# ACAMDA: Improving Data Efficiency in Reinforcement Learning Through Guided Counterfactual Data Augmentation

**Yuewen Sun[1, 2], Erli Wang[3*], Biwei Huang[4], Chaochao Lu[5], Lu Feng[3], Changyin Sun[6], Kun Zhang[1, 2]**

[1]Mohamed bin Zayed University of Artificial Intelligence
[2]Carnegie Mellon University
[3]NEC Labs, China
[4]University of California San Diego
[5]Shanghai Artificial Intelligence Laboratory
[6]Anhui University
yuewen.sun@mbzuai.ac.ae, wang_erli@nec.cn, bih007@ucsd.edu, luchaochao@pjlab.org.cn, feng_lu@nec.cn, cysun@ahu.edu.cn, kunz1@cmu.edu

## Abstract

Data augmentation plays a crucial role in improving the data efficiency of reinforcement learning (RL). However, the generation of high-quality augmented data remains a significant challenge. To overcome this, we introduce ACAMDA (Adversarial Causal Modeling for Data Augmentation), a novel framework that integrates two causality-based tasks: causal structure recovery and counterfactual estimation. The unique aspect of ACAMDA lies in its ability to recover temporal causal relationships from limited non-expert datasets. The identification of the sequential cause-and-effect allows the creation of realistic yet unobserved scenarios. We utilize this characteristic to generate guided counterfactual datasets, which, in turn, substantially reduces the need for extensive data collection. By simulating various state-action pairs under hypothetical actions, ACAMDA enriches the training dataset for diverse and heterogeneous conditions. Our experimental evaluation shows that ACAMDA outperforms existing methods, particularly when applied to novel and unseen domains.

## Introduction

Reinforcement learning (RL), an important approach for sequential decision-making, aims to develop policies that guide the agent to learn optimal actions through trial-and-error interactions with the environment. Despite many recent advances, RL suffers from the challenge of data inefficiency. The performance of the policy is highly dependent on the quantity and quality of the training data, which limits its practical scaling and hinders generalization (Kamthe and Deisenroth 2018). To tackle this challenge, various data augmentation techniques have been developed to improve data diversity by manipulating the empirical dataset (Laskin et al. 2020). For instance, data augmentation in hyperbolic space can outperform benchmarks in speech, text, and visualization (Sawhney et al. 2021), while image operations can optimize control performance (Hendrycks et al. 2019).

*Corresponding author.


While traditional data augmentation techniques have successfully increased the diversity of offline data (Yang et al. 2022), they often overlook the impact of transition dynamics on policy optimization. In RL, the accuracy of policy estimation depends on the precise estimation of potential future states, and this estimation is strongly influenced by the actions taken and the observed current state. Unlike traditional augmentation techniques, which mainly focus on diversifying existing data without explicitly considering the transition dynamics of states, actions, and next states, counterfactual technique offers a valuable solution to address this limitation. They involve generating reachable data points by manipulating input variables to observe potential outcomes. Specifically, by manipulating the state and action variables to observe potential next states, it offers a promising approach to improve the quality of data tailored to policy objectives (Lu et al. 2020; Pitis, Creager, and Garg 2020). This ability helps reveal how different actions affect outcomes.

However, a challenge arises in how to accurately estimate the transition dynamics. On the one hand, existing methods often rely heavily on data-driven approaches that emphasize statistical patterns over causality. Some counterfactual augmentation approaches focus on correlations and overlook the causal relations when estimating the transition dynamics (Joshi and He 2021). This may lead to unreliable conclusions due to spurious correlations or chance. Furthermore, ignoring causality compromises the quality of augmented data, as it fails to capture the true causal mechanisms that accurately represent system dynamics, leading to inadequate and inexplicable guidance for counterfactual data augmentation, particularly in heterogeneous tasks. On the other hand, ignoring model bias, which is the discrepancy between the true values and estimates, may hinder counterfactual learning. Recent advances in counterfactual augmentation using generative models aim to predict the outcomes of hypothetical actions (Lu et al. 2020). However, these models often suffer from inaccurate predictions and unstable control performance due to imperfect representations, limited data, and intrinsic uncertainties. Addressing both causality and model

bias is essential for a balanced and effective approach to accurate transition dynamics estimation.

In this work, we present an end-to-end framework that incorporates causal relationship recovery and model estimation, enabling guided counterfactual data augmentation at the individual level. We novelly incorporate temporal causality for accountable generalization across heterogeneous domains. A challenge in counterfactual data generation is that ignoring domain constraints can inadvertently lead to cross-domain information contamination during training. To address this issue, we introduce a low-dimensional change factor, $\lambda$, into the model's dynamics and create a generalized dynamics prediction function suitable for heterogeneous domains, which aligns with related research (Lu et al. 2020; Huang et al. 2021; Yao, Chen, and Zhang 2022; Sun, Zhang, and Sun 2023). By integrating causality, we are able to generate realistic counterfactuals that accurately reflect the effects of interventions, thereby enhancing the generalization of decision-making processes. Additionally, during the model estimation phase, we introduce a discrepancy loss to effectively reduce bias. Finally, following Pearl's approach (Pearl et al. 2000), our counterfactual mechanism complements the original training dataset at the individual level, thus increasing the generalization and fairness of our generative models.

The contributions of the paper are summarized as follows.

- We investigate temporal causal relations to guide our counterfactual data augmentation. By integrating the change factor $\lambda$, we create a generalized dynamics prediction function that can effectively adapt to heterogeneous scenarios.
- We propose an end-to-end framework that includes model estimation and counterfactual learning, aiming to facilitate counterfactual data augmentation. A discrepancy loss is introduced to reduce model bias and stabilize the training process which distinct from (Lu et al. 2020).
- We demonstrate that under weak conditions, the counterfactual outcomes in heterogeneous cases are identifiable. This theoretical result provides a solid foundation for our approach and supports the use of counterfactual operations in diverse scenarios.

## Related Work

**Model estimation in RL** In model-based RL, the environment is represented by an estimated functional model, and the policy is optimized based on that model. One approach to this estimation problem is to consider the Markov property (i.e., the future and past are conditionally independent given the present) and unmeasured noise. This can be formulated as $\boldsymbol{s}_{t+1} = f(\boldsymbol{s}_t, \boldsymbol{a}_t, \epsilon_{t+1})$, where $f$ is the transition function, and the noise $\epsilon_{t+1}$ is considered under uncertainty. Pioneering work in this area has primarily focused on estimating accurate dynamic models in stationary cases. Gaussian process-based methods, such as PILCO (Deisenroth and Rasmussen 2011), learn probabilistic dynamic models and incorporate model uncertainty into long-term planning. While effective for learning from scratch with minimal data, these methods often struggle with high-dimensional and discontinuous dynamics (Roberto et al. 2016). While neural network-based models (Draeger, Engell, and Ranke 1995; Gal, McAllister, and Rasmussen 2016) have shown their superior scalability to sophisticated inputs, they are prone to overfitting on small datasets and minor errors will compound over the horizon. In addition, real-world data distributions tend to drift over time and exhibit heterogeneity with domain-varying means and variances. Existing solutions focus on either eliminating heterogeneity, i.e., periodic distillation (Kang et al. 2023) or change detection (Igl et al. 2020). Such methods can uncover some heterogeneous regions but overlook the intrinsic connection and interplay among different states and actions, which fails to resolve the contradiction between specification and generalization.

**Data Augmentation in RL** The most straightforward way of data augmentation is to prioritize synthetic examples for the minority class to address class imbalance. Recent advances focus on introducing interpretability into the data generation process. CoDA (Pitis, Creager, and Garg 2020) introduces a local causal model into MDP to decompose the whole task as locally independent mechanisms, and counterfactuals are generated from a local structure to speed up the training efficiency. MoCoDA (Pitis et al. 2022) extends CoDA to the generalization case based on a learned locally factored dynamic model, and generates counterfactual transitions for data augmentation. However, such methods do not explicitly engage in Pearl-style counterfactual reasoning and fail to consider domain-specific information embedded in exogenous noise variables. To handle the global model directly, CTRL (Lu et al. 2020) uses the structural causal model (Pearl 1980), which can describe the causal mechanism of the system, to model the state dynamics and perform counterfactual reasoning to solve the biased policy problem. However, it assumes predefined causal structure, which may not be realistic in some cases.

**Causal Discovery and Counterfactual Inference** Causal discovery aims to identify causal structure from observational data and provides possible guidelines for investigating interpretability. Standard techniques include constraint-based methods (e.g., PC (Spirtes et al. 2000)), score-based methods (e.g., GES (Chickering 2002)), and function-based methods (e.g., LiNGAM (Shimizu et al. 2006; Zhang and Hyvarinen 2009; Bühlmann, Peters, and Ernest 2014)). Given the causal structure, we can perform counterfactual inference and test what outcomes would have occurred had some preconditions been different (Pearl 1980). It is the ability to imagine alternative possibilities that are different from current observations (Morgan and Winship 2015). For instance, given a control situation *turn right at point A then arrive at point B*, we make a counterfactual decision *turn left at point A*, and the alternative outcome would become *arrive at point C*. From a state transition dynamics perspective, a counterfactual instance defines how the subsequent state would change if the agent took an alternative action (based on empirical data), and such inference is exactly based on the same condition for particular individuals rather than simulating a different action at the population level.

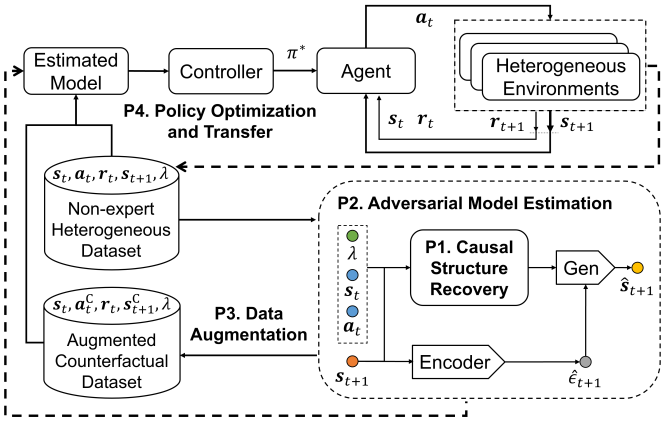



Figure 1: ACAMDA framework. P1. Causal structure is learned that encodes the relationships between variables and provides interpretation for the counterfactual analysis. P2. Adversarial model takes the causally related variables as inputs and estimates the transition dynamics with specified noise terms. P3. Counterfactual dataset is generated by hypothetical changing actions to realize data augmentation. P4. The policy is optimized based on the augmented dataset and transferred to the target domain.

## Preliminaries

**Problem Formulation** A Markov Decision Process (MDP) (Bellman 1957) provides the mathematical framework in the context of RL to model the environment in which an agent learns to make decisions based on states, actions, and rewards. In our work, we consider the case of *heterogeneous* environments, where the transition distribution varies across different environments, while the underlying causal mechanisms of the reward variables remain fixed. Our customized MDP formulation can be characterized as $(\mathcal{S}, \mathcal{A}, \mathbb{P}, R, \gamma, \lambda)$, where $\mathcal{S} \in \mathbb{R}^{d_s}$ and $\mathcal{A} \in \mathbb{R}^{d_a}$ represent the state and action spaces respectively, $\mathbb{P}$ denotes the transition probability, $R \in \mathbb{R}$ is the immediate reward, $\gamma \in [0, 1]$ is the discount factor, and $\lambda \in \mathbb{N}^+$ is the heterogeneous factor, i.e., the domain index. The policy $\boldsymbol{a} = \pi_\lambda(\boldsymbol{s})$ is a mapping from state $\boldsymbol{s}$ to action $\boldsymbol{a}$. The objective is to find an optimal policy that maximizes the expected discounted rewards $\mathbb{E}_{\pi_\lambda}\left[\sum_{t=0}^{T} \gamma^t R_t\right]$ under a specific horizon $T$.

Here, we address the challenge of characterizing heterogeneity in a structured manner. Real-world scenarios often exhibit distribution shifts (Li et al. 2019; Zhao et al. 2018), and understanding these shifts is crucial for effective model adaptation. The concept of sparse mechanism shift (Schölkopf et al. 2021) posits that distribution shifts arise from a limited set of changing causal conditions. Additionally, causally factorized distributions are known to change in a minimal and sparse manner (Ghassami et al. 2018). Based on these insights, we hypothesize that domain variations can be effectively captured by a changeable yet detectable factor $\lambda$. This factor remains constant within a given domain but varies across different domains.

**Data Generation Process** Here we observe a collection of trajectories $\mathcal{H}$ from a set of $\mathcal{E}$ heterogeneous environments that $\mathcal{H} = \{\mathcal{H}^1, \ldots, \mathcal{H}^{\mathcal{E}}\}$. Each trajectory $\mathcal{H}^e$ from the environment $e$ contains sequences of tuples $\{\boldsymbol{s}_t^e, \boldsymbol{a}_t^e, \boldsymbol{s}_{t+1}^e; \lambda^e\}_{t=1}^T$ following the transition distribution $\mathbb{P}^e(\boldsymbol{s}_{t+1}^e = s' | \boldsymbol{s}_t^e = s, \boldsymbol{a}_t^e = a)$, or simplified as $\mathbb{P}_a^e[s'|s]$. Since we consider the causal structures are fixed and the dynamic outcome in each domain is independent with each other, the distribution $\mathbb{P}_a[s'|s]$ can be written as

$$\mathbb{P}_a[s'|s] = \prod_{e \in \mathcal{E}} \mathbb{P}^e(\boldsymbol{s}_{t+1}^e | \mathrm{Pa}(\boldsymbol{s}_{t+1}^e)), \tag{1}$$

where $\mathrm{Pa}(\boldsymbol{s}_{t+1}^e) \subseteq \{\boldsymbol{s}_t^e, \boldsymbol{a}_t^e\}$ denotes the direct causes of $\boldsymbol{s}_{t+1}^e$. Here we assume causal sufficiency, that there are no latent confounders or hidden common causes that influence two or more observed variables in the system. Then the transition dynamics can be written as

$$\boldsymbol{s}_{t+1} = f(\mathrm{Pa}(\boldsymbol{s}_{t+1}), \epsilon_{t+1}; \lambda), \tag{2}$$

where $f$ represents smooth invertible functions, and the noise terms $\epsilon_{t+1}$ are independent of $\mathrm{Pa}(\boldsymbol{s}_{t+1})$. We aim to realize guided counterfactual data augmentation to improve data efficiency and achieve adaptive control that generalizes well to unseen environments.

## Proposed Method

The proposed framework is illustrated in Figure 1. The four steps are complementary and naturally integrated in an end-to-end manner for adaptive control. (P1) *Causal Structure Recovery* involves the recovery of temporal causal relations based on sparsity-promoting penalties to guide augmentation and improve generalization. (P2) *Adversarial Model Estimation* employs an improved Bidirectional Conditional Causal Generative Adversarial Network (BiCCGAN) to estimate dynamics $f_{\text{est}}$ by taking causally related elements as inputs. The bidirectional framework (Jaiswal et al. 2018) facilitates both forward and inverse mappings, allowing for individual-level counterfactual inference to augment the original dataset. The combination of adversarial training and an improved loss function effectively reduces bias. (P3) *Counterfactual-based Data Augmentation* generates alternative states $\boldsymbol{s}_{t+1}^C$ to augment training data and improve data efficiency. (P4) *Policy Optimization and Transfer* uses the augmented dataset to optimize the policy in a standard model-based fashion and transfer it to the new domains.

### Causal Structure Recovery

Given the offline collection of trajectories from heterogeneous environments, in this phase, we employ neural networks to recover the temporal causal relations and provide a clear and structured understanding of how states and actions at time $t$ influence states at time $t+1$. Following previous work (Tank et al. 2018), we construct individual neural networks for each element in $\boldsymbol{s}_{t+1}$, using inputs of $\boldsymbol{s}_t, \boldsymbol{a}_t$ along with $\lambda$. This construction is crucial for disentangling the influence of past inputs on future outputs, allowing us to understand how actions and states at time $t$ affect outcomes at time $t+1$. The incorporation of $\lambda$ allows for adaptation according to different environmental domains, which

enhances the adaptability of our model. We then implement component-wise MLP architectures to learn nonlinear transition dynamics. To these architectures, we apply sparsity-promoting penalties, such as convex group-lasso penalties, on the weights. These penalties encourage certain sets of weights to be zero, which in turn induces sparsity in the learned relationships. Sparse models are easier to interpret and are particularly useful for identifying causal relations in our temporal data. By simplifying the network in this manner, we can more easily identify and understand the key relationships between variables over time. Subsequently, we utilize causal graphical models, a powerful tool for representing probabilistic relationships in a visual format, to illustrate the dependency structure from time $t$ to $t+1$. Representing the joint distribution in this compact and factorized way not only simplifies the complexity of the system but also provides a clear picture of how different factors interact over time. This representation helps decompose complex systems and isolate the independent influencing factors, thereby improving policy adaptation in new scenarios.

## Adversarial Model Estimation

Here we introduce BiCCGAN, a novel adversarial framework that integrates causal structure to accurately estimate the transition model for counterfactual analysis. The design of the proposed BiCCGAN focuses on four key aspects: First, its bidirectional structure efficiently maps data to noise space, which is crucial for generating valid counterfactual instances. Second, by conditioning on states and actions at time $t$ to generate states at time $t+1$, it aligns with natural data generation processes, thereby improving reconstruction accuracy. Third, by using the GAN framework, BiCCGAN implicitly models complex data distributions, generating high-quality and diverse counterfactual instances. Lastly, the integration of causal structure not only improves interpretability but also facilitates adaptation to heterogeneous environments. Moreover, to address the limitations of traditional adversarial training, such as parameter oscillation and instability, we extend the work in (Lu et al. 2020) and provide an additional loss to stabilize the training process.

**Observations** BiCCGAN is based on several observations that, though relatively simple, are critical for good performance. (1) The causal structure supports the guided generation of counterfactual instances based on given observations. Thus, the inputs of BiCCGAN, filtered by the recovered causal structure, are variables causally related to the output. (2) An additional metric to measure the distance between the estimated and ground-truth distributions would stabilize the training process. Since the counterfactual instances are generated based on the estimated transition dynamics, the quality of the augmentation strongly depends on the estimation performance, while traditional GAN-based objectives often suffer from poor training stability due to the adversarial training process. Considering the smoothness of the transition dynamics, we propose an improved twin loss objective includes a Bias Factor Loss (BFL), which minimizes discrepancies, and an Adversarial Factor Loss (AFL), which promotes robustness. This improved objective offers a flexible way to balance training stability with augmentation robustness. Specifically, for any time $t$, a training minibatch is given as $(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1}, \lambda)$. The update processes in the bidirectional architecture are:

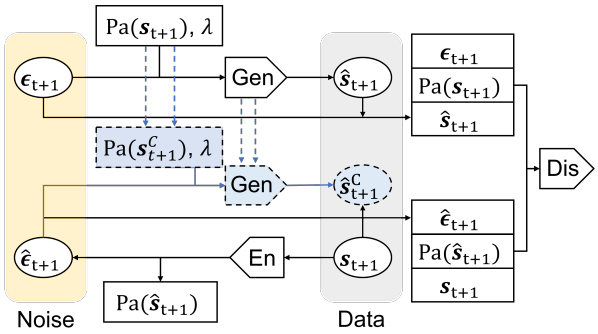



Figure 2: Adversarial model estimation and counterfactual-based data augmentation. Evidence $\boldsymbol{s}_{t+1}$ are used to estimate noise $\hat{\epsilon}_{t+1}$ and parent $\mathrm{Pa}(\hat{\boldsymbol{s}}_{t+1})$. Then the generative model is modified by alternating the parent $\mathrm{Pa}(\boldsymbol{s}^C_{t+1}) \leftarrow \mathrm{Pa}(\boldsymbol{s}_{t+1})$. Finally the counterfactual outcome is computed by the modified model as $\boldsymbol{s}^C_{t+1}$.

$$\hat{\epsilon}_{t+1} = E(\boldsymbol{s}_{t+1}), \tag{3}$$

$$\hat{\boldsymbol{s}}_{t+1} = G(\mathrm{Pa}(\boldsymbol{s}_{t+1}), \epsilon_{t+1}, \lambda), \tag{4}$$

where $G$ (generator) and $E$ (encoder) are the forward and inverse mappings, and $\hat{\boldsymbol{s}}_{t+1}$ and $\hat{\epsilon}_{t+1}$ are the forward and backward outputs, respectively. Since the heterogeneity is highly dependent on the heterogeneous factor, we also embed $\lambda$ in the data generation process.

**Objectives** We introduce AFL, a traditional GAN-based objective, to distinguish between true and generated samples. Furthermore, to reduce the model bias in the estimation process, we add a $L_2$ term between estimated states and true states $\mathcal{L}_{BFL} = ||\hat{\boldsymbol{s}}_{t+1} - \boldsymbol{s}_{t+1}||_2$ as an explicit mechanism that guides BiCCGAN to better approximate the transition dynamics. In general, the objective is

$$\min_{G,E} \max_{D} V(D,G,E) = \underbrace{\mathcal{L}_{DE} + \mathcal{L}_{DG}}_{AFL} + \beta \mathcal{L}_{BFL}, \tag{5}$$

where $\mathcal{L}_{DE} = \mathbb{E}\big[\log D(\boldsymbol{s}_{t+1}, \mathrm{Pa}(\hat{\boldsymbol{s}}_{t+1}), \hat{\epsilon}_{t+1})\big]$ and $\mathcal{L}_{DG} = \mathbb{E}\big[\log(1 - \alpha D(\hat{\boldsymbol{s}}_{t+1}, \mathrm{Pa}(\boldsymbol{s}_{t+1}), \epsilon_{t+1}))\big]$ are the components of AFL. The importance weight $\beta$ modulates the learning constraints applied to the model, balancing the training focus between reconstruction and robustness. $\beta$ can be selected as a constant or a parameter, and selecting a larger $\beta$ will encourage the model to learn a more accurate representation. $\alpha = 1 - \mathrm{norm}||\hat{\boldsymbol{s}}_{t+1} - \boldsymbol{s}_{t+1}||_1$ represents the normalized $L_1$ term used for regularization. Intuitively, we have the following different scenarios in qualitative terms:

- If the generator cannot fool the discriminator (i.e., $D(\cdot) \to 0$), $\alpha$ will have no effect.
- If the generator successfully fools the discriminator (i.e., $D(\cdot) \to 1$), the discrepancy between the true and generated samples is small. In this case, $\alpha$ approaches 1, and therefore, the value of $\alpha D(\cdot)$ is close to 1.

## Counterfactual-based Data Augmentation

Here, we combine the observational dataset $\mathcal{H}$ with the counterfactual dataset $\mathcal{H}^+$ to achieve data augmentation. Counterfactual inference refers to the ability to reason about the outcomes of alternative states or actions that could have been taken if everything else remained the same. It focuses on specific conditions for individual cases, rather than simulating different states or actions at the population level. The augmented outcomes can serve as additional trajectories, thereby improving the data efficiency and fairness of generative models. In this work, we explore the impact of certain actions to determine the next alternative state if the action had been different. We simulate hard interventions by fixing the value of action $\boldsymbol{a}_t$ to $\boldsymbol{a}_t^C$, with the alternative action $\boldsymbol{a}_t^C \in \mathcal{A}$ being randomly selected. Specifically, we follow Pearl's calculus of interventions which provides a standard procedure for counterfactual reasoning (Pearl 1980). The realization of this approach is depicted in Figure 2.

- **Abduction**: Determine the value of noise $\epsilon$ based on the factual evidence $\boldsymbol{s}_{t+1} = m, \mathrm{Pa}(\boldsymbol{s}_{t+1}) = n$.
- **Action**: Remove the structural equations for variables in $\mathrm{Pa}(\boldsymbol{s}_{t+1})$ and modify the model with alternate instance $\mathrm{Pa}(\boldsymbol{s}_{t+1}^C) = n'$.
- **Prediction**: Use the modified model and $\epsilon$ to compute the counterfactual outcome $\boldsymbol{s}_{t+1}^C = m'$.

In the RL paradigm, *abduction* explains the indeterminacy in the transition model $P(\boldsymbol{s}_{t+1}|\mathrm{Pa}(\boldsymbol{s}_{t+1}))$ by considering the current state $\boldsymbol{s}_t$ and action $\boldsymbol{a}_t$. *Action* alters the counterfactual action to $\boldsymbol{a}_t^C \leftarrow \boldsymbol{a}_t$, aligning it with a hypothetical antecedent. *Prediction* predicts the alternative next state $\boldsymbol{s}_{t+1}^C$ based on the revised understanding of the past and the alternative action $\boldsymbol{a}_t^C$.

## Policy Optimization and Transfer

After estimating the transition dynamic and augmenting the dataset, we optimize the policy network via RL, employing it as a warm start in the target domain. The control objective is to maximize the expected total reward, also known as the value function, by following best policy at each time step. Typically, each policy induces a value function $V$, and the best policy leads to an optimal value function $V^*$ satisfying $V^*(\boldsymbol{s};\lambda) = \max_{a\in\mathcal{A}}\left[r + \gamma \sum_{\boldsymbol{s}'\in\mathcal{S}} p(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a};\lambda)V^*(\boldsymbol{s}';\lambda)\right]$, where $r$ is the immediate reward, and $p(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a};\lambda)$ is the estimated probability of state transition. It's important to note that our policy optimization approach is general enough to be applied to any RL algorithm. For illustration purposes, we apply a customized Deep Q-Network (DQN) (Mnih et al. 2013) for policy optimization. The adapted DQN involves two main phases to maximize the optimal Q-value. First, the agent interacts with the learned model, performing actions and storing the experienced tuples in a replay buffer. Then, we randomly select a batch from the buffer and perform a gradient descent update on the policy. To summarize, we first optimize the policy $\pi^*$ on the learned model, then transfer $\pi^*$ directly to the target task as a warm-start to improve generalization efficiency.

## Theoretical Analysis

We further provide an identifiability theorem of the counterfactual operation.

**Theorem 1** *Assume a transition dynamic as given in Eq.* (2)*, where $f$ is an unknown, smooth, and strictly monotonic transition function. Let $\boldsymbol{a}^C$ denote the counterfactual action randomly sampled from the action space when an intervention is applied. Then the identifiability of the counterfactual outcome $\boldsymbol{s}_{t+1,\boldsymbol{a}_t^C}^C|\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1}, \lambda$ is ensured.*

**Proof sketch:** Given the monotonic property of the function $f^i$, we can derive the value of the noise term as $\hat{\epsilon}_{t+1}^i = {f^i_{\boldsymbol{s}_t,\boldsymbol{a}_t,\lambda}}^{-1}(\boldsymbol{s}_{t+1})$, where ${f^i_{\boldsymbol{s}_t,\boldsymbol{a}_t,\lambda}}^{-1}$ is the inverse of $f^i$ for the fixed tuple $(\boldsymbol{s}_t, \boldsymbol{a}_t, \lambda)$. Since the value of $\lambda$ remains constant within a single domain, and the counterfactual operation is performed independently within each domain, it can be considered a constant. Subsequently, the problem can be reduced to $\hat{\epsilon}_{t+1}^i = {h^i_{\boldsymbol{s}_t,\boldsymbol{a}_t}}^{-1}(\boldsymbol{s}_{t+1})$. Based on the monotonicity property of composite functions, $f$ implies the monotonicity of $h$. According to Theorem 1 in (Lu et al. 2020), the monotonicity condition imposed on $h$ with respect to $\epsilon_{t+1}$ ensures the recovery of the noise term, which in turn guarantees the identifiability of the counterfactual outcome.

# Experiments

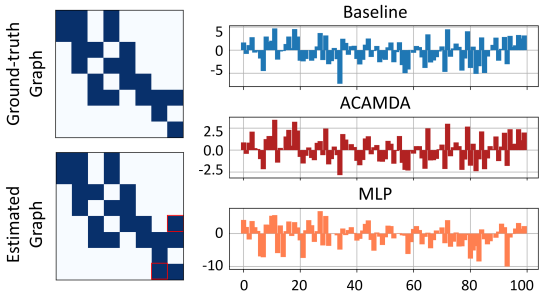
We evaluate ACAMDA on synthetic, control, and inventory tasks, in each case finding that ACAMDA can successfully recover temporally-causal structure, learn policies under variant noises and generalizes the new domains under limited data. Ablation study is further analyzed to demonstrate the advantages of causal investigation and improved loss. Below we outline our experimental design and results.

**Datasets** (1) Synthetic dataset is a variant of Vector AutoRegressive (VAR) model (Lütkepohl 2005) of order $p = 1$ with Gaussian noise. We varied the matrix strength across domains to create heterogeneity. (2) Control task is a variant of a classical control benchmark called CartPole (CP) (Brockman et al. 2016), where an unactuated joint attaches a pole to a cart, and the goal is to prevent it from falling. To mimic the realistic situation, we make two variants named CP-Noisy (with Gaussian noise) and CP-Windy (with multi-modal noise), and created both stationary dataset (SD) and heterogeneous dataset (HD) to test the feasibility. (3) Inventory task can be formulated as a sequential decision-making model with Laplacian noise. It aims to determine the optimal reorder for various commodities at a single store and we assume the strengths of interaction (between commodities) are varied across domains.

**Metrics and Baselines** We evaluated our approach and baselines using different metrics: (1) Structural Hamming Distance (SHD) (Norouzi, Fleet, and Salakhutdinov 2012) measures the distance between the estimated and ground-truth causal structure. (2) Mean Squared Error (MSE) (Das, Jiang, and Rao 2004) calculates the average of squared errors, quantifying model bias and evaluating the estimation performance of the learned model. (3) Pearson Correlation

| Methods | Metric | CP-Noisy: SD | CP-Noisy: HD | CP-Windy: SD | CP-Windy: HD | Inventory |
|---|---|---|---|---|---|---|
| **V-DQN** | Sample number | 81899 | 77544 | 77224 | 75612 | 83564 |
| | Success number | 182±16 | **200** | **200** | 104±47 | 100±49 |
| | AAR | **169.026±16.667** | **192.755±6.149** | **199.936±0.054** | 116.061±32.52 | 248.735±7.266 |
| **D-MLP** | Sample number | 2000 | 6000 | 2000 | 6000 | 8000 |
| | Success number | 40±21 | 0 | **200** | 0 | 0 |
| | AAR | 76.116±10.097 | 34.938±3.238 | 145.755±9.006 | 59.832±12.381 | 236.238±3.777 |
| **P-MLP** | Sample number | 2000 | 6000 | 2000 | 6000 | 8000 |
| | Success number | 103±36 | 76±40 | 108±45 | 3±2 | 150±42 |
| | AAR | 116.491±27.765 | 106.041±28.17 | 130.34±29.153 | 48.539±3.023 | 258.289±10.884 |
| **CoGAN** | Sample number | 2000 | 6000 | 2000 | 6000 | 8000 |
| | Success number | 62±40 | 10±7 | 50±42 | 66±33 | 50±42 |
| | AAR | 88.871±33.976 | 50.36±9.846 | 78.564±37.274 | 79.944±15.315 | 238.492±4.861 |
| **BiCoGAN** | Sample number | 2200 | 6600 | 2200 | 6600 | 8800 |
| | Success number | 23±15 | 12±10 | 0 | 0 | 50±42 |
| | AAR | 44.556±9.016 | 32.114±10.582 | 18.96±3.121 | 27.185±6.838 | 251.717±8.102 |
| **ACAMDA** | Sample number | 2200 | 6600 | 2200 | 6600 | 8800 |
| | Success number | **189±10** | **196±3** | **200** | **196±2** | **200** |
| | AAR | **157.658±13.502** | **185.592±7.692** | **161.935±12.396** | **132.135±8.084** | **301.362±8.536** |

Table 1: Control performance comparisons on CartPole-related tasks and the inventory task (5 seeds). ACAMDA outperforms all the baselines in the heterogeneous dataset including the bound baseline (V-DQN).




(a) Causal skeleton (b) State distribution

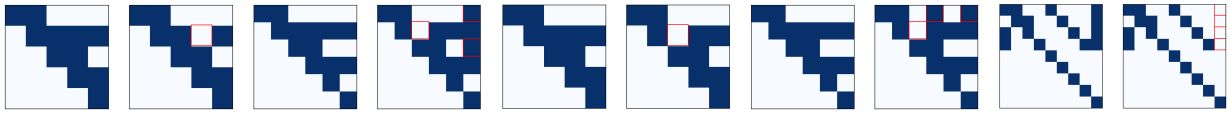
Figure 3: Synthetic dataset. (a) SHD=2 out of 64. (b) PCC comparisons between ACAMDA (0.97) and MLP (0.66).

Coefficient (PCC) (Cohen et al. 2009) measures the distribution discrepancy and quantifies the efficiency of data augmentation. (4) Success number counts the number of successes over trials to evaluate policy learning performance. (5) Average Accumulative Reward (AAR) specifies the return and quantifies the performance of the RL algorithm. We adopt the following baselines: V-DQN (Mnih et al. 2015), D-MLP (Lu et al. 2020), P-MLP (Lu et al. 2020), Conditional GAN (Mirza and Osindero 2014) and BiCoGAN (Jaiswal et al. 2018). V-DQN serves as the upper bound since the agent can interact with environments until it converges without data limits. For the ablation study, we compare ACAMDA with its variants that lack causal structure knowledge (named Add-BFL) and those without the improved loss (named Add-Causal).

**Causal Recovery Performance** To demonstrate the effectiveness of the counterfactual outcome, we performed experiments on a synthetic dataset, comparing the estimated distributions of $\boldsymbol{s}_{t+1}$ given $\boldsymbol{s}_t$ and $\boldsymbol{a}_t$. We used PCC to measure discrepancies between distributions (Figure 3(b)). The PCC between the ground truth and counterfactual state is $0.97$, significantly outperforming the MLP-based method's $0.66$.

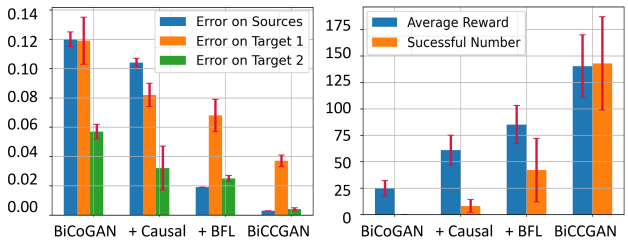
Furthermore, we present the estimated causal skeleton for all datasets in Figure 4. The columns and rows visually illustrate sparse causal connections between variables at time $t$ and $t+1$. Causal relations, represented as adjacency matrices (where $A_{ij} = 1$ indicates a connection from $j$ to $i$), are shown in blue for 1 and white for 0. The left and right subfigures correspond to the ground truth and estimated causal graph, respectively, with red rectangles highlighting the differences. In each graph, the first $S$ columns represent state elements, followed by $A$ columns for action elements. For heterogeneous datasets, an additional column in the causal graph denotes heterogeneity. For example, in Figure 4 (b), the first five columns represent four states (position $x$, velocity $\dot{x}$, angle $\theta$ and angle-velocity $\dot{\theta}$) and one action, respectively, while the last column indicates the auxiliary variable $\lambda$. ACAMDA is observed to successfully recover over 90% of the actual relations, demonstrating effectiveness even in datasets collected from heterogeneous environments. The aforementioned two pieces of evidence show that ACAMDA not only recovers ground-truth distributions effectively but also identifies the causal skeleton with high accuracy.

**Control Performance** Table 1 presents sample number, success number, and AAR for scenarios and methods, where the running policy is purely trained on the learned model and directly transfers to the new domain. In general, ACAMDA successfully learns policies under variant noises and generalizes the new domains under limited non-expert data. It has shown that the proposed method achieves comparable results to other leading model-based approaches when the testing domain is identical to the training domain. The im-



(a) CP-Noisy: SD (b) CP-Noisy: HD (c) CP-Windy: SD (d) CP-Windy: HD (e) Inventory

Figure 4: Ground truth graphs (left) and recovered graphs (right) on all datasets, indicating the successful recovery on the causal relations. (a) SHD=1 out of 25, (b) SHD=3 out of 36, (c) SHD=1 out of 25, (d) SHD=3 out of 36, (e) SHD=5 out of 81.




(a) Model bias (b) Performance

Figure 5: Ablation results. The causal knowledge and BFL help reduce model bias and improve control performance.

proving performance becomes clear and significant when the policy adapts to novel unseen scenarios, which verifies the generalization and data efficiency of the proposed method.

In detail, we have the following observations. (1) Compared with the model-free method (i.e., V-DQN) with unlimited interactions, ACAMDA improves the data efficiency to a large margin. It reduces the source samples on all datasets by 36 times (in CP-Noisy: SD), 11 times (in CP-Noisy: HD), 34 times (in CP-Windy: SD), 10 times (in CP-Windy: HD), and 8 times (in inventory). (2) ACAMDA achieved the best performance compared with the baselines under limited samples in stationary cases. For example, ACAMDA improved AAR by 35.3% (versus P-MLP) on CP-Noisy: SD and by 11.1% (versus D-MLP) on CP-Windy: SD, which are the best learning baselines in the stationary cases. (3) ACAMDA achieved almost the same average reward and success number with V-DQN and outperformed other baselines in nonstationary cases. For example, ACAMDA reaches the highest success number and reward, followed by V-DQN, CoGAN, D-MLP, P-MLP, and BiCoGAN in CP-Windy: HD. The reason is that ACAMDA can recover the heterogeneous mechanism in the causal structure and realize the model adaptation successfully to new domains, while other methods overlook the causal relations among domains and fail to model the new environment.

**Ablation Study** We demonstrate the advantages of BiCCGAN with an embedded factorized causal structure using CP-Noisy: HD as an example and conducting an ablation analysis over the source domain and two target domains. The main differences between the proposed BiCCGAN and the BiCoGAN lie in two aspects: (1) BiCCGAN enhances the loss objective by incorporating BFL, an explicit mechanism that helps to better encode of extrinsic factors; and (2) BiCCGAN integrates causal knowledge as a prior, and filters parents to be the only inputs. For the ablation study, we use four different network architectures: 1) BiCoGAN as baseline; 2) BiCCGAN without BFL (Add-Causal); 3) BiCCGAN without causal prior (Add-BFL); 4) BiCCGAN. All configurations within this study are kept the same unless stated otherwise, and the demonstration process is run three times under different random seeds to ensure generalization. The results are shown in Figure 5.

First, we analyze the importance of BFL and its influence on control performance, aiming to avoid causal knowledge bias. Figure 5 (a) shows the model bias under different cases on both source and target domains, indicating that: (1) BFL helps to train BiCCGAN better than BiCoGAN in both source and target domains. The introduction of BFL reduced the model bias on average by 84% (on the source domain), 43% (on the target domain 1), and 56% (on the target domain 2), while the average reward improved by 250%. (2) Simply adding BFL may underestimate the dynamics in the new domains. The reason is that BFL encourages model fitting but lacks expert knowledge to guide distribution shifts in heterogeneous environments. This observation motivates the introduction of causal structure into the system.

We further analyze the importance of causal knowledge and its influence on control performance. We see that incorporating causal knowledge into the data generation process can reduce the model bias in both the source and target domains by an average of 13% (on the source domain), 31% (on the target domain 1), and 44% (on the target domain 2). Although the model bias is large in the source, it can achieve comparable performance with Add-BFL architecture in the target domains, further proving the transferability. The twofold combination, which is BiCCGAN, achieved the best implementation in both model bias and control performance. In conclusion, causal knowledge helps a lot in improving the transferability and interpretability, and BFL loss is necessary to reduce the model bias.

## Conclusion

We propose a novel method to alleviate the data inefficiency problem in control tasks across heterogeneous domains. ACAMDA combines causal recovery with guided counterfactual data augmentation to realize sequential decision-making across heterogeneous environments in a data-driven manner, so that non-expert datasets can be used to disentangle the causal mechanism, leading to tremendous cost savings in collecting data from multiple sources. We hope that this new advancement will further advance the practicality of RL and allow more widespread applications of this robust approach to decision-making in the presence of uncertainty.

## Acknowledgments

Changyin Sun would like to acknowledge the support by the National Natural Science Foundation of China No. 62236002 and No. 61921004.

## References

Bellman, R. 1957. A Markovian decision process. *Journal of mathematics and mechanics*, 679–684.

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. OpenAI Gym. *arXiv: Learning*.

Bühlmann, P.; Peters, J.; and Ernest, J. 2014. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6): 2526–2556.

Chickering, D. M. 2002. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov): 507–554.

Cohen, I.; Huang, Y.; Chen, J.; Benesty, J.; Benesty, J.; Chen, J.; Huang, Y.; and Cohen, I. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, 1–4.

Das, K.; Jiang, J.; and Rao, J. 2004. Mean squared error of empirical predictor.

Deisenroth, M.; and Rasmussen, C. E. 2011. PILCO: A Model-Based and Data-Efficient Approach to Policy Search. In *Proceedings of the 28th International Conference on Machine Learning*, 465–472.

Draeger, A.; Engell, S.; and Ranke, H. 1995. Model predictive control using neural networks. *IEEE Control Systems Magazine*, 15(5): 61–66.

Gal, Y.; McAllister, R.; and Rasmussen, C. E. 2016. Improving PILCO with Bayesian neural network dynamics models. In *Data-Efficient Machine Learning workshop, ICML*, volume 4, 25.

Ghassami, A.; Kiyavash, N.; Huang, B.; and Zhang, K. 2018. Multi-domain causal structure learning in linear systems. *Advances in neural information processing systems*, 31.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2019. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*.

Huang, B.; Feng, F.; Lu, C.; Magliacane, S.; and Zhang, K. 2021. Adarl: What, where, and how to adapt in transfer reinforcement learning. *arXiv preprint arXiv:2107.02729*.

Igl, M.; Farquhar, G.; Luketina, J.; Böhmer, W.; and Whiteson, S. 2020. Transient non-stationarity and generalisation in deep reinforcement learning. *arXiv preprint arXiv:2006.05826*.

Jaiswal, A.; AbdAlmageed, W.; Wu, Y.; and Natarajan, P. 2018. Bidirectional Conditional Generative Adversarial Networks. In *Asian Conference on Computer Vision*, 216–232.

Joshi, N.; and He, H. 2021. An investigation of the (in) effectiveness of counterfactually augmented data. *arXiv preprint arXiv:2107.00753*.

Kamthe, S.; and Deisenroth, M. 2018. Data-efficient reinforcement learning with probabilistic model predictive control. In *International conference on artificial intelligence and statistics*, 1701–1710. PMLR.

Kang, S.; Kweon, W.; Lee, D.; Lian, J.; Xie, X.; and Yu, H. 2023. Distillation from Heterogeneous Models for Top-K Recommendation. In *Proceedings of the ACM Web Conference 2023*, 801–811.

Laskin, M.; Lee, K.; Stooke, A.; Pinto, L.; Abbeel, P.; and Srinivas, A. 2020. Reinforcement learning with augmented data. *Advances in neural information processing systems*, 33: 19884–19895.

Li, D.; Zhao, D.; Zhang, Q.; and Chen, Y. 2019. Reinforcement learning and deep learning based lateral control for autonomous driving [application notes]. *IEEE Computational Intelligence Magazine*, 14(2): 83–98.

Lu, C.; Huang, B.; Wang, K.; Hernández-Lobato, J. M.; Zhang, K.; and Schölkopf, B. 2020. Sample-efficient reinforcement learning via counterfactual-based data augmentation. *arXiv preprint arXiv:2012.09092*.

Lütkepohl, H. 2005. *New introduction to multiple time series analysis*. Springer Science & Business Media.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; Petersen, S.; Beattie, C.; Sadik, A.; Antonoglou, I.; King, H.; Kumaran, D.; Wierstra, D.; Legg, S.; and Hassabis, D. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533.

Morgan, S. L.; and Winship, C. 2015. *Counterfactuals and causal inference*. Cambridge University Press.

Norouzi, M.; Fleet, D. J.; and Salakhutdinov, R. R. 2012. Hamming distance metric learning. *Advances in neural information processing systems*, 25.

Pearl, J. 1980. Causality: models, reasoning, and inference.

Pearl, J.; et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2).

Pitis, S.; Creager, E.; and Garg, A. 2020. Counterfactual Data Augmentation using Locally Factored Dynamics. In *Advances in Neural Information Processing Systems*, volume 33, 3976–3990. Curran Associates, Inc.

Pitis, S.; Creager, E.; Mandlekar, A.; and Garg, A. 2022. MoCoDA: Model-based Counterfactual Data Augmentation. *arXiv preprint arXiv:2210.11287*.

Roberto, C.; Jan, P.; Carl, E. R.; and Marc, P. D. 2016. Manifold Gaussian processes for regression. In *In International Joint Conference on Neural Networks (IJCNN)*, 3338–3345.

Sawhney, R.; Thakkar, M.; Agarwal, S.; Jin, D.; Yang, D.; and Flek, L. 2021. HYPMIX: Hyperbolic Interpolative Data Augmentation. In *Proceedings of the 2021 Conference on*

*Empirical Methods in Natural Language Processing*, 9858–9868.

Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Toward Causal Representation Learning. *Proceedings of the IEEE*, 109(5): 612–634.

Shimizu, S.; Hoyer, P. O.; Hyvärinen, A.; Kerminen, A.; and Jordan, M. 2006. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10).

Spirtes, P.; Glymour, C. N.; Scheines, R.; and Heckerman, D. 2000. *Causation, prediction, and search*. MIT press.

Sun, Y.; Zhang, K.; and Sun, C. 2023. Model-Based Transfer Reinforcement Learning Based on Graphical Model Representations. *IEEE Transactions on Neural Networks and Learning Systems*, 34(2): 1035–1048.

Tank, A.; Covert, I.; Foti, N.; Shojaie, A.; and Fox, E. 2018. Neural Granger Causality. *arXiv preprint arXiv:1802.05842*.

Yang, S.; Xiao, W.; Zhang, M.; Guo, S.; Zhao, J.; and Shen, F. 2022. Image data augmentation for deep learning: A survey. *arXiv preprint arXiv:2204.08610*.

Yao, W.; Chen, G.; and Zhang, K. 2022. Temporally Disentangled Representation Learning. *arXiv preprint arXiv:2210.13647*.

Zhang, K.; and Hyvarinen, A. 2009. On the identifiability of the post-nonlinear causal model. In *Uncertainty in Artificial Intelligence*.

Zhao, X.; Zhang, L.; Ding, Z.; Xia, L.; Tang, J.; and Yin, D. 2018. Recommendations with negative feedback via pairwise deep reinforcement learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1040–1048.