maximize returns. Regarding task inference, VariBAD [Zintgraf *et al.*, 2020] adopted Bayes-Adaptive (BAMDP) [Duff, 2002] and meta-trained a Variational Auto-Encoder (VAE) [Kingma and Welling, 2013] to extract task representations from historical trajectories. The VAE consists of an encoder $q_\phi(m|\tau_{:t})$ for task representation and decoders to forecast future rewards and states, contributing to the reconstruction loss that is used during meta-training. The training objective is

$$\mathcal{L}_{\text{VAE}}(\phi, \theta) = \mathbb{E}_{p(M)} \left[\sum_{t=0}^{H^+} ELBO_t(\phi, \theta)\right] \tag{2}$$

where

$$\begin{aligned} ELBO_t = &\mathbb{E}_{p(M)} \left[\mathbb{E}_{q_\phi(m|\tau_{:t})} \left[\log p_\theta\left(\tau_{:H+} \mid m\right)\right]\right. \\ &\left. -KL\left(q_\phi\left(m \mid \tau_{:t}\right) \| q_\phi\left(m \mid \tau_{:t-1}\right)\right)\right], \end{aligned} \tag{3}$$

H^+ is the horizon in the BAMDP. The objective is to maximize evidence lower bound (ELBO), comprising a reconstruction term for the trajectory and a KL divergence relative to the previous posterior.

Similarly, we adopt this method to generate task embedding at the current time step, utilizing historical information up to this point. In contrast, we have additionally improved the robustness of the task inference module.

4 Methodology

In this section, we outline our methodology, encompassing the training of both the task inference module and the conditional policy module in an offline setting. The section begins with an overview of our proposed method, followed by an in-depth explanation of each component within the training segment.

4.1 Method Overview

Our method comprises a three-phase process: (1) Collecting offline multi-task datasets, which include trajectories for training both the task inference and conditional policy modules. (2) Training the task inference module to generate embeddings that encapsulate the attributes of each task. (3) Training the conditional policy module to execute actions based on the task embedding and environmental state. As depicted in Figure 1, our proposed methodology is outlined. During the offline data collection phase, we gathered trajectory data at various checkpoints. This data, capturing a spectrum of tasks linked to either reward or dynamic variation, was collected utilizing the Soft Actor-Critic (SAC) algorithm [Haarnoja *et al.*, 2018]. In the task inference module's training phase, our approach mirrors that in [Zintgraf *et al.*, 2020], employing a VAE to transform trajectory data into task embeddings. For online adaptation, the architecture of the encoder is based on a recurrent neural network. To enhance the robustness of the task inference module, we have integrated InfoNCE loss [Oord *et al.*, 2018] to distinguish between different task embeddings. Additionally, we have developed a novel method for constructing negative sample pairs, aimed specifically at improving the differentiation of more similar task embeddings. During the training phase of the conditional policy module, we utilize the characteristic of task similarity to augment learning efficiency. Specifically, we introduce an innovative network architecture for training the actor-critic network. This architecture merges a shared foundational layer for common feature extraction across tasks with task-specific upper layers based on hypernetworks, tailored to process features according to each task's attributes. In the testing phase, our method can achieve online adaptation. It infers the task representation based on historical trajectory up to the current moment and subsequently generates the corresponding task-specific policy.

4.2 Adaptive Policy Generation

Earlier approaches [Rakelly *et al.*, 2019b; Yuan and Lu, 2022; Gao *et al.*, 2023] implemented a conditional policy reliant on environmental states and task representations. However, this strategy does not fully exploit the essence of meta-RL, which involves learning across a multitude of similar tasks. Specifically, it fails to account for the explicit utilization of shared features among similar tasks and the advantages of leveraging task properties for task-specific control, both of which are essential for enhancing learning efficiency and adaptation performance. Concerning these potential issues, and drawing inspiration from multi-task learning [Hessel *et al.*, 2019; Zhang and Yang, 2021], we propose a novel method that explicitly utilizes shared features among similar tasks and leverages task properties to achieve task-specific control. We posit that the network's lower layers should extract common features across tasks, whereas the upper layer should specialize in particular tasks. To achieve this specialization in the upper layer, we propose the implementation of a hypernetwork designed to generate network parameters specifically tailored to task properties. Consequently, task representations can be fed into the hypernetwork, facilitating the creation of task-specific control policies.

To elucidate the operational principles of our proposed method, consider integrating a hypernetwork, denoted as $\mathcal{H}$, which derives the parameter set for the final layer of a policy network. This framework posits an environmental state represented by $s \in \mathcal{S}$ and a task-specific attribute denoted by $z \in \mathcal{Z}$. The architecture under consideration is a three-layer policy network designed to generate actions in response to given states. The operational dynamics of the network can be formalized by the recurrence relation: $x_{i+1} = F_{\theta_i}(x_i)$, where $x_1 = s$ represents the initial state, and $x_4 = a$ denotes the resulting action. Here, the action a is derived using parameters θ_3, where θ_3 is obtained through the relation $\theta_3 = \mathcal{H}(z)$. Thus, the transformations within the network can be explicitly described by $x_2 = F_{\theta_1}(s)$, $x_3 = F_{\theta_2}(x_2)$, and $a = F_{\theta_3}(x_3) = F_{\mathcal{H}(z)}(x_3)$. This network framework is designed to facilitate the implementation of specialized policies that are adaptive to varying task properties.

In conclusion, considering the intrinsic characteristics of meta-RL across similar tasks and the benefit of leveraging task properties to achieve task-specific control, we have effectively created a conditional policy module that enhances learning efficiency and adaptation performance in new tasks. Furthermore, our network architecture is sufficiently versatile to be adaptable for use in an online meta-RL setting.

4.3 Robust Contrastive Task Representation Learning

As the generation of effective task-specific control policies depends on accurate task representations, similar to methods described in [Li *et al.*, 2021; Wang *et al.*, 2023b; Gao *et al.*, 2023], we use contrastive learning to strength the robustness and accuracy of task representations. To improve the efficiency of training data utilization in contrastive learning, we have adopted the more efficient InfoNCE loss [Oord *et al.*, 2018]. This method involves considering an anchor sample x, a positive sample x^+, and multiple negative samples $\{x_i^-\}_{i=1}^{N-1}$.

We designate task embedding z_t at a certain time step as the anchor point x and select the task embedding from the same task at a different time step as the positive sample x^+. Task embeddings from distinct tasks serve as the negative samples $\{x_i^-\}_{i=1}^{N-1}$. However, the question arises: how many negative samples should be sampled from different tasks? Prior studies have implemented these loss functions by uniformly sampling an equal number of negative samples from different tasks. Nevertheless, this method could lead to models that struggle to distinguish between the embeddings of more similar tasks.

Ideally, the model should be compelled to concentrate on tasks with high similarity, which are more challenging to distinguish. To attain this goal, we propose that varying numbers of negative samples should be selected based on the task's similarity. To determine weights for the quantity of negative samples to be sampled from different tasks, we employ the Gaussian kernel function. By taking into account known properties of different tasks, such as the speed value in the Half-Cheetah-Vel task, we are then able to calculate weights for the sampling of negative sample amounts from these tasks. Yet, it should be noted that in cases of tasks with very high similarity, full separation may not be imperative.

$$b_{ij} = e^{(-\alpha * \|\mathbf{p}_i - \mathbf{p}_j\|^2)} \tag{4}$$

where b_{ij} represents the weight of task j relative to task i, $\mathbf{p}_i$ and $\mathbf{p}_j$ denote the attribute vectors of tasks i and j respectively. A larger value of α assign higher weights to tasks with greater similarities.

The final contrastive learning objective is mathematically formulated as:

$$\mathcal{L}_{\text{contra}} = -\frac{1}{N_{\text{tasks}}} \sum_{i=0}^{N_{\text{tasks}}} \log \frac{\exp(x_i \cdot x_i^+ / \tau)}{\sum_{j=0, j\neq i}^{N_{\text{tasks}}} \sum_{k=0}^{N_{ij}^-} \exp(x_i \cdot x_{ijk}^- / \tau)} \tag{5}$$

Here N_{ij}^- is defined as $b_{ij} * N_{\text{total}}^-$, where N_{total}^- represents the total number of negative samples corresponding to a task, N_{tasks} indicates the number of sampled tasks, and x_{ijk}^- denotes the k-th negative sample in task j corresponding to the sample from task i. Consequently, the final objective is obtained as $\mathcal{L}(\phi, \theta) = \mathcal{L}_{\text{VAE}} + \mathcal{L}_{\text{contra}}$.

5 Experiments

We propose a method for generating adaptive policies based on hypernetworks, coupled with robust task representations that leverages contrastive learning. Through our experiments, we aim to answer the following questions: (1) Is our methodology capable of efficiently transitioning to online testing of new tasks following offline training, particularly concerning learning efficiency and adaptation performance? (2) Does our approach demonstrate zero-shot adaptation to new tasks? (3) Does our innovative combination of an efficient conditional policy module with a hypernetwork-based architecture and an improved task inference module based on contrastive learning enhance the performance in offline meta-RL? To address the first and second questions, we conduct comparative evaluations of our method against others in various environments. For the third question, ablation studies are conducted to elucidate the impact of our proposed improvements. Additionally, we present the learned task representations as a qualitative assessment of our proposed method.

5.1 Experimental Settings

Environments. We evaluate our proposed method on various benchmarks frequently utilized for assessing meta-RL performance: (1) **Point-Robot-Sparse** is a 2D navigation task in continuous space, where the point robot must reach a goal position located on a semi-circle. (2) **Half-Cheetah-Vel** involves controlling the velocity of a half-cheetah robot to match a target speed in continuous space. (3) **Walker-Rand-Params** requires the agent to move forward as quickly as possible. (4) **Hopper-Rand-Params**, similar to Walker-Rand-Params, focuses on controlling the agent to attain the highest possible velocity. The variations in these tasks are characterized as follows: for (1), the goal positions differ for each task; (2) vary in target velocity; (3) and (4) feature dynamic functions that alter based on environmental conditions,

indicating robots with varying physical parameters such as mass, inertia, friction, and stamping.

Offline Data Collection. In each environment, multiple training tasks are sampled from the task distribution. The Soft Actor-Critic (SAC) algorithm [Haarnoja *et al.*, 2018] is then employed on each task to train an agent for task completion. To create diverse datasets across various environments, we roll out policies at different timesteps to collect trajectory data.

Baselines. To demonstrate the effectiveness of our method, we compare it with the following methods: (1) **BoReL** [Dorfman *et al.*, 2021] is an offline version of VariBAD [Zintgraf *et al.*, 2020]. To ensure a fair comparison, we utilize a variant of BOReL that does not incorporate oracle reward functions, as delineated in the original paper [Dorfman *et al.*, 2021]. (2) **FOCAL** [Li *et al.*, 2021] employs metric learning to train the context encoder, thereby attracting representations of identical tasks and repelling those of different tasks. (3) **CORRO** [Yuan and Lu, 2022] utilizes either CVAE [Sohn *et al.*, 2015] or noise addition to generate negative samples, applying the InfoNCE loss [Oord *et al.*, 2018] for training the context encoder. (4) **CSRO** [Gao *et al.*, 2023] innovates with a max-min mutual information representation learning mechanism and develops a non-prior context collection strategy to address the context shift problem in OMRL.

To guarantee a fair comparison, each method is implemented using identical offline datasets. To better study the issue of acquiring task representations and learning efficiently from offline multi-task datasets, the use of offline RL methods is intentionally excluded in all approaches.

5.2 Comparison Unseen Tasks Adaptation Performance

To evaluate the performance of our method, we compare R2PGO with other methods in four environments. In the online adaptation phase, each method needs to explore new tasks and collect context information to update understanding of these new tasks. We plot the mean and standard deviation curves of returns across five random seeds in Figure 2. The performance is measured by the average return over all the test tasks.

Figure 2 demonstrates that R2PGO outperforms all baselines across the four environments, particularly in Point-Robot-Sparse, Walker-Rand-Params, and Hopper-Rand-Params. The experimental results substantiate the effectiveness of our proposed method, achieving optimal learning efficiency and adaptation performance.

5.3 Comparison Zero-shot Adaptation Results

To evaluate our method's zero-shot adaptation capabilities, we compared it with other methods in various environments during the initial episodes. It is important to note that both our method and BOReL [Dorfman *et al.*, 2021] focus on the agent's performance in the first two episodes, hence we lack data on the trained policy's performance in subsequent episodes.

Table 1 illustrates that R2PGO can adapt to new tasks and attain respectable performance within the first episode, surpassing other baselines across all environments. Although

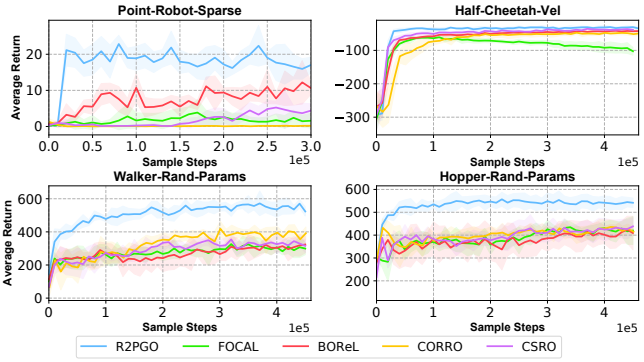

Figure 2: The average return for unseen tasks in the second rollout during the online test phase, compared with baselines in four environments, is shown. These include Point-Robot-Sparse and Half-Cheetah-Vel, with variable reward functions, and Walker-Rand-Params and Hopper-Rand-Params, featuring changing dynamic functions.

BOReL operates under the same setting as our method and achieves relatively good performance, it falls short of attaining the best performance in the first episode. Other methods, such as FOCAL, CORRO, and CSRO, do not reach satisfactory performance in the initial episode and necessitate significantly more environmental interactions in each new task to achieve such performance.

5.4 Ablation

To validate our proposed method, we conducted ablation experiments. The efficient adaptive policy generation, rooted in our novel architecture using hypernetworks, and the robust task representation learning via contrastive learning, are crucial elements of our method. We compare the methods with those without adaptive policy generation and robust contrastive task representation components, to show the effect of each component. We conducted experiments on four environments, and the results are shown in Figure 3. In the four environments, the absence of both components was found to result in the lowest average return. However, incorporating RCTR led to an enhanced average return. Notably, adding APG significantly improved both the average return and learning efficiency in most environments. This enhancement is attributed to our innovative network structure, which adeptly mines and utilizes shared features across tasks and achieves task-specific control, thereby augmenting learning efficiency and adaptability to new tasks. Specifically, in the Point-Robot-Sparse environment, our robust contrastive task representation method demonstrates exceptional efficacy in identifying new tasks. This capability is especially valuable in sparse reward environments, where discerning the properties of new tasks poses significant challenges.

5.5 Visualization

To demonstrate the quality of the learned task representations, t-SNE [Van der Maaten and Hinton, 2008] is employed to map the task embedding vectors into 2D space, enabling the visualization of these task representations. For each testing task, 100 transitions from the meta-testing phase are sampled

Environment	Algorithms	Episode 1	Episode 2	Episode 3	Episode 4	Episode 5
Half-Cheetah-Vel	R2PGO (Ours)	-32.3±1.5	**-28.9**±1.2	-	-	-
	BOReL	-40.8±2.1	-41.7±3.4	-	-	-
	FOCAL	-185.6±19.3	-80.5±5.7	-87.5±5.0	-91.9±4.9	-94.9±4.8
	CORRO	-333.6±48.4	-48.0±6.4	-55.2±17.1	-71.8±20.5	-62.8±14.9
	CSRO	-193.0±24.0	-38.9±1.6	-37.9±1.8	-37.7±2.0	-37.5±1.9
Walker-Rand-Params	R2PGO (Ours)	520.9±29.0	**545.3**±23.5	-	-	-
	BOReL	283.1±55.3	321.7±49.7	-	-	-
	FOCAL	270.4±9.8	318.6±29.9	316.7±24.0	347.1±28.3	334.6±49.2
	CORRO	190.7±13.9	412.5±41.6	391.9±45.2	409.2±51.7	405.2±34.1
	CSRO	270.2±43.1	317.2±26.5	317.5±18.8	317.3±34.4	328.3±23.5
Hopper-Rand-Params	R2PGO (Ours)	**539.7**±22.0	539.1±26.8	-	-	-
	BOReL	377.5±56.2	404.3±65.0	-	-	-
	FOCAL	371.9±32.6	421.7±48.0	412.3±58.3	424.5±53.4	425.9±50.6
	CORRO	196.7±8.0	422.7±25.4	437.5±30.6	437.8±25.2	436.1±26.8
	CSRO	385.1±29.6	424.8±21.5	426.4±11.7	433.2±19.1	434.9±28.1

Table 1: The average test performance across three different MuJoCo environments is reported, with each environment being trained separately using five seeds per method. To demonstrate their adaptability to unseen tasks, the meta-trained policies were rolled out over five episodes.

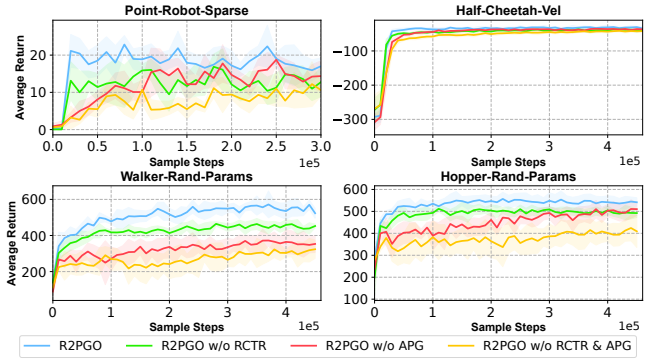

Figure 3: Ablation study conducted across four environments comparing R2PGO against methods lacking Robust Contrastive Task Representation (RCTR) and Adaptive Policy Generation (APG).

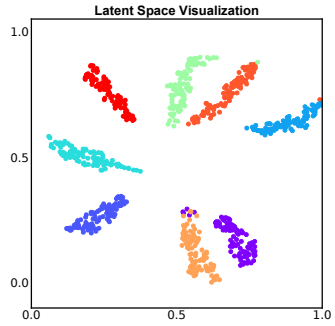

Figure 4: The t-SNE visualization of the learned task representation space in the Half-Cheetah-Vel environment is presented. Each point in the visualization represents an embedding vector extracted from transitions and is color-coded based on task property.

to visualize the task embeddings. As depicted in Figure 4, our method effectively clusters the task embeddings of the same task and distinguishes between embeddings from different tasks.

6 Conclusion and Future Work

In this study, we address the challenge of enhancing learning efficiency in multi-task datasets within an offline meta-RL setting, while also focusing on the development of robust task representations that enable the generation of effective task-specific control policies. We introduce R2PGO, a novel offline meta-RL algorithm that is rooted in a context based meta-RL framework. Specifically, we develop a unique network architecture that effectively leverages task similarity for efficient learning and generates task-specific control policies for enhanced zero-shot adaptation to new tasks. This architecture integrates a shared foundational layer that extracts common features across various tasks with specialized upper layers powered by hypernetworks, which are customized to process features according to each task's specific attributes. Additionally, to achieve robust task representations for generating effective task-specific control policies, we employ contrastive learning and introduce a novel method for constructing negative sample pairs based on task similarity. Experimental results demonstrate that our method significantly improves learning efficiency in large multi-task datasets and supports zero-shot adaptation to new tasks. Moreover, it outperforms previous methods across a range of challenging domains.

While our work has achieved significant progress, it has not addressed the issue that the training and testing distributions are identical, an assumption that likely does not hold in real-world generalization contexts. Future efforts will be dedicated to addressing the challenge of out-of-distribution generalization, thereby enhancing the adaptability of the agent in dynamic and open-ended environments.

References

[Ball *et al.*, 2021] Philip J Ball, Cong Lu, Jack Parker-Holder, and Stephen Roberts. Augmented world models facilitate zero-shot dynamics generalization from a single offline environment. In *International Conference on Machine Learning*, pages 619–629. PMLR, 2021.

[Beck *et al.*, 2023a] Jacob Beck, Matthew Thomas Jackson, Risto Vuorio, and Shimon Whiteson. Hypernetworks in meta-reinforcement learning. In *Conference on Robot Learning*, pages 1478–1487. PMLR, 2023.

[Beck *et al.*, 2023b] Jacob Beck, Risto Vuorio, Evan Zheran Liu, Zheng Xiong, Luisa Zintgraf, Chelsea Finn, and Shimon Whiteson. A survey of meta-reinforcement learning. *arXiv preprint arXiv:2301.08028*, 2023.

[Bellman, 1966] Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

[Beukman *et al.*, 2023] Michael Beukman, Devon Jarvis, Richard Klein, Steven James, and Benjamin Rosman. Dynamics generalisation in reinforcement learning via adaptive context-aware policies. *arXiv preprint arXiv:2310.16686*, 2023.

[Dorfman *et al.*, 2021] Ron Dorfman, Idan Shenfeld, and Aviv Tamar. Offline meta reinforcement learning–identifiability challenges and effective data collection strategies. *Advances in Neural Information Processing Systems*, 34:4607–4618, 2021.

[Duan *et al.*, 2016] Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. Rl ^{2}: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

[Duff, 2002] Michael O'Gordon Duff. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. University of Massachusetts Amherst, 2002.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[Fujimoto and Gu, 2021] Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021.

[Fujimoto *et al.*, 2019] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pages 2052–2062. PMLR, 2019.

[Gao *et al.*, 2023] Yunkai Gao, Rui Zhang, Jiaming Guo, Fan Wu, Qi Yi, Shaohui Peng, Siming Lan, Ruizhi Chen, Zidong Du, Xing Hu, et al. Context shift reduction for offline meta-reinforcement learning. *arXiv preprint arXiv:2311.03695*, 2023.

[Ha *et al.*, 2016] David Ha, Andrew Dai, and Quoc V. Le. Hypernetworks, 2016.

[Haarnoja *et al.*, 2018] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

[Hessel *et al.*, 2019] Matteo Hessel, Hubert Soyer, Lasse Espeholt, Wojciech Czarnecki, Simon Schmitt, and Hado Van Hasselt. Multi-task deep reinforcement learning with popart. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3796–3803, 2019.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Kumar *et al.*, 2019] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 32, 2019.

[Kumar *et al.*, 2020] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

[Lange *et al.*, 2012] Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning: State-of-the-art*, pages 45–73. Springer, 2012.

[Li *et al.*, 2020] Jiachen Li, Quan Vuong, Shuang Liu, Minghua Liu, Kamil Ciosek, Henrik Christensen, and Hao Su. Multi-task batch reinforcement learning with metric learning. *Advances in Neural Information Processing Systems*, 33:6197–6210, 2020.

[Li *et al.*, 2021] Lanqing Li, Rui Yang, and Dijun Luo. FOCAL: Efficient fully-offline meta-reinforcement learning via distance metric learning and behavior regularization. In *International Conference on Learning Representations*, 2021.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[Pong *et al.*, 2022] Vitchyr H Pong, Ashvin V Nair, Laura M Smith, Catherine Huang, and Sergey Levine. Offline meta-reinforcement learning with online self-supervision. In *International Conference on Machine Learning*, pages 17811–17829. PMLR, 2022.

[Rakelly *et al.*, 2019a] Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International conference on machine learning*, pages 5331–5340. PMLR, 2019.

[Rakelly *et al.*, 2019b] Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International conference on machine learning*, pages 5331–5340. PMLR, 2019.

[Sarafian *et al.*, 2021] Elad Sarafian, Shai Keynan, and Sarit Kraus. Recomposing the reinforcement learning building

blocks with hypernetworks. In *International Conference on Machine Learning*, pages 9301–9312. PMLR, 2021.

[Shinzaki *et al.*, 2021] Masao Shinzaki, Yusuke Koda, Koji Yamamoto, Takayuki Nishio, Masahiro Morikura, Yushi Shirato, Daisei Uchida, and Naoki Kita. Zero-shot adaptation for mmwave beam-tracking on overhead messenger wires through robust adversarial reinforcement learning. *IEEE Transactions on Cognitive Communications and Networking*, 8(1):232–245, 2021.

[Sohn *et al.*, 2015] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[Wang *et al.*, 2016] Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

[Wang *et al.*, 2023a] Jianhao Wang, Jin Zhang, Haozhe Jiang, Junyu Zhang, Liwei Wang, and Chongjie Zhang. Offline meta reinforcement learning with in-distribution online adaptation. *arXiv preprint arXiv:2305.19529*, 2023.

[Wang *et al.*, 2023b] Mingyang Wang, Zhenshan Bing, Xiangtong Yao, Shuai Wang, Huang Kai, Hang Su, Chenguang Yang, and Alois Knoll. Meta-reinforcement learning based on self-supervised task representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10157–10165, 2023.

[Xu *et al.*, 2023] Mengdi Xu, Yuchen Lu, Yikang Shen, Shun Zhang, Ding Zhao, and Chuang Gan. Hyper-decision transformer for efficient online policy adaptation. *arXiv preprint arXiv:2304.08487*, 2023.

[Yuan and Lu, 2022] Haoqi Yuan and Zongqing Lu. Robust task representations for offline meta-reinforcement learning via contrastive learning. In *International Conference on Machine Learning*, pages 25747–25759. PMLR, 2022.

[Zhang and Yang, 2021] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021.

[Zintgraf *et al.*, 2020] Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. In *International Conference on Learning Representation (ICLR)*, 2020.