Cross-Domain Generalization with Reverse Dynamics Models in Offline Model-Based Reinforcement Learning

Yana Stoyanova and Maryam Tavakol
Eindhoven University of Technology, Eindhoven, The Netherlands

Keywords: Offline Model-Based Reinforcement Learning, Reverse Dynamics Models, Cross-Domain Generalization, Context-Awareness.

Abstract: Recent advancements in offline reinforcement learning (RL) have enabled automation in many real-world applications, where online interactions are often infeasible or costly, especially in high-stakes problems like healthcare or robotics. However, most algorithms are developed and evaluated in the same environment, which does not reflect the ever-charging nature of our world. Hence, beyond dealing with the distributional shift between the learning policy and offline data, it is crucial to account for domain shifts. Model-based offline RL (MBORL) methods are generally preferred over model-free counterparts for their ability to generalize beyond the dataset by learning (forward) dynamics models to generate new trajectories. Nevertheless, these models tend to overgeneralize in out-of-support regions due to limited samples. In this paper, we present a safer approach to balance conservatism and generalization by learning a *reverse* dynamics model instead, that can adapt to environments with varying dynamics, known as cross-domain generalization. We introduce CARI (Context-Aware Reverse Imaginations), a novel approach that incorporates context-awareness to capture domain-specific characteristics into the reverse dynamics model, resulting in more accurate models. Experiments on four variants of Hopper and Walker2D demonstrate that CARI consistently matches or outperforms state-of-the-art MBORL techniques that utilize a reverse dynamics model for cross-domain generalization.

1 INTRODUCTION

Reinforcement learning stands at the forefront of contemporary research in the fields of artificial intelligence, captivating the attention of scientists, engineers, and practitioners alike. This paradigm represents a pivotal departure from traditional approaches, introducing a dynamic framework that enables intelligent agents to learn from interaction with their environment. Specifically, in online reinforcement learning, agents learn and adapt in real time, by continuously interacting with the environment and updating their strategies as new information becomes available (Sutton and Barto, 2018). However, despite its huge prospects to revolutionize different fields and industries, RL has remained mainly in the realm of research and experimentation. This is because in most real-world scenarios, simulation or online interaction with the environment is usually impractical, costly and/or dangerous (Prudencio et al., 2022). Therefore, offline reinforcement learning (Sutton and Barto, 2018), where the agent learns from a precollected offline dataset, is an appealing alternative especially in high-stakes domains such as healthcare (Liu et al., 2020a) or robotics (Singh et al., 2021). However, learning from such a static dataset is a very challenging task, because the agent needs to find a balance between increased generalization and avoiding unwanted behaviors outside of distribution (distributional shift) (Prudencio et al., 2022).

Usually, RL agents are broadly separated into two different categories, namely model-free reinforcement learning (MFRL) agents and model-based reinforcement learning (MBRL) agents (Sutton and Barto, 2018), and each category tackles the distributional shift issue in various ways. Importantly, most MFRL approaches introduce policy constrains and aggressive regularization techniques, therewith limiting their task generalization. On the other hand, MBRL methods are shown to exhibit better generalization abilities over their model-free counterparts, but they usually depend on (accurate) uncertainty estimation. A way to refrain from such unwanted behaviors is to learn a reverse dynamics model for generating reverse imaginations, which has been shown to provide informed data augmentation and enable

Stoyanova, Y. and Tavakol, M.
Cross-Domain Generalization with Reverse Dynamics Models in Offline Model-Based Reinforcement Learning.
DOI: 10.5220/0013097200003890
In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 2*, pages 57-68
ISBN: 978-989-758-737-5; ISSN: 2184-433X

conservative generalization beyond the offline dataset (Wang et al., 2021; Lyu et al., 2022). Nonetheless, in the offline MBRL context this line of work is very recent and more research is needed to explore its full potential.

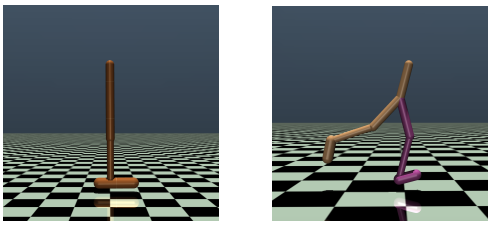
Having the ability to generalize to new scenarios is one of the most important requirements for the safe deployment of RL agents, particularly in high-stakes domains. However, in the majority of research works, agents are trained and tested on offline datasets that focus on singleton environments, where all trajectories are from the same environment with the same dynamics (Mediratta et al., 2023). Thus, the generalization capabilities of offline RL agents to new environments (with different initial states, transition functions, or reward functions), also referred to as cross-domain generalization, remain underexplored. In fact, this makes RL agents practically ill-equipped and unprepared for the ever-changing world.

So far offline MBRL approaches that incorporate learning a reverse dynamics model have not yet been studied in the context of cross-domain generalization, despite emerging as a promising direction. Motivated by this lack of research and driven by the urge to contribute making RL agents reliable for real-life deployment, the goal of this work is to bridge the gap between cross-domain generalization and MBORL approaches that learn reverse dynamics models by proposing a novel framework. Therefore, the contributions of this work are two-fold:

- Performance comparison between existing offline MBRL approaches that learn reverse dynamics models, namely Reverse Offline Model-based Imagination (ROMI), Confidence-Aware Bidirectional offline model-based Imagination (CABI) and Backwards Model-based Imitation Learning (BMIL), with respect to cross-domain generalization on Hopper and Walker2D environments (Figure 1). An in-depth analysis of the findings is performed, focusing on the specific characteristics inherent to each approach that either facilitate or impede generalization.
- A novel offline MBRL framework, based on reverse dynamics models, named Context-Aware Reverse Imagination (CARI), for improved cross-domain generalization.

2 RELATED WORK

Reinforcement learning is typically conceptualized in two main settings, namely online and offline. In online RL, agents learn by directly interacting with the environment in real-time. The agent takes actions based on its current policy, observes the resulting state transitions, and receives immediate feedback in the form of rewards. These experiences are then used to update the agent's policy or value function, driving iterative learning. However, online RL poses safety concerns in certain domains, making it very restricted or even infeasible to use for some real-world applications (Dulac-Arnold et al., 2019; Levine et al., 2020).



(a) Hopper (b) Walker2D

Figure 1: Experimental environments.

In contrast, offline RL operates on fixed datasets of previously collected experiences, without the need for direct interaction with the environment during the learning process (Levine et al., 2020). These datasets are typically pre-collected from sources such as historical data or unknown behavioural policies. Contrary to its online counterpart, offline RL is particularly suitable for the real world as there is no need for access to the environment. Nonetheless, one of the main challenges of this setting, known as distributional shift, is finding a balance between increased generalization and avoiding previously unseen (out-of-distribution) states and actions (Levine et al., 2020; Wang et al., 2021; Kidambi et al., 2020).

In this paper, the focus is on offline reinforcement learning, due to its practical applicability in real-world scenarios.

2.1 Model-Based versus Model-Free Methods in Offline RL

Generally, reinforcement learning algorithms are separated into two main categories, namely model-free and model-based (Remonda et al., 2021). These two methodologies diverge in their strategies for acquiring and utilizing information to optimize decision-making processes within an environment. Model-free reinforcement learning prioritizes empirical learning through direct interaction with the environment. Rather than formulating an explicit representation of the environment's dynamics, model-free algorithms focus on learning optimal policies solely based on observed experiences. Although generally favoured due to its simplicity, MFRL has major limitations when

it comes to sample efficiency, as such algorithms often require extensive exploration or datasets to converge to optimal policies, leading to slow learning rates (Luo et al., 2022; Remonda et al., 2021). Some offline model-free RL algorithms deal with out-of-distribution (OOD) actions by constraining the policy search within the support of the static offline dataset via importance sampling (Precup et al., 2001; Sutton et al., 2015; Liu et al., 2019; Nachum et al., 2019; Gelada and Bellemare, 2019) or policy constraints (Fujimoto et al., 2018; Kumar et al., 2020; Wu et al., 2019; Peng et al., 2019; Kostrikov et al., 2021; Wang et al., 2020; Laroche and Trichelair, 2017; Liu et al., 2020b). Other algorithms learn conservative critics (Lu et al., 2021; Ma et al., 2021b; Kumar et al., 2021; Ma et al., 2021a), quantify uncertainty (Wu et al., 2021; Zanette et al., 2021; Deng et al., 2021) or model the trajectories in a sequential manner (Chen et al., 2021; Meng et al., 2021). Nonetheless, due to the heavy constraints and the sample complexity issue, these methods have poor generalization capabilities (Wang et al., 2021; Yarats et al., 2019).

Model-based RL, on the other hand, emphasizes the construction of an explicit model representing the environment's dynamics. This model captures the transitions between states and the corresponding rewards, enabling agents to simulate potential trajectories and plan actions accordingly. Such methods attain excellent sample efficiency. By leveraging the learned model, model-based methods can perform more effective planning and decision-making, leading to potentially faster convergence (Luo et al., 2022). Moreover, MBRL algorithms have the added benefit of generalizing knowledge more effectively across similar states or tasks, as the learned model encapsulates the underlying dynamics of the environment (Wang et al., 2021). Nonetheless, such algorithms are generally more computationally expensive and their performance heavily relies on the accuracy of the dynamics model (Remonda et al., 2021).

As distributional shift remains the most influential problem in offline model-based reinforcement learning, recent advancements in the field have focused on addressing namely this issue. Some offline MBRL algorithms handle it by introducing constraints on the model through modifications in state transition dynamics, reward functions, or value functions (He, 2023; Janner et al., 2021; Li et al., 2022; Matsushima et al., 2020; Yu et al., 2021; Rigter et al., 2022; Bhardwaj et al., 2023). Another widely adopted strategy in recent model-based RL literature is to further mitigate distributional shift by learning ensembles of (typically forward) dynamics models, used to estimate uncertainty (Yu et al., 2020; Yu et al., 2021; Kidambi et al., 2020; Rigter et al., 2022; Lowrey et al., 2018; Ovadia et al., 2019; Diehl et al., 2021). These uncertainty estimates encourage the agent to stay in states of low uncertainty by heavily penalizing it when visiting areas, where the model is uncertain. However, (inaccurate) uncertainty estimates can lead to overgeneralization in out-of-support regions (Wang et al., 2021). In other words, such conservatism quantifications can overestimate some unknown states and mislead forward model-based imaginations to undesired areas (Wang et al., 2021). A newly researched way of mitigating distributional shift in the context of offline MBRL is the learning of reverse dynamics models (Wang et al., 2021; Lyu et al., 2022; Park and Wong, 2022; Jain and Ravanbakhsh, 2023), as it adds a new layer of conservatism.

2.2 Reverse Dynamics Models

The idea of learning a reverse dynamics model (also called backward dynamics model) to generate imagined reversed samples first emerged in the literature of online RL algorithms (Holyoak and Simon, 1999; Goyal et al., 2018; Edwards et al., 2018; Lai et al., 2020; Lee et al., 2020). It has been shown to speed up learning, improve sample efficiency by aimed exploration, benefit planning for credit assignment (Hasselt et al., 2019; Chelu et al., 2020) and robustness (Jafferjee et al., 2020). Lai et al. (2020) (Lai et al., 2020) utilize a reverse model to reduce the dependence on accuracy in forward model predictions. Having similar motivation, Lee et al. (2020) (Lee et al., 2020) learn a backward dynamics model to capture contextual information while mitigating the risk of overly focusing on predicting only the forward dynamics. In contrast to the backward model in online RL, Wang et al. (2021) (Wang et al., 2021) propose to diversify the offline dataset with reverse imaginations to induce conservatism bias with data augmentation. Their proposed framework, called Reverse Offline Model-based Imagination (ROMI) marks the start of reverse dynamics modelling in the realm of MBORL.

ROMI learns a reverse dynamics model in conjunction with a novel reverse policy, which can generate rollouts leading to the target goal states within the offline dataset. The authors show that these reverse imaginations provide informed data augmentation, therewith diversifying the offline dataset. Based on this idea, Lyu et al. (2022) (Lyu et al., 2022) developed a method, that incorporates learning both a forward and a reverse dynamics model (also known as bidirectional dynamics model) with the purpose of introducing conservatism into transition. Their framework, called Confidence-Aware Bidirectional offline

model-based Imagination (CABI) is based on a double checking mechanism, which ensures the forward imagination is reasonable by generating a reverse imagination from it. In other words, only samples that both the forward and reverse models agree on are trusted and therefore, included in the augmented dataset. Park & Wong (2022) (Park and Wong, 2022), propose a method for goal-conditioned RL, called Backwards Model-based Imitation Learning (BMIL), which utilizes a reverse-time generative dynamics model that can generate possible paths leading the agent back onto the dataset trajectories. BMIL pairs a backwards dynamics model with a policy and is being trained on both offline data and imagined model rollouts. These reverse rollouts provide useful information since every rollout ends within the support of the offline dataset.

2.3 Cross-Domain Generalization

Another area of research, primary explored in online RL, is the topic of cross-domain generalization, defined as generalization across environments with varying transition dynamics, initial states or reward functions (Mediratta et al., 2023). One noteworthy contribution to the field is a method called Context-aware Dynamics Model (CaDM), where context refers to the dynamics of the environment (e.g. pole lengths in CarPole or body mass in Hopper, Figure 2) (Lee et al., 2020). First, it uses a context encoder to capture the contextual information from a recent experience. Then, an online adaptation to the unseen environment dynamics is performed by conditioning the forward dynamics model on the context. While there is a large body of work focused on evaluating and improving the generalization of online RL approaches (Packer et al., 2018; Cobbe et al., 2018; Zhang et al., 2018; Cobbe et al., 2019; Kuttler et al., 2020; Raileanu et al., 2021; Jiang et al., 2021; Raileanu and Fergus, 2021), just recently it began attracting more research interest in the offline setting, which is particularly suitable for the real-world deployment of RL. Specifically, the majority of current research is centralized around developing better performing offline RL methods (Mediratta et al., 2023; Levine et al., 2020; Prudencio et al., 2022; Fujimoto et al., 2018; Wu et al., 2019; Agarwal et al., 2019; Nair et al., 2020; Fujimoto and Gu, 2021; Zanette et al., 2021; Rashidinejad et al., 2021; Zhang et al., 2021; Lambert et al., 2022; Yarats et al., 2022; Brandfonbrener et al., 2022; Cheng et al., 2022), rather than better generalizable offline RL agents. Mediratta et al. (2023) (Mediratta et al., 2023) demonstrated that some benchmarked offline RL methods, namely Batch-Constrained deep Q-learning (BCQ) (Fujimoto et al., 2018), Conservative Q-Learning (CQL) (Kumar et al., 2020), Implicit Q-Learning (IQL) (Kostrikov et al., 2021), Behavioral Cloning Transformer (BCT) (Chen et al., 2021), and Decision Transformer (DT) (Chen et al., 2021), exhibit poor cross-domain generalization capabilities, highlighting the need for developing offline learning methods which generalize better to new environments. Another important finding of this work is that the diversity of the data, rather than its size, improves performance on new environments (Mediratta et al., 2023).

To bridge the gap between cross-domain generalization and offline RL, Liu et al. (2022) (Liu et al., 2022) propose Dynamics-Aware Rewards Augmentation (DARA). DARA is evaluated according to their newly introduced cross-domain setup, consisting of offline RL datasets with dynamics (mass, joint) shift compared to the original D4RL datasets. The modified datasets are used for training, while the original D4RL datasets are used for testing, therewith evaluating cross-domain generalization capabilities due to the varying transition dynamics. By augmenting rewards in the training dataset, DARA can acquire an adaptive policy in testing time, which results in consistently stronger performance when compared to prior offline RL methods (Liu et al., 2022). DARA is similar to another method designed for the cross-domain setup it introduces, namely Beyond OOD State Actions (BOSA). The core concept of BOSA is to address the intrinsic offline extrapolation error by focusing on OOD state-actions and OOD transition dynamics. The aim is to filter out offline transitions that could lead to a shift in state-actions or a mismatch in transition dynamics (Liu et al., 2023).

3 THEORETICAL BACKGROUND

This section outlines the theoretical background relevant to the newly proposed method, Context-Aware Reverse Imagination (CARI). Consequently, a use case describing the general idea of ROMI is provided, as CARI is largely based on ROMI.

3.1 Preliminaries

In general, reinforcement learning addresses the problem of learning to control a dynamical system. The dynamical system is defined by a fully-observed or partially-observed Markov decision process (MDP).

Following the definitions in Sutton and Barto (1998) (Sutton and Barto, 2018), the Markov decision process is defined as a tuple $M = (S, A, p, r, \gamma, \rho_0)$,

(a) Thin (b) Normal (c) Thick

Figure 2: Different body masses of Hopper.

where S is a set of states $s \in S$, which may be either discrete or continuous (i.e. multi-dimensional vectors), A is a set of actions $a \in A$, which similarly can be discrete or continuous, p defines a conditional probability distribution of the form $p(s'|s,a)$, which describes the dynamics of the system, where s' is the next state after taking action a at the current state s. $r : S \times A \to \mathbb{R}$ defines a reward function, $\gamma \in (0,1]$ is a scalar discount factor and ρ_0 defines the initial state distribution. In the case of offline RL, the trajectories represented by the tuples (s,a,r,s') are stored in a static dataset $\mathcal{D}_{\text{env}}$. The goal of RL is to optimize a policy $\pi(a|s)$ that maximizes the expected discounted return defined as $\mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(s_t,a_t)]$.

3.2 Problem Formulation

The standard offline RL framework is considered, where an agent learns from an offline static dataset. Formally, the problem is formulated as an MDP, following the definitions introduced in Section 3.1. This is further tackled in the context of model-based RL by learning either a forward dynamics model $f = \hat{p}(s'|s,a)$ or a reverse (backward) dynamics model $b = \hat{p}(s|s',a)$, which approximate the true transition dynamics $p(s'|s,a)$ and $p(s|s',a)$, respectively. In order to address the problem of cross-domain generalization, the distribution of MDPs is further considered, where the transition dynamics $p_c(s'|s,a)$ varies according to a context c. For instance, consider a change in the transition dynamics of Hopper by modifying its environment parameters (e.g.mass, Figure 2).

This study concerns the development of an offline MBRL approach that learns a reverse dynamics model, capable of generalization, which is robust to such dynamics changes, i.e., approximating a distribution of transition dynamics. Specifically, given a set of training environments with contexts sampled from $p_{train}(c)$, the aim is to learn a reverse dynamics model that can retain good cross-domain generalization, i.e. produce accurate predictions for test environments with unseen contexts sampled from $p_{test}(c)$.

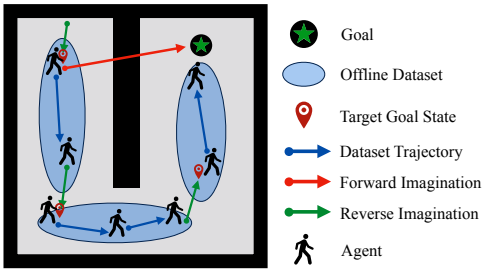


Figure 3: General idea of ROMI.

3.3 General Idea of ROMI

To illustrate the general idea of ROMI, consider Figure 3. The agent navigates in a U-shaped maze to reach the goal. The precollected offline dataset is represented by the blue ovals and its trajectories are denoted by blue arrows. This dataset does not contain any trajectories that hit the walls and thus, the agent will be unaware of them during the learning process. Hence, in this scenario, a standard (forward) dynamics model might greedily generate an invalid imagination with overestimated value with the aim of finding a better route to the goal state (red arrow). Contrary to forward imaginations, ROMI generates trajectories that lead to the target goal states within the offline dataset (green arrows). Further, those reverse imaginations can be connected with the dataset trajectories to create more diverse or even optimal policies.

Translating this example to a real-world scenario, consider an offline dataset collected by an expert behaviour policy, where the agent is a vacuum cleaning robot. The robot's task is to learn to reach its charging station (i.e. the goal state) when its battery life is low. Thus, in this case, the agent has to learn this task only from successful trajectories that avoid bumping into walls or furniture. When using a forward dynamics model, the agent can generate aggressive rollouts from the dataset to outside areas. Such forward imaginations can potentially discover better routes to the charging station, but can also guide the vacuum cleaner towards an obstacle, rather than the charger, due to overestimation. If the agent has learned a reverse dynamics model instead, then the reverse imaginations generate possible traces leading to targeted states inside the offline dataset, therefore providing a conservative way of augmenting the offline dataset. These are useful not only because the agent will not be guided towards obstacles, which in turn will impede learning, but also because such conservative rollouts can merge existing trajectories eventually composing an optimal path to the charger.

4 TOWARD CROSS-DOMAIN GENERALIZATION

In the offline RL setting, the agent has access only to a given dataset, without the option to perform online exploration. In this setting, MBRL algorithms are usually challenged by (i) the limited samples of the given dataset, (ii) the uncertainties in the out-of-support areas and (iii) the inability to correct model inaccuracies by online interaction. Thus, it is of great importance to augment the dataset, while also keeping conservative generalization, as increasing the diversity of the data, rather than its size, is shown to improve performance with respect to generalization (Mediratta et al., 2023). In this section, Context-Aware Reverse Imagination (CARI) framework is introduced, that combines model-based imaginations with model-free offline policy learners. CARI builds upon ROMI by incorporating context-awareness into the dynamics model, while keeping the other two components, namely reverse rollout policy and the generation of rollouts for augmenting the offline dataset (see Figure 4). Thus, both frameworks promote diverse augmentation of model-based rollouts and enable conservative generalization of the generated imaginations. However, CARI captures the local dynamics (i.e. the context), and then predicts the previous state conditioned on it. This is favorable because learning a global model that can generalize across different dynamics is a challenging task.

4.1 Illustrative Example

As mentioned CARI is based on ROMI, which means that the augmentation properties of the latter are preserved. Hence why CARI retains the same superiority as ROMI, compared to a method that generates forward imaginations, as illustrated in Section 3.3 (see Figure 3). To demonstrate this further, consider the following example: Assume state s_{in} has a dataset trajectory leading to the goal (denoted by s_{goal}):

Dataset trajectory: $\langle s_{in}, \cdots, s_{goal} \rangle$

Suppose s_{in} is part of a forward imaginary trajectory (s_{in} is the starting state of the trajectory, while s_4 is the last state) and its reverse counterpart (s_4 is the starting state of the trajectory and s_{in} is the last state). Formally:

Forward imagination: $\langle s_{in}, a_1^f, s_1, a_2^f, s_2, a_3^f, s_3, a_4^f, s_4 \rangle$
Reverse imagination: $\langle s_4, a_4^b, s_3, a_3^b, s_2, a_2^b, s_1, a_1^b, s_{in} \rangle$

Note that both trajectories visit the same sequence of states but in reversed order, hence why the actions differ. During the training process, the reverse rollout will expand the dataset trajectory from $\langle s_{in}, \cdots, s_{goal} \rangle$ to $\langle s_4, \cdots, s_{in}, \cdots, s_{goal} \rangle$. In other words, the sequence $\langle s_1 : s_4 \rangle$ can now reach the goal. Thus, the policy learning in this task can be enhanced by the reverse rollout through the reverse imagination of $\langle s_1 : s_4 \rangle$. However, for the forward rollout, there are three cases for the state s_4:

(i) State s_4 is out-of-support and its value is overestimated: If this is the case, the forward rollout can mislead the policy from s_{in} to s_4 and impede the learning process by diverging from the goal. On the other hand, the reverse trajectory cannot make such a negative impact since s_{in} does not reach s_4 (but rather, s_4 reaches s_{in}).

(ii) State s_4 is out-of-support and its value is not overestimated: In this case, the forward trajectory does not impact the policy learning, as the policy will not go to s_4 with a lower value. As elaborated above, the reverse rollout will benefit the learning, because of its effective trajectory expansion.

(iii) State s_4 is in the support of the dataset: If s_4 has a trajectory leading to the goal, then the forward imaginary trajectory can improve the learning by selecting the better trajectory between $\langle s_{in}, \cdots, s_{goal} \rangle$ and $\langle s_{in}, \cdots, s_4, \cdots, s_{goal} \rangle$. In this case, the reverse augmentation has a single successful trajectory starting at state s_{in}, $\langle s_{in}, \cdots, s_{goal} \rangle$. However, for this situation neither of the models has to deal with the conservatism issue, since s_4 is within the dataset support.

4.2 Context-Aware Reverse Imagination (CARI)

Training the Context-Aware Reverse Dynamics Model. To make the dynamics model context-aware, CARI separates the task of reasoning about the environment dynamics into (i) learning the dynamics-specific information (a latent vector c), and (ii) predicting the next state conditioned on the latent vector (transition inference). This is done by introducing an additional neural network head to the standard two-head architecture (one head for transition inference and one for the variance). The loss function puts pressure on the context head to produce a context vector that improves prediction accuracy, as similarly done in (Lee et al., 2020). The reverse model estimates the reverse dynamics model $\hat{p}(s|s', a, c)$ and rewards model $\hat{r}(s, a)$ simultaneously. For simplicity, the dynamics and reward function are unified into the reverse model $p(s, r|s', a, c)$. The output predictions are explicitly conditioned on the context vector. If the context vector does not capture relevant information,

the model's predictions will be inaccurate, leading to higher loss (Eq. 1). This unified model represents the probability of the current state and immediate reward conditioned on the next state, current action and learned context. It is parameterized by ϕ and optimized by maximizing the log-likelihood:

$$\mathcal{L}^{bwd}(\phi) = \mathop{\mathbb{E}}_{(s,a,r,s')\sim\mathcal{D}_{\texttt{env}}} [-\log \hat{p}_\phi(s,r,|s',a,c)] \quad (1)$$

Training the Reverse Rollout Policy. Just like ROMI, diversity in reverse model-based imaginations near the dataset is encouraged by training a generative model $\hat{G}_\theta(a|s')$, which samples diverse reverse actions from $\mathcal{D}_{env}$ using stochastic inference. Specifically, a conditional variational autoencoder is utilized to train the diverse rollout policy , represented by $\hat{G}_\theta(a|s')$. The rollout policy is trained to maximize the variational lower bound:

$$\mathcal{L}_{\texttt{bvae}}(\theta) = \mathop{\mathbb{E}}_{(s,a,r,s')\sim\mathcal{D}_{\texttt{env}},z\sim\hat{\mathbb{E}}_\omega(s',a)} \Big[\Big(a-\hat{D}_\zeta(s',z)\Big)^2 + D_{KL}\Big(\hat{E}_\omega(s',a)\|\mathcal{N}(0,\boldsymbol{I})\Big)\Big], \quad (2)$$

where $\hat{E}_\omega(s',a)$ is an action encoder that produces latent vector z under the multivariate normal distribution $\mathcal{N}(0,\boldsymbol{I})$ with $\boldsymbol{I}$ being an identity matrix, and $\hat{D}_\zeta(s',z)$ is the action decoder.

Combination with Model-Free Algorithms. Based on the learned context-aware reverse dynamics model and the reverse rollout policy, CARI can generate reverse rollouts. These reverse imaginations are collected and stored in a model-based buffer $\mathcal{D}_{\texttt{model}}$. This buffer is further combined with the original offline dataset $\mathcal{D}_{\texttt{env}}$ to compose the final augmented dataset $\mathcal{D}_{\texttt{total}}$, i.e. $\mathcal{D}_{\texttt{total}} = \mathcal{D}_{\texttt{env}} \cup \mathcal{D}_{\texttt{model}}$. By design, $\mathcal{D}_{\texttt{total}}$ is obtained before the policy learning stage, therefore CARI can be combined with any model-free offline RL algorithm such as BCQ (Fujimoto et al., 2018).

Figure 4 illustrates the CARI framework, while Algorithm 1 details its training procedure.

Input: Offline dataset $\mathcal{D}_{\texttt{env}}$, rollout horizon h, the number of iterations C_ϕ, C_θ, T, learning rates α_ϕ, α_θ, offline MFRL algorithm
Result: $\pi_{\texttt{out}}$
Randomly initialize reverse model params ϕ;
for $i = 0,...,C_\phi - 1$ **do**
  Prepare inputs (s',a) and targets (s,r) from the dataset $\mathcal{D}_{\texttt{env}}$;
  Get context latent vector c;
  Condition the next state prediction on the context latent vector c;
  Compute $\mathcal{L}^{bwd}$;
  Update $\phi \leftarrow \phi - \alpha_\phi \Delta_\phi \mathcal{L}^{bwd}$;
end
Randomly initialize rollout policy params θ;
for $i = 0,...,C_\theta - 1$ **do**
  Compute $\mathcal{L}_{\texttt{bvae}}$ using the dataset $\mathcal{D}_{\texttt{env}}$;
  Update $\theta \leftarrow \theta - \alpha_\theta \Delta_\theta \mathcal{L}_{\texttt{bvae}}$;
end
Initialize the replay buffer $\mathcal{D}_{\texttt{model}} \leftarrow \emptyset$;
for $i = 0,...,T-1$ **do**
  Sample target state s_{t+1} from $\mathcal{D}_{\texttt{env}}$;
  Generate $\{(s_{t-i}, a_{t-i}, r_{t-i}, s_{t+1-i})\}_{i=0}^{h-1}$ from s_{t+1} by drawing samples from the dynamics model and rollout policy;
  $\mathcal{D}_{\texttt{model}} \leftarrow \mathcal{D}_{\texttt{model}} \cup \{(s_{t-i}, a_{t-i}, r_{t-i}, s_{t+1-i})\}_{i=0}^{h-1}$;
end
Compose the final $\mathcal{D}_{\texttt{total}} \leftarrow \mathcal{D}_{\texttt{env}} \cup \mathcal{D}_{\texttt{model}}$;
Combine the offline MFRL algorithm to derive the final policy $\pi_{\texttt{out}}$ using $\mathcal{D}_{\texttt{total}}$;

Algorithm 1: CARI.

5 EMPIRICAL EVALUATION

This section outlines the conducted experiments and provides an answer to the following questions: (i) How well do existing offline MBRL approaches, that learn reverse dynamics models, perform with respect to cross-domain generalization (see Table 1)? (ii) What approach-specific characteristics hinder or contribute to cross-domain generalization abilities, and what are the associated trade-offs (see Section 5.2.1)? (iii) Does CARI outperform all methods considered in this study (see Table 1)?

5.1 Experimental Setup

The considered methods, namely ROMI, CABI and BMIL, and the proposed framework, CARI, are evaluated against each other in cross-domain offline RL settings, following (Liu et al., 2022; Liu et al., 2023). Importantly, ROMI, CABI and CARI are model-based methods that incorporate a model-free component. Thus, for the aims of this study all three methods are combined with BCQ as a model-free policy learner. All methods share the same values for the same set of hyperparameters across all environment variants. The rollout length is set to 5 for all environments and all methods. The study focuses on two Gym-MuJoCo environments, Hopper and Walker2D (see Figure 1), such that each environment has four versions: Random, Medium, Medium-Expert and Medium-Replay. The Random datasets contain experiences collected with a random policy. The Medium datasets contain experiences from an early-stopped

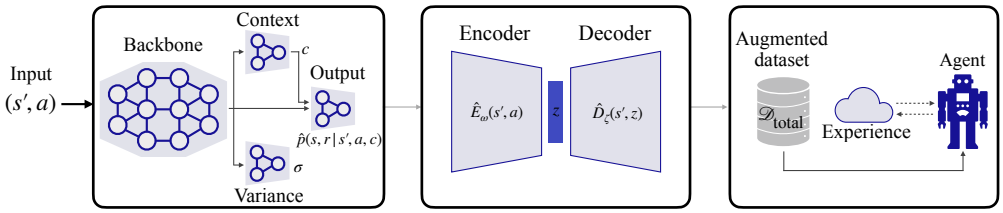


Figure 4: Illustration of CARI. First a context-aware reverse dynamics model is learned (left), which takes the next state and current action as input. Then a diverse rollout policy, used to sample diverse rollouts, is trained, represented by a CVAE (middle). Finally, the generated rollouts serve as augmentations to the original dataset, used by a MFRL algorithm to train the learning policy (right).

SAC policy. The Medium-Replay datasets record the samples in the replay buffer during the training of the Medium SAC policy. The Medium-Expert datasets mix sub-optimal samples with samples generated from an expert policy. As the main idea of cross-domain generalization is to have the ability to retain strong performance in previously unseen environmental settings, then the experimental setup involves both a source and a target dataset. The source dataset is used for training, while the target one is used for testing, such that both datasets differ in terms of their environmental dynamics. This study adopts the proposed setup by Liu et al. (2022) (Liu et al., 2022), which involves using offline samples from D4RL as target offline dataset. For the source dataset, the authors have changed the body mass of agents or added joint noise to the motion, and, similar to D4RL, collected the Random, Medium, Medium-Replay and Medium-Expert offline datasets for the environments (Liu et al., 2022). However, for the purposes of this study only the datasets with shifted body masses are considered for both Hopper and Walker2D. Specifically, the source datasets are comprised of 10% D4RL data and 100% of the collected source offline data (Liu et al., 2022).

Each of the methods is evaluated on the eight variants on five different random seeds. The evaluation criterion is the normalized scores metric, suggested by D4RL benchmark (Fu et al., 2020), where 0 indicates a random policy performance and 100 corresponds to an expert performer. The code of this work is available at https://github.com/YanasGH/CDG.

5.2 Overall Performance

The three baselines that learn a reverse dynamics model, namely BMIL, CABI and ROMI, are compared against a cross-domain baseline, DARA. Like CABI and ROMI, DARA can incorporate a model-free part. Thus, DARA+BCQ (referred to as DARA) is selected to ensure fair comparison, as both ROMI and CABI are combined with BCQ for the purposes of this study. The results for DARA are directly obtained from the source paper (Liu et al., 2022).

For seven out of the eight environment variants, DARA outperforms BMIL. Only for Walker2D Medium-Expert the performance of BMIL exceeds the one of DARA. CABI performs better than BMIL and manages to improve the scores achieved by DARA on four environments. However, it matches the performance of DARA for one environment, while it underperforms for the three left. ROMI, on the other hand, strongly outperforms DARA for all environments.

Compared to the strongest performer of the three reverse baselines, namely ROMI, CARI manages to further improve the score for both Hopper Medium-Expert and Walker2D Medium-Replay. Complete results of all methods considered are given in Table 1, while the following sections contain an elaborate discussion.

5.2.1 Approach-Specific Characteristics Influencing Cross-Domain Generalization

To provide a multifaceted view of the algorithms' cross-domain generalization capabilities, four versions of each environment are considered. The version with the most unstructured data is the Random one. Hence why the goal-conditioned method, BMIL, fails. BMIL relies on successful trajectories to learn a task effectively. For this reason, it is expected that this method fails to score high in general. CABI, on the contrary, manages to improve the performance compared to BMIL, especially in Medium and Medium-Expert environments. Notably, these respectively high scores highlight CABI's ability to leverage more structured datasets effectively. However, CABI still does not convincingly outperform DARA. One possible explanation for this could be the influence of the forward dynamics model in CABI in combination with the fact that CABI is not specifically designed for cross-domain settings, unlike DARA. CABI incorporates a double-checking mechanism to perform the data augmentation such that the only admitted imaginations are the ones confirmed by both the forward

Table 1: Performance of BMIL, CABI, ROMI, DARA and CARI for each of the environments, on the normalized return metric. Results are averaged over five random seeds and standard deviation is reported (highest scores in bold). Also, results of FOMI are included, such that its scores are compared only against ROMI's and FOMI's outperforming results are in italic.

Environment		BMIL	CABI	ROMI	DARA	CARI	FOMI
Hopper	Random	2.3 ± 0.0	13.1 ± 0.3	**24.2 ± 3.5**	9.7	**27.3 ± 2.3**	12.7 ± 1.8
	Medium	14.5 ± 0.1	40.6 ± 2.7	**47.1 ± 5.2**	38.4	**46.3 ± 0.9**	*70.1 ± 8.7*
	Medium-Expert	48.4 ± 1.2	63.6 ± 3.5	106.3 ± 1.4	84.2	**110.1 ± 0.5**	95.2 ± 6.3
	Medium-Replay	12.3 ± 0.1	30.3 ± 1.1	**46.2 ± 3.4**	32.8	**48.3 ± 1.9**	42.6± 5.8
Walker2D	Random	1.4 ± 0.0	0.9 ± 0.2	**12.4 ± 8.0**	4.8	**9.9 ± 4.5**	5.4 ± 0.2
	Medium	37.0 ± 2.5	80.5 ± 0.8	**87.0 ± 0.5**	52.3	**87.5 ± 0.3**	87.2 ± 0.2
	Medium-Expert	61.9 ± 3.8	88.3 ± 0.3	**93.3 ± 0.3**	57.2	**92.8 ± 0.6**	93.1 ± 0.3
	Medium-Replay	10.0 ± 0.0	62.6 ± 2.7	84.1 ± 0.5	15.1	**85.9 ± 1.2**	84.3± 1.5

and reverse dynamics models. Therefore, this added conservatism can serve as an explanation of the questionable performance of CABI.

On the other hand, the most distinguished method, ROMI, does not depend only on successful trajectories, neither does it learn a forward dynamics. Notably, in the Random environments, ROMI performs impressively well. This serves as an indication that this method learns to generalize very well despite unstructured data. Importantly, the strong cross-generalization abilities of ROMI are present even in structured data, as it is the case for the Medium and Medium-Expert environments for both Hopper and Walker2D. These overall high scores serve as a proof of concept that the combination of learning a reverse dynamics model and diverse generative rollout policy leads to an augmented dataset with extreme diversity, needed for strong cross-domain generalization.

In analyzing the comparative performance of CARI against the reverse baselines across Hopper and Walker2D domains, several noteworthy observations arise that highlight its relative strengths and weaknesses in terms of cross-domain generalization capabilities. The results from the Hopper domain underscore CARI's superior performance across varying dataset conditions. Specifically, CARI achieves the highest performance in the Medium-Expert condition (110.1 ± 0.5), demonstrating a marked advantage over BMIL, CABI, ROMI, and DARA in this engineered data scenario. This robust performance suggests that CARI is particularly adept at leveraging the structural insights provided by structured data, thereby achieving higher efficacy in task execution. Since CARI is incremental work of ROMI, these results showcase the positive impact of context-awareness. In comparison, while ROMI also exhibits strong performance in the Medium-Expert condition, its results are characterized by higher variability. This variability points to potential sensitivity to data randomness, which CARI appears to handle more effectively.

Notably, CARI maintains a more stable performance profile across all conditions, indicating a balanced trade-off between adaptability and consistency. CABI does not reach the same level of excellence as CARI, regardless of its strong performance in Walker2D Medium and Walker2D Medium-Expert variants. Similarly, the results of BMIL, despite being the most stable of all, are not comparable to the ones of CARI, as the performance gap is quite pronounced in every setting.

In summary, CARI demonstrates a clear edge over the other methods in terms of cross-domain generalization abilities due to its consistently high performance across both domains. While other methods like ROMI and CABI show strong performance in specific conditions, their variability and sensitivity to data structure limit their overall efficacy compared to CARI. These findings highlight CARI's superior balance of performance, adaptability, and consistency, thereby making it a preferred choice for tasks requiring robust cross-domain generalization.

5.3 Ablation

Two ablation studies are performed to provide a deeper insight into (i) the effect of a reverse dynamics model (versus its forward counterpart), as well as (ii) the influence of rollout length on CARI.

5.3.1 Ablation Study with Model-Based Imagination

To investigate whether ROMI's strong performance is only due to the reverse model-based imagination, an ablation study is performed to compare ROMI with its forward counterpart, namely Forward Offline Model-based Imagination (FOMI). Specifically, the reverse imaginations are substituted with ones in the forward direction. To study the performance of FOMI, the same two environments with the same four variants are used. To ensure fair comparison, FOMI is combined with BCQ, as done in ROMI.

Table 1 shows that ROMI consistently matches or outperforms FOMI for all environments except for

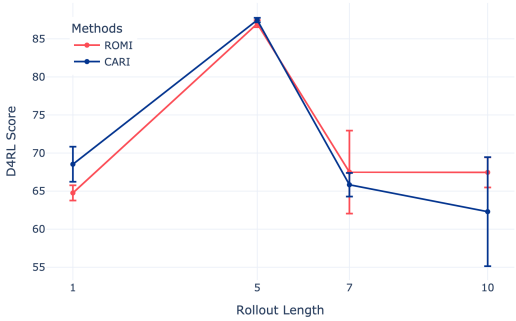


Figure 5: Performance comparison between ROMI and CARI with varying rollout length for Walker2D Medium.

Hopper Medium. This performance gap is particularly pronounced in the Random environments, underscoring the limitations of forward dynamics models in handling unstructured settings. These overall results imply that the reverse imaginations play a crucial role for ROMI when it comes to cross-domain generalization. However, it should be noted that although FOMI comes generally short compared to ROMI, it still retains very strong performance. The most plausible reason for this is the diverse rollout policy.

5.3.2 Rollout Horizon Length Experimentation

Additional experiments are conducted to investigate the effect of rollout length more clearly. This hyperparameter is varied on the Walker2D Medium environment, as the results for CARI and ROMI are comparable. The results are vizualized in Figure 5, which illustrates the aggregated normalized scores over five runs, where the error bars are the standard deviation. The findings point that both methods follow the same trend, and their respective performances peak at rollout length of 5. Nonetheless, it can be seen that CARI performs slightly better compared to ROMI for shorter rollout length, while ROMI seems to have a more accurate model for longer horizon imagination. It should be noted that for this specific environment it seems the learning of the context does not provide a significant impact on the performance for any horizon length, possibly due to the transitions being more predictable.

6 CONCLUSION

This paper investigates the cross-domain generalization capabilities of BMIL, CABI and ROMI. While BMIL struggled to outperform the cross-domain baseline DARA, CABI and ROMI showed consistent improvements. Most notably, ROMI with its combination of learning a reverse dynamics model and diverse generative rollout policy lead to an augmented dataset with extreme diversity, needed to ensure strong generalization abilities. These results serve as a proof of concept that offline MBRL methods that learn a reverse dynamics model exhibit overall good capabilities with respect to cross-domain generalization. Importantly, this paper introduces CARI, a novel MBORL framework, that builds on ROMI by making the reverse dynamics model context-aware, therewith improving its cross-domain generalization.

As future work, one potential direction is to combine CARI with other MFRL algorithms, such as CQL, as this can further improve its performance. Another interesting possibility is to explore trajectory augmentation techniques for further diversification of the augmented dataset.

REFERENCES

Agarwal, R., Schuurmans, D., and Norouzi, M. (2019). An Optimistic Perspective on Offline Reinforcement Learning. In *International Conference on Machine Learning*.

Bhardwaj, M., Xie, T., Boots, B., Jiang, N., and Cheng, C.-A. (2023). Adversarial Model for Offline Reinforcement Learning. *ArXiv*, abs/2302.11048.

Brandfonbrener, D., Bietti, A., Buckman, J., Laroche, R., and Bruna, J. (2022). When does return-conditioned supervised learning work for offline reinforcement learning? *ArXiv*, abs/2206.01079.

Chelu, V., Precup, D., and Hasselt, H. V. (2020). Forethought and Hindsight in Credit Assignment. *ArXiv*, abs/2010.13685.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. (2021). Decision Transformer: Reinforcement Learning via Sequence Modeling. In *Neural Information Processing Systems*.

Cheng, C.-A., Xie, T., Jiang, N., and Agarwal, A. (2022). Adversarially Trained Actor Critic for Offline Reinforcement Learning. *ArXiv*, abs/2202.02446.

Cobbe, K., Hesse, C., Hilton, J., and Schulman, J. (2019). Leveraging Procedural Generation to Benchmark Reinforcement Learning. In *International Conference on Machine Learning*.

Cobbe, K., Klimov, O., Hesse, C., Kim, T., and Schulman, J. (2018). Quantifying Generalization in Reinforcement Learning. In *International Conference on Machine Learning*.

Deng, Z.-H., Fu, Z., Wang, L., Yang, Z., Bai, C., Wang, Z., and Jiang, J. (2021). SCORE: Spurious COrrelation REduction for Offline Reinforcement Learning. *ArXiv*, abs/2110.12468.

Diehl, C. P., Sievernich, T., Krüger, M., Hoffmann, F., and Bertram, T. (2021). UMBRELLA: Uncertainty-Aware Model-Based Offline Reinforcement Learning Leveraging Planning. *ArXiv*, abs/2111.11097.

Dulac-Arnold, G., Mankowitz, D. J., and Hester, T. (2019). Challenges of Real-World Reinforcement Learning. *ArXiv*, abs/1904.12901.

Edwards, A. D., Downs, L., and Davidson, J. C. (2018). Forward-Backward Reinforcement Learning. *ArXiv*, abs/1803.10227.

Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. (2020). D4RL: Datasets for Deep Data-Driven Reinforcement Learning. *ArXiv*, abs/2004.07219.

Fujimoto, S. and Gu, S. S. (2021). A Minimalist Approach to Offline Reinforcement Learning. *ArXiv*, abs/2106.06860.

Fujimoto, S., Meger, D., and Precup, D. (2018). Off-Policy Deep Reinforcement Learning without Exploration. In *International Conference on Machine Learning*.

Gelada, C. and Bellemare, M. G. (2019). Off-Policy Deep Reinforcement Learning by Bootstrapping the Covariate Shift. In *AAAI Conference on Artificial Intelligence*.

Goyal, A., Brakel, P., Fedus, W., Singhal, S., Lillicrap, T. P., Levine, S., Larochelle, H., and Bengio, Y. (2018). Recall Traces: Backtracking Models for Efficient Reinforcement Learning. *ArXiv*, abs/1804.00379.

Hasselt, H. V., Hessel, M., and Aslanides, J. (2019). When to use parametric models in reinforcement learning? *ArXiv*, abs/1906.05243.

He, H. (2023). A Survey on Offline Model-Based Reinforcement Learning. *ArXiv*, abs/2305.03360.

Holyoak, K. J. and Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General*, 128:3–31.

Jafferjee, T., Imani, E., Talvitie, E. J., White, M., and Bowling, M. (2020). Hallucinating Value: A Pitfall of Dyna-style Planning with Imperfect Environment Models. *ArXiv*, abs/2006.04363.

Jain, V. and Ravanbakhsh, S. (2023). Learning to Reach Goals via Diffusion. *ArXiv*, abs/2310.02505.

Janner, M., Li, Q., and Levine, S. (2021). Offline Reinforcement Learning as One Big Sequence Modeling Problem. In *Neural Information Processing Systems*.

Jiang, M., Dennis, M., Parker-Holder, J., Foerster, J. N., Grefenstette, E., and Rocktaschel, T. (2021). Replay-Guided Adversarial Environment Design. In *Neural Information Processing Systems*.

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. (2020). MOReL : Model-Based Offline Reinforcement Learning. *ArXiv*, abs/2005.05951.

Kostrikov, I., Nair, A., and Levine, S. (2021). Offline Reinforcement Learning with Implicit Q-Learning. *ArXiv*, abs/2110.06169.

Kumar, A., Agarwal, R., Ma, T., Courville, A. C., Tucker, G., and Levine, S. (2021). DR3: Value-Based Deep Reinforcement Learning Requires Explicit Regularization. *ArXiv*, abs/2112.04716.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. (2020). Conservative Q-Learning for Offline Reinforcement Learning. *ArXiv*, abs/2006.04779.

Kuttler, H., Nardelli, N., Miller, A. H., Raileanu, R., Selvatici, M., Grefenstette, E., and Rocktäschel, T. (2020). The NetHack Learning Environment. *ArXiv*, abs/2006.13760.

Lai, H., Shen, J., Zhang, W., and Yu, Y. (2020). Bidirectional Model-based Policy Optimization. *ArXiv*, abs/2007.01995.

Lambert, N., Wulfmeier, M., Whitney, W. F., Byravan, A., Bloesch, M., Dasagi, V., Hertweck, T., and Riedmiller, M. A. (2022). The Challenges of Exploration for Offline Reinforcement Learning. *ArXiv*, abs/2201.11861.

Laroche, R. and Trichelair, P. (2017). Safe Policy Improvement with Baseline Bootstrapping. In *International Conference on Machine Learning*.

Lee, K., Seo, Y., Lee, S., Lee, H., and Shin, J. (2020). Context-aware Dynamics Model for Generalization in Model-Based Reinforcement Learning. *ArXiv*, abs/2005.06800.

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. *ArXiv*, abs/2005.01643.

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. (2022). Settling the Sample Complexity of Model-Based Offline Reinforcement Learning. *ArXiv*, abs/2204.05275.

Liu, J., Zhang, H., and Wang, D. (2022). DARA: Dynamics-Aware Reward Augmentation in Offline Reinforcement Learning. *ArXiv*, abs/2203.06662.

Liu, J., Zhang, Z., Wei, Z., Zhuang, Z., Kang, Y., Gai, S., and Wang, D. (2023). Beyond OOD State Actions: Supported Cross-Domain Offline Reinforcement Learning. *ArXiv*, abs/2306.12755.

Liu, S., See, K. C., Ngiam, K. Y., Celi, L. A., Sun, X., and Feng, M. (2020a). Reinforcement Learning for Clinical Decision Support in Critical Care: Comprehensive Review. *Journal of Medical Internet Research*, 22.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. (2019). Off-Policy Policy Gradient with Stationary Distribution Correction. *ArXiv*, abs/1904.08473.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. (2020b). Provably Good Batch Off-Policy Reinforcement Learning Without Great Exploration. In *Neural Information Processing Systems*.

Lowrey, K., Rajeswaran, A., Kakade, S. M., Todorov, E., and Mordatch, I. (2018). Plan Online, Learn Offline: Efficient Learning and Exploration via Model-Based Control. *ArXiv*, abs/1811.01848.

Lu, C., Ball, P. J., Parker-Holder, J., Osborne, M. A., and Roberts, S. J. (2021). Revisiting Design Choices in Offline Model Based Reinforcement Learning. In *International Conference on Learning Representations*.

Luo, F., Xu, T., Lai, H., Chen, X.-H., Zhang, W., and Yu, Y. (2022). A Survey on Model-based Reinforcement Learning. *Sci. China Inf. Sci.*, 67.

Lyu, J., Li, X., and Lu, Z. (2022). Double Check Your State Before Trusting It: Confidence-Aware Bidirectional Offline Model-Based Imagination. *ArXiv*, abs/2206.07989.

Ma, X., Yang, Y., Hu, H., Liu, Q., Yang, J., Zhang, C., Zhao, Q., and Liang, B. (2021a). Offline Reinforcement Learning with Value-based Episodic Memory. *ArXiv*, abs/2110.09796.

Ma, Y. J., Jayaraman, D., and Bastani, O. (2021b). Conservative Offline Distributional Reinforcement Learning. In *Neural Information Processing Systems*.

Matsushima, T., Furuta, H., Matsuo, Y., Nachum, O., and Gu, S. S. (2020). Deployment-Efficient Reinforcement Learning via Model-Based Offline Optimization. *ArXiv*, abs/2006.03647.

Mediratta, I., You, Q., Jiang, M., and Raileanu, R. (2023). The Generalization Gap in Offline Reinforcement Learning. *ArXiv*, abs/2312.05742.

Meng, L., Wen, M., Yang, Y., Le, C., Li, X., Zhang, W., Wen, Y., Zhang, H., Wang, J., and Xu, B. (2021). Offline Pre-trained Multi-Agent Decision Transformer: One Big Sequence Model Tackles All SMAC Tasks. *ArXiv*, abs/2112.02845.

Nachum, O., Chow, Y., Dai, B., and Li, L. (2019). DualDICE: Behavior-Agnostic Estimation of Discounted Stationary Distribution Corrections. *ArXiv*, abs/1906.04733.

Nair, A., Dalal, M., Gupta, A., and Levine, S. (2020). Accelerating Online Reinforcement Learning with Offline Datasets. *ArXiv*, abs/2006.09359.

Ovadia, Y., Fertig, E., Ren, J. J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J. (2019). Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. In *Neural Information Processing Systems*.

Packer, C., Gao, K., Kos, J., Krähenbühl, P., Koltun, V., and Song, D. X. (2018). Assessing Generalization in Deep Reinforcement Learning. *ArXiv*, abs/1810.12282.

Park, J. Y. and Wong, L. L. S. (2022). Robust Imitation of a Few Demonstrations with a Backwards Model. *ArXiv*, abs/2210.09337.

Peng, X. B., Kumar, A., Zhang, G., and Levine, S. (2019). Advantage-Weighted Regression: Simple and Scalable Off-Policy Reinforcement Learning. *ArXiv*, abs/1910.00177.

Precup, D., Sutton, R. S., and Dasgupta, S. (2001). Off-Policy Temporal Difference Learning with Function Approximation. In *International Conference on Machine Learning*.

Prudencio, R. F., Maximo, M. R., and Colombini, E. L. (2022). A Survey on Offline Reinforcement Learning: Taxonomy, Review, and Open Problems. *IEEE transactions on neural networks and learning systems*, PP.

Raileanu, R. and Fergus, R. (2021). Decoupling Value and Policy for Generalization in Reinforcement Learning. *ArXiv*, abs/2102.10330.

Raileanu, R., Goldstein, M., Yarats, D., Kostrikov, I., and Fergus, R. (2021). Automatic Data Augmentation for Generalization in Reinforcement Learning. In *Neural Information Processing Systems*.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. J. (2021). Bridging Offline Reinforcement Learning and Imitation Learning: A Tale of Pessimism. *IEEE Transactions on Information Theory*, 68:8156–8196.

Remonda, A., Veas, E. E., and Luzhnica, G. (2021). Acting upon Imagination: when to trust imagined trajectories in model based reinforcement learning. *ArXiv*, abs/2105.05716.

Rigter, M., Lacerda, B., and Hawes, N. (2022). RAMBO-RL: Robust Adversarial Model-Based Offline Reinforcement Learning. *ArXiv*, abs/2204.12581.

Singh, B., Kumar, R., and Singh, V. P. (2021). Reinforcement learning in robotic applications: a comprehensive survey. *Artificial Intelligence Review*, 55:945 – 990.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA.

Sutton, R. S., Mahmood, A. R., and White, M. (2015). An Emphatic Approach to the Problem of Off-policy Temporal-Difference Learning. *J. Mach. Learn. Res.*, 17:73:1–73:29.

Wang, J., Li, W., Jiang, H., Zhu, G., Li, S., and Zhang, C. (2021). Offline Reinforcement Learning with Reverse Model-based Imagination. In *Neural Information Processing Systems*.

Wang, Z., Novikov, A., Zolna, K., Springenberg, J. T., Reed, S. E., Shahriari, B., Siegel, N., Merel, J., Gulcehre, C., Heess, N. M. O., and de Freitas, N. (2020). Critic Regularized Regression. *ArXiv*, abs/2006.15134.

Wu, Y., Tucker, G., and Nachum, O. (2019). Behavior Regularized Offline Reinforcement Learning. *ArXiv*, abs/1911.11361.

Wu, Y., Zhai, S., Srivastava, N., Susskind, J. M., Zhang, J., Salakhutdinov, R., and Goh, H. (2021). Uncertainty Weighted Actor-Critic for Offline Reinforcement Learning. In *International Conference on Machine Learning*.

Yarats, D., Brandfonbrener, D., Liu, H., Laskin, M., Abbeel, P., Lazaric, A., and Pinto, L. (2022). Don't Change the Algorithm, Change the Data: Exploratory Data for Offline Reinforcement Learning. *ArXiv*, abs/2201.13425.

Yarats, D., Zhang, A., Kostrikov, I., Amos, B., Pineau, J., and Fergus, R. (2019). Improving Sample Efficiency in Model-Free Reinforcement Learning from Images. In *AAAI Conference on Artificial Intelligence*.

Yu, T., Kumar, A., Rafailov, R., Rajeswaran, A., Levine, S., and Finn, C. (2021). COMBO: Conservative Offline Model-Based Policy Optimization. In *Neural Information Processing Systems*.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J. Y., Levine, S., Finn, C., and Ma, T. (2020). MOPO: Model-based Offline Policy Optimization. *ArXiv*, abs/2005.13239.

Zanette, A., Wainwright, M. J., and Brunskill, E. (2021). Provable Benefits of Actor-Critic Methods for Offline Reinforcement Learning. In *Neural Information Processing Systems*.

Zhang, A., Ballas, N., and Pineau, J. (2018). A Dissection of Overfitting and Generalization in Continuous Reinforcement Learning. *ArXiv*, abs/1806.07937.

Zhang, C., Kuppannagari, S. R., and Prasanna, V. K. (2021). BRAC+: Improved Behavior Regularized Actor Critic for Offline Reinforcement Learning. *ArXiv*, abs/2110.00894.