ERCI: An Explainable Experience Replay Approach with Causal Inference for Deep Reinforcement Learning

Jingwen Wang, Dehui Du*, Lili Tian, Yikang Chen, Yida Li, YiYang Li

Shanghai Key Laboratory of Trustworthy Computing, East China Normal University, 200062, China
dhdu@sei.ecnu.edu.cn

Abstract

Deep reinforcement learning (DRL) has gained significant attention in autonomous systems, yet its black-box nature and lack of explainability hinder user trust in safety-critical domains such as autonomous driving. Existing experience replay approaches enhance sample efficiency but often fail to capture the internal causality of training data, leading to a convoluted training process that is difficult for humans to explain. In this work, we introduce Experience Replay with Causal Inference (ERCI), an explainable approach that integrates time series representation and causal inference to offer human-aligned explanations for DRL. Specifically, ERCI 1) introduces a novel multivariate time series representation to extract explainable Time Series Causal Factors (TSCF) from experimental data and 2) leverages internal causality in TSCFs with causal inference as a crucial standard for experience replay in DRL training. We evaluate ERCI using multiple baseline algorithms across diverse environments. Results show that ERCI provides human-aligned explanations and further improves sample efficiency through enhanced explainability. Notably, ERCI outperforms other state-of-the-art approaches by 15% in average performance, highlighting its effectiveness and generalizability.

Introduction

Reinforcement Learning (RL) enables agents to interact with the environment iteratively, acquiring an optimal policy to maximize the cumulative expected reward. However, traditional RL techniques face limitations in handling high-dimensional state spaces, generalizing models, and learning complex tasks. To address these issues, Deep Reinforcement Learning (DRL) (Mnih et al. 2013) introduces deep neural networks, which can more effectively manage complex state representations and action spaces. Despite these advancements, DRL models are often considered black boxes (Heuillet, Couthouis, and Díaz-Rodríguez 2021), requiring extensive interactions to obtain a limited number of effective samples. Experience replay approaches (Lin 1992) can improve sample efficiency but often fail to capture the internal causality within the experiences, leading to an inscrutable training and decision-making process. This inscrutability makes it difficult for humans to understand and explain DRL model decisions in practical applications, thereby limiting their application scenarios. Particularly in domains such as autonomous driving, the complex and dynamic road environment and potential accident risks demand high explainability of vehicle behavior, which is essential for gaining user trust in these safety-critical systems.

Therefore, eXplainable Reinforcement Learning (XRL) has become a research focus, aiming to provide explanations for model predictions to ensure the reliability of model decisions. Due to the complexity and opacity of neural network models in DRL, many existing studies focus on post-hoc explainability for different aspects of XRL (Arrieta et al. 2020; Heuillet, Couthouis, and Díaz-Rodríguez 2021). These include representation learning approaches based on states, actions, and policies (Raffin et al. 2019), simultaneous learning of policies and explanations (Juozapaitis et al. 2019), multi-objective learning approaches (Beyret, Shafti, and Faisal 2019; Cideron et al. 2019) based on Hindsight Experience Replay (Andrychowicz et al. 2017; Beyret, Shafti, and Faisal 2019; Cideron et al. 2019), global reward allocation explanation approaches based on Shapley values (Wang et al. 2020), and explanation approaches based on saliency maps in image data (Selvaraju et al. 2017). Causal inference (Pearl and Mackenzie 2018) is an emerging research direction in the field of model explanation. Unlike traditional statistics, causal inference is closely related to human cognitive psychology and analyzes events through logical chains, making it well-suited for explaining decisions (Kuang et al. 2020; Young et al. 2016). For example, Madumal combined decision trees with causal models to generate counterfactual explanations (Madumal et al. 2020), and other researchers have reshaped reward functions to add exploratory rewards, allowing agents to intervene in behaviors to optimize causal models (Volodin, Wichers, and Nixon 2020). However, there is currently a lack of research on explaining the experience replay process through a causal inference lens, which presents a significant research gap. Addressing this gap can enhance the explainability of the model from the perspective of data sampling, which is fundamental to DRL training.

In addition, the low sample efficiency in DRL remains a pressing issue. In complex application domains like robotic control and autonomous driving, DRL requires substantial interaction experience to update models. This requirement is often hard to fulfill due to sparse rewards and the imbal-

*Corresponding author.

anced distribution of observations in large state spaces (Zeng et al. 2023). Existing approaches such as reward shaping (Ng, Harada, and Russell 1999), imitation learning (Christiano et al. 2017), transfer learning (Vecerik et al. 2017), and meta-learning (Wang et al. 2016) enhance DRL's sample efficiency in various ways. The experience replay approach focused on in this paper thoroughly mines historical samples to improve sample efficiency by reusing existing experience data. However, current experience replay approaches often rely on traditional dimensional correlations and do not fully leverage the internal temporal and causal information within historical experiences, which hampers the potential for further improvement in sample efficiency.

To solve these issues, this paper proposes an approach that integrates time series representation and causal inference to explain the model-free DRL training process in a way that aligns with human values, which also leverages internal causality within DRL experience data to improve sample efficiency. We hypothesize that repetitive "patterns" in the DRL training process exhibit temporal and causal correlations. We segment the interaction data into meaningful patterns, treated as temporal factors, and use causal inference to establish relationships among them. To integrate the relationships into the DRL training process, we introduce a new experience replay architecture that considers the causality between policy objectives and temporal factors during experience replay. We explored experiments in several common DRL environments, demonstrating the explainability, overall performance, and scalability of our approach. To the best of our knowledge, this is the first work attempting to enhance the explainability of experience replay with causal inference. Our main contributions are as follows:

- We introduce an innovative multivariate time series representation approach, which uses two-stage analysis to extract explainable Time Series Causal Factors (TSCF) from time series data. It captures the complex internal causality within the multivariate time series, thereby simplifying the information in long sequences and addressing issues of non-uniform time steps.
- We propose Experience Replay with Causal Inference (ERCI). It's a novel approach utilizing causal inference techniques to analyze the causality within historical experience data and uses causal strength as a key criterion for sampling in experience replay. This significantly improves the data utilization in traditional experience replay, providing a more human-aligned method for agents to learn strategies and adapt to environments.
- We evaluated the explainability and effectiveness of ERCI through several experiments. Using the simulator CARLA, we analyzed overtaking scenarios with causal models to provide human-aligned explanations. We also compared our approach with state-of-the-art DRL models in Highway-Env and Gym environments, particularly showing that ERCI boosts the average score of the baseline approach by 20.70%. Additionally, we extended our approach to other experience replay frameworks to demonstrate its scalability.

Related Work

Multivariate Time Series Representation

The multivariate time series data encountered during DRL training is highly complex. To enhance the efficiency of analysis, it is necessary to perform dimensionality reduction and re-represent the data. Traditional approaches for representing time series data include piece-wise approximation, identification of key points, and symbolic representation (Wilson 2017). However, these approaches often modify the original data structure, obscuring crucial latent patterns and lacking explainability. Therefore, we aim to uncover hidden information through data mining techniques. In the field of pattern recognition, Matrix Profile-based approaches have shown excellent results (Yeh, Kavantzas, and Keogh 2017). In the clustering field, David Hallac introduced the Toeplitz Inverse Covariance-based Clustering (TICC) algorithm (Hallac et al. 2017), which employs Markov Random Fields (MRF) to define each cluster within multi-dimensional time series, ensuring that dimensionality reduction does not compromise data explainability.

Causal Discovery

Identifying causality is a fundamental issue in time series data mining. However, the high dimensionality and lengthy sequences in multivariate time series make this task particularly challenging. Causal discovery algorithms can uncover potential causality beyond mere correlations. The Granger causality test, introduced by Clive Granger (Granger 1969), is a classic technique for determining relationships between time series dimensions. Other notable algorithms include the Peter-Clark algorithm (Spirtes and Glymour 1991), and the Fast Causal Inference algorithm (Spirtes, Glymour, and Scheines 2000). Building on prior research, the Greedy Fast Causal Inference (GFCI) algorithm (Ogarrio, Spirtes, and Ramsey 2016) addresses the influence of unmeasured confounders. Independent Component Analysis (Hyvärinen and Oja 2000) has also been extended to identify causal relationships in multivariate time series. However, most existing research focuses on causal relationships between features, overlooking the internal causality within the time series.

Experience Replay

The formal study of experience replay was initiated by Lin et al. (1992) and has since been widely applied in various DRL models (Mnih et al. 2013, 2015; Silver et al. 2014; Mnih et al. 2016). One issue with uniform random sampling in classic experience replay is that it masks differences in correlations between experiences, leading to suboptimal updates. Some studies have improved the experience replay algorithm to address this deficiency. Schaul et al. (2016) proposed a Prioritized Experience Replay (PER) algorithm based on Temporal Difference (TD) error. Andrychowicz et al. (2017) introduced the hindsight experience replay algorithm, effectively sampling experiences from sparse and binary rewards. Zhang et al. (2017) observed a significant impact of large replay buffer size on results and proposed combined experience replay to mitigate the effects. Fedus et al. (2020) demonstrated that increasing replay capacity

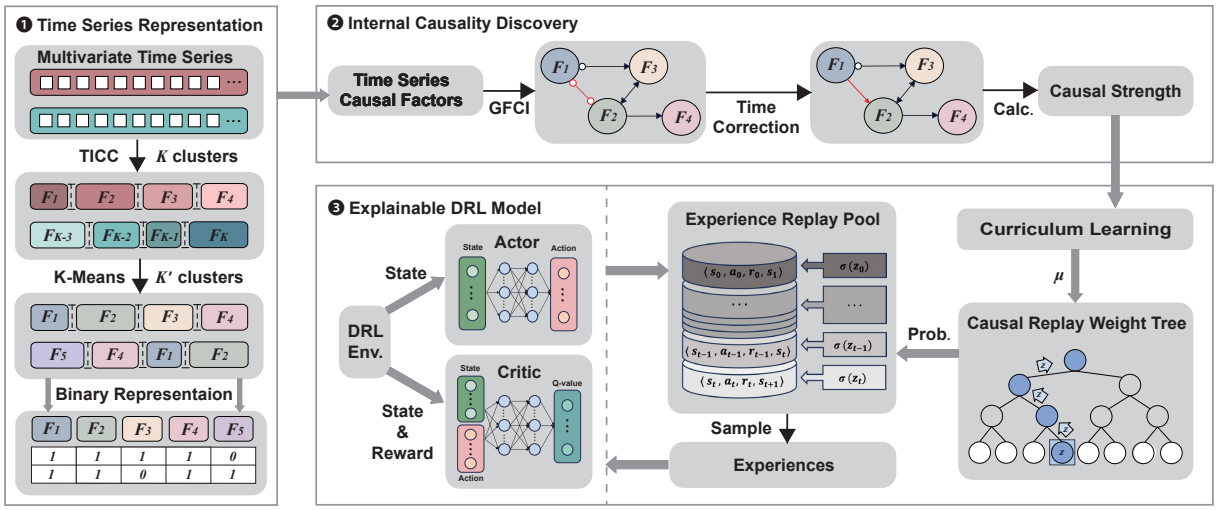


Figure 1: Architecture and workflow of ERCI.

and reducing the age of the oldest policy can enhance agent performance. Sullivan et al. (2023) investigated the impact of reducing experience replay capacity in Deep Q-learning agents and visualized experience replay decisions to improve interpretability and resource efficiency in DRL applications. However, these prior works do not leverage causal information within experiences and lack human-aligned explanations. Our approach differs by explaining meaningful time subsequences and enhancing the explainability of experience replay with a causal inference perspective, applicable to most DRL models using experience replay techniques.

Methodology

Overview

We used the Actor-Critic (AC) framework (Konda and Tsitsiklis 1999) as the foundation and optimized the experience replay process within the standard DRL model. Figure 1 provides an overview of ERCI, which comprises the following modules: ❶ Time Series Representation, ❷ Internal Causality Discovery, and ❸ Explainable DRL Model.

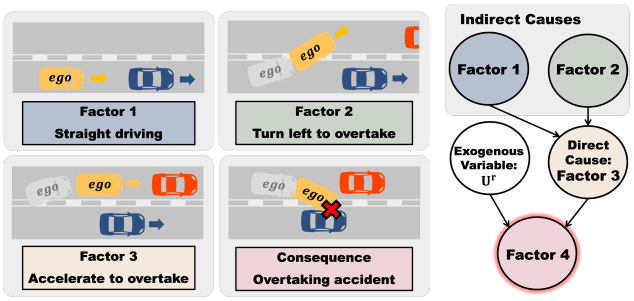
Module ❶ represents the interaction data between the DRL environment and the agent as a multivariate time series. Using two-stage clustering, it extracts internal TSCFs and re-dimensionally represents the data, preserving explainability and highlighting significant patterns and trends. Next, module ❷ employs the do-operator and GFCI algorithm to uncover causal relationships between the TSCFs that align with human understanding and calculate causal strengths for sampling during the experience replay process. Module ❸ describes the DRL training process. The left part involves the AC architecture, which combines policy learning and value learning, interacting with the environment and storing experiences. The critic evaluates the actor's decisions and optimizes the model parameters. The right part maps the causal strengths from module ❷ to the experience replay pool, optimizes experience weight updates through a curriculum learning strategy, and uses a causal replay weight tree to enhance sample efficiency, integrating the internal causality of time series into the experience update process. The DRL training concludes when it converges or reaches a specified number of iterations.

Multivariate Time Series Representation

The raw time series typically encompasses extensive information. The goal of time series representation is to simplify this complex data into an explainable format. Current representation algorithms typically require equal-length sequences or use approaches like truncation, interpolation, or padding. However, these approaches may fail to capture the nonlinear structures and dynamic patterns in the data. Our work introduces a novel multivariate time series representation approach. We define time series with action time series data abstracted from historical actions in DRL, which can defined as $T_A = [a_1, a_2, \ldots, a_n] \in \mathbb{R}^{n \times d}$, where n denotes the length of T_A, $a_i \in \mathbb{R}^d$ is the i-th multidimensional action vector of T_A, and d represents the feature dimension. The features of time series are $\{x_1, x_2, \ldots, x_d\}$. We conceptualize time series as a set of discrete, non-overlapping, explainable patterns, thereby achieving dimensionality reduction in storage. These patterns, referred to as TSCFs, collectively form the set $\mathcal{F}$ as defined in Equation 1. Each F_i represents the i-th segment of the time series.

$$\mathcal{F} = \{F_i \subseteq \mathcal{T}_A \mid \mathcal{T}_A = \bigcup_i F_i,\ F_i \cap F_j = \varnothing,\ i \neq j\}. \quad (1)$$

In summary, TSCFs are defined as a set of subsequences with internal semantics and causal correlations. This segmentation approach breaks down the entire dataset into a series of subsequences, each possessing unique features and semantically distinct from others. This approach not only simplifies the complexity of the original time series but also provides a more explainable foundation for subsequent causal inference. To demonstrate the explainability of TSCFs, we designed a safety-critical overtaking scenario in the autonomous driving system, as illustrated in Figure 2.



(a) (b)

Figure 2: Analysis of the overtaking scenario. As shown in (a), the entire scenario can be divided into four explainable TSCFs based on the temporal order of actions, whose relationships can be analyzed as a causal graph in (b).

Suppose we want to trace the causes and avoid similar collisions from a human perspective in future scenarios. In that case, we should consider slowing down to avoid the collision in Factor 3 or trace back to the more fundamental cause, abandoning the overtaking intention in Factor 2. In summary, meaningful subsequences can make the research results more explainable. Therefore, we should regard TSCFs as fundamental units for further utilization and analysis.

To construct the aforementioned TSCFs, we propose a two-stage representation approach. In the first stage, we use the TICC algorithm to extract the internal causality within the time series data, ensuring that the TSCFs have explainable features. The TICC algorithm aims to cluster T_A into K clusters, with each cluster represented by a Precision Matrix (PM) $\Theta_i \in \mathbb{R}^{dw_i \times dw_i}$, where d is the feature dimension of T_A and w_i is the sequence length of the i-th T_A. The PM, derived from the inverse of the covariance matrix, indicates the conditional independence structure among different clusters. Formally, it is defined in Equation 2:

$$\text{COV}(X_i, X_j) = 0 \Leftrightarrow X_i \perp X_j \mid X_{V \setminus \{i,j\}}, \tag{2}$$

where X_i represents the i-th cluster in the PM, and V is the set of all clusters. Based on the expression of MRF, this problem can be transformed into finding a convergent $L(\Theta, P)$ as the clustering result Θ by applying sparsity and temporal continuity constraints to the PM and using maximum likelihood estimation. The final optimization objective can be defined as in Equation 3:

$$L(\Theta, P) = \text{argmin} \sum_{i=1}^{K} \Bigg[\|\lambda \circ \Theta_i\|_1 + \sum_{X_t \in P_i} \Big(-\ell\ell(X_t, \Theta_i) + \beta \mathbb{1}\{X_{t-1} \notin P_i\} \Big) \Bigg], \tag{3}$$

where λ and β are control parameters.

In the second stage, we apply the K-Means algorithm and Dynamic Time Warping (DTW) measure (Berndt and Clifford 1994) to merge the initial clusters with high similarity between individual time series samples. We group all subsequences from the previous stage, randomly select K' subsequences as initial cluster centers, and calculate the DTW distance between each remaining subsequence and each cluster center, assigning them to the nearest cluster center until the algorithm converges or reaches the maximum iterations.

The final result is regarded as TSCFs and mapped onto binary variables. We use the binary variable $\alpha_{i,j} \in \{0, 1\}$ to represent whether TSCF F_i is present in the time series T_A, where $1 \leq i \leq K'$. According to Equation 1, the combination of all TSCFs forms a complete binary time series.

Internal Causal Inference within Time Series

Our approach focuses on internal temporal correlations and causality within time series to clarify explainable common patterns, setting it apart from traditional techniques. We designate TSCF as the treatment variable to observe its impact on the outcome variable, the reward obtained from each training iteration in DRL. By employing the GFCI algorithm, we construct a Partial Ancestral Graph (PAG) (Zhang 2008) to aid in the discovery of causal relationships and use the Bayesian Information Criterion to assess the fit and complexity of the causal model within the PAG. According to the PAG definition, there are four types of causal relationships, with only the type "$\rightarrow$" indicating deterministic causality, while the other three types remain uncertain. Our internal causal inference approach performs correction and optimization on each type, as detailed in Figure 3.

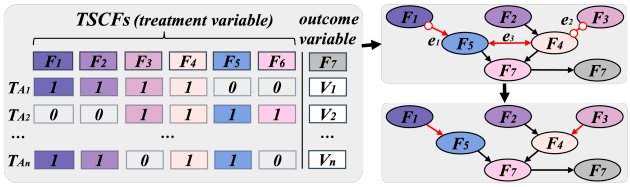


Figure 3: Illustration of the optimization process in the PAG model. The left panel shows TSCFs F_1 to F_6 and the outcome variable F_7 in binary multivariate time series. The right panel shows uncertain causal relationships with GFCI and highlights the optimization results, emphasizing uncertain direct causality e_1, temporal uncertainty e_2, and bidirectional uncertainty e_3.

- **Uncertain Direct Causality (A $\circ\!\!\rightarrow$ B)**: For e_1, where F_1 and F_5 have an uncertain direct causal relationship, we assume a causal relationship exists to capture all potential factors that might influence reward changes, even if it includes some non-causal associations.
- **Temporal Uncertainty (A $\circ\!\!-\!\!\circ$ B)**: For e_2 between F_3 and F_4, we calculate the probabilities of F_3 occurring before F_4 and vice versa, following the principle of Granger causality, which asserts that future events cannot causally influence the past. By evaluating these probabilities. The variable with the higher likelihood of occurring first is designated as the causal variable.
- **Bidirectional Uncertainty (A $\leftrightarrow$ B)**: For e_3 between F_5 and F_4, we do not consider bidirectional relationships

but focus on identifying unidirectional causal relationships that are more likely to affect the study outcomes.

For the optimized PAG graph, we select all paths ending at the outcome variable F_7 and intervene on all variables along those paths to calculate their causal strength with F_7. If there are indirect causal relationships between the treatment variables and F_7, we traverse all existing edges between them, aggregating their causal strengths to determine the overall causal strength. This approach aims to address the limitations of binary variable representation in capturing temporal sequences, thereby providing a more detailed and human-aligned causal inference in multivariate time series.

Experience Replay Model with Causal Inference

From a human perspective, as shown in Figure 2 for the overtaking scenario, altering the decision-making process of Factor 3 could prevent accidents. Additionally, modifying Factor 2 to account for a failed left turn would preclude Factor 3, thus avoiding the accident. Similarly, from the perspective of a DRL agent, placing greater emphasis on Factor 2 in this scenario can significantly benefit the DRL training process, minimizing reward penalties due to accidents.

In DRL, experience data is defined as the sequence (s_t, a_t, r_t, s_{t+1}), where s_t is the state, a_t is the action, r_t is the reward, and s_{t+1} is the next state. We extract actions and rewards temporally from the experience replay buffer to construct an action time series for each training episode. Using the proposed time series representation approach, we can extract the TSCF set, which serves as the treatment variables in the internal causal inference process within the time series. To adapt to DRL, the outcome variable in the PAG is adjusted to cumulative reward returns.

First, consider the causal strength of a single path from treatment node A to outcome node R. Let τ_i represent the causal strength of the i-th edge in the path. The causal strength of the path τ_{path} is given by Equation 4:

$$\tau_{\text{path}} = \prod_i \tau_i. \tag{4}$$

The total causal strength $\tau_{A \to R}$ from A to R is the sum of the causal strengths of all possible paths $\mathcal{P}$, as defined in Equation 5:

$$\tau_{A \to R} = \sum_{\text{path} \in \mathcal{P}} \tau_{\text{path}}. \tag{5}$$

Next, to leverage this causal relationship in DRL, we enhance the experience replay sampling algorithm by using $\tau_{A \to R}$ as a measure of experience sampling priority. Each experience in the replay buffer is assigned a corresponding causal strength value through the aforementioned calculation. To ensure all experiences have an opportunity to be sampled during training and avoid prolonged sampling from local regions, we apply the softmax function to transform the causal strengths into "causal weights" w_i, where $w_i \in (0, 1)$. During model updates, experiences are prioritized based on these causal weights, enhancing the efficiency of weighted experience sampling.

Finally, to adapt to the increasingly complex training process in the DRL environment, we use a curriculum learning strategy to update the experience weights. Initially, higher causal weights are assigned and gradually reduced, enhancing initial training efficiency. Over time, this approach shifts the model's focus to new samples. This process is governed by a hyperparameter μ, calculated as Equation 6:

$$\mu = \frac{1}{\epsilon_m} \eta \sqrt{\epsilon_m^2 - \epsilon_c^2}, \tag{6}$$

where ϵ_m denotes the maximum number of training epochs, ϵ_c represents the current number of training epochs, and η stands for the proportional constant of the curriculum learning control parameter. In addition, We employ a sum-tree structure, referred to as the causal replay weight tree, for storing and retrieving causal weights, enabling efficient weighted random sampling. The time complexity for locating weights at the leaf nodes is $O(\log n)$.

Experiments

Experimental Setup

To evaluate the effectiveness of our DRL-based ERCI approach, we conducted experiments in diverse DRL environments and compared our approach with other state-of-the-art approaches. The experiments are designed to assess ERCI's performance and its impact on DRL explainability and sample efficiency. We performed them on a system with NVIDIA GeForce RTX 3090 and Python 3.8. Specifically, we aim to address the following research questions:

- **RQ1 (Explainability)**: How does ERCI improve the explainability of the DRL training process in a way that aligns with human understanding, allowing humans to comprehend and interpret its operations?
- **RQ2 (Overall Performance)**: How does the DRL training performance of ERCI, which incorporates explainability, compare to state-of-the-art approaches across diverse environments?
- **RQ3 (Scalability)**: Does ERCI scale well and maintain consistent performance across various DRL models?

Dataset Our DRL environment is constructed with the Highway-Env (Leurent 2018) and Gym (Brockman et al. 2016) libraries. In the Highway-Env, we experiment with vehicle-based traffic flow and complex highway scenarios, including Highway, Intersection, and Racetrack. The autonomous vehicle is tasked with efficiently completing the specified tasks while avoiding collisions with other vehicles. The reward is defined as a combination of speed and collision avoidance, as detailed in Equation 7:

$$R(s, a) = a \cdot \frac{v - v_{\min}}{v_{\max} - v_{\min}} - b \cdot \text{collision}, \tag{7}$$

where the collision penalty is set to -1, and a and b representing scaling coefficients. In the Gym environment, we performed experiments in the Reacher and Humanoid settings, both of which rely on the Mujoco engine and feature continuous action spaces.

Baselines Since ERCI is primarily based on the AC architecture, we employed two state-of-the-art baselines for training across various DRL environments: the Deep Deterministic Policy Gradient algorithm (DDPG) (Lillicrap et al. 2015) and the Twin Delayed Deep Deterministic Policy Gradient algorithm (TD3) (Fujimoto, Hoof, and Meger 2018). Both DDPG and TD3 are exemplary models in the field of DRL, especially suitable for problems involving continuous action spaces. Therefore, in our experiments, the action type was set to continuous. The network models used in the experiments consisted of 3 to 6 fully connected layers, with a training batch size of 256.

To enhance training efficiency and avoid recalculating TSCF during each iteration, we designed a temporary experience buffer with a capacity of 100,000. This buffer synchronizes and updates with the main experience replay buffer, which handles regular experience storage. The smaller temporary buffer manages the activation of the internal causality discovery module by storing only indices. When the buffer is full, the module processes the data, updates the main pool with causal weights, and resets the temporary buffer to acquire new entries.

Metrics Based on prior research, we adopted widely used DRL training performance metrics, including Average Score (AS) and Best Score (BS), which respectively reflect the real-time performance within the same training cycle and the highest achievable score within the given training session. Additionally, we introduced supplementary metrics, namely Step of Average Score (SAS) and Average Cumulative Score (ACS). SAS represents the minimum number of steps required for the algorithm to reach the average score, highlighting learning efficiency and convergence speed. ACS, on the other hand, calculates the cumulative average score from the start to each time point, measuring the overall performance stability throughout the training process.

Discussion

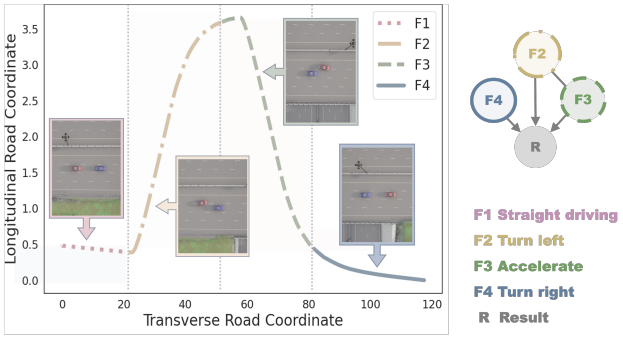
To more intuitively demonstrate the effectiveness of ERCI, we concentrated our experiments on the initial 1000 training episodes. We conducted multiple trials to obtain average values, and the results are detailed in Table 1. Our experiments showed that the ERCI approach effectively addresses the challenges faced by DRL with causal inference, improving both explainability and sample efficiency.

RQ1: Explainability To illustrate the explainability of our approach, we used the CARLA simulator to simulate the overtaking scenario in Figure 2. We prepared a dataset containing 300 instances. Our reward is defined as Equation 8:

$$2\|\mathbf{L_c} - \mathbf{L_d}\|_2 + \text{collision}, \tag{8}$$

where $\|\mathbf{L_c} - \mathbf{L_d}\|$ represents the L_2 norm between the current position of T_S and the task destination, with the collision penalty set to -1.

We analyzed the data using the temporal and causal modules within the ERCI framework. Figure 4a displays a representative sample trajectory and its TSCF representation. The target vehicle moves in the positive x-axis direction, while



(a) (b)

Figure 4: TSCFs and PAG for a typical vehicle trajectory sample. In (a), the ego-vehicle successfully overtakes by moving from the positive to the negative direction of the x-axis. In (b), the PAG is generated with internal causality discovery, with treatments corresponding to the variables in (a) and explanatory labels arranged in the temporal order.

the y-axis indicates the road's cross-section. The trajectory segment $F1$ represents the preparation stage for overtaking. Since the ego vehicle's initial position and driving direction remain constant across scenarios, the trajectories in this stage show limited variation. $F2$ illustrates the change in lane to the left, $F3$ represents the acceleration to overtake, and $F4$ represents the return to the main lane.

Using the optimized internal causality discovery algorithm, we generated the PAG graph in Figure 4b, showing the relationships between TSCFs and the overtaking result. Due to the initial environmental settings, $F1$ shows minimal variation across all samples and thus has no causal relationship with other variables. However, the TSCFs for the remaining stages, $F2$, $F3$, and $F4$ directly influence the overtaking result. Additionally, an unobserved confounding variable exists between $F2$ and $F3$, representing the overtaking intention defined by the scenario, and affects both factors.

The above analysis demonstrates that the ERCI framework enhances the explainability of the DRL training process by revealing clear causal relationships and ensuring the model's experiences and operations are understandable. This alignment with human intuition facilitates better comprehension and explainability of the model's decisions.

RQ2: Overall Performance We evaluated ERCI's training performance in various environments using DDPG and TD3 models, comparing the base models with their ERCI-enhanced versions. Specifically, we compared DDPG with DDPG-ERCI and TD3 with TD3-ERCI. The results are shown in rows 2 to 5 of Table 1. ERCI consistently outperforms the baseline models, with significant improvements in AS across all environments. A detailed analysis of each scenario follows. Our approach surpasses the baselines in most metrics in the Highway and Intersection scenarios, demonstrating its effectiveness in dynamic and complex traffic conditions. Performance consistently exceeds the baseline in the Racetrack, showing significant improvements in complex, extended time-step environments. The performance metrics

Models	Highway				Intersection				Racetrack				Reacher				Humanoid			
	AS	BS	SAS	ACS	AS	BS	SAS	ACS	AS	BS	SAS	ACS	AS	BS	SAS	ACS	AS	BS	SAS	ACS
DDPG	25	**29**	159	11835	19	**23**	**142**	9187	20	27	516	9055	-20	-8	192	-14679	282	430	377	115002
DDPG-ERCI	**26**	29	**59**	**13137**	**20**	23	194	**9382**	**33**	**42**	**178**	**15393**	**-18**	**-8**	**190**	**-12706**	**324**	**454**	**338**	**133718**
TD3	24	30	**329**	10659	16	19	400	7271	62	75	219	28153	-14	-11	**110**	-7498	305	522	**449**	108124
TD3-ERCI	**25**	**32**	330	**10869**	**17**	**19**	**357**	**7846**	**72**	**85**	**146**	**32660**	**-13**	**-9**	184	**-7309**	**343**	**689**	507	**119837**
DDPG-PER	26	29	**123**	12358	15	20	430	6527	31	37	297	14624	-20	-13	150	-13121	242	407	**344**	91525
DDPG-PERCI	**27**	**29**	136	**13095**	**17**	**23**	**400**	**7661**	**38**	**50**	**295**	**16537**	**-18**	**-9**	**144**	**-12336**	**400**	**589**	348	**156936**
TD3-PER	**27**	29	326	**12907**	18	**31**	479	8825	57	67	**229**	26411	-15	-10	116	-8372	299	532	**414**	115655
TD3-PERCI	24	**30**	**314**	10716	**19**	27	**396**	**9133**	**65**	**79**	268	**29449**	**-12**	**-9**	**115**	**-7011**	**381**	**613**	474	**154343**

Table 1: Training performance comparison of ERCI across different environments.

in Reacher are almost universally better, highlighting our advantage in precise motion tasks. In the Humanoid scenario, our approach performs highly in almost all metrics except SAS, excelling in high-dimensional action spaces. We conducted a comprehensive evaluation of DDPG-ERCI and TD3-ERCI with their baselines from four perspectives. AS reflects the agent's average performance level, BS reflects the best performance during training, SAS reflects the convergence speed, and ACS considers the cumulative performance, reflecting the trend of performance changes. The results in Table 2 show significant improvements in each metric across five environments. Notably, the AS score is 20.7% higher than the DDPG baseline. Our approach's effectiveness lies in identifying TSCFs that enhance rewards and focusing on sampling experiences based on them. Additionally, the model can quickly identify reasons for score decreases and make corrections, leading to higher overall scores and faster convergence for the optimized DRL model.

Model	AS ↑	BS ↑	SAS ↓	ACS ↑
DDPG-ERCI	20.70%	13.80%	-22.99%	22.96%
TD3-ERCI	8.92%	14.96%	-7.28%	7.85%

Table 2: Performance enhancements of ERCI over baseline.

Based on the performance of ERCI in all environments, the explainable experience replay approach with causal inference for DRL proposed in this paper significantly outperforms the baseline in overall performance across five environments. The decrease in SAS, in particular, reflects that the explainable ERCI approach can more effectively guide training in the early stages of DRL, improving sampling efficiency and thus achieving faster convergence.

RQ3: Scalability To demonstrate the scalability of our approach, we extended it to the PER algorithm, resulting in the Prioritized Experience Replay with Causal Inference (PERCI) approach. In PER, priorities are defined by TD-Error, which measures the difference between the predicted value and the observed value for a given state and action. The update rates of causal weights and TD-Errors in the experience replay pool are not synchronized. To align them, we proposed a hierarchical sampling approach where causal weights and TD- Errors are stored separately in two arrays of the same size. Based on which whenever the causal weights are updated, they are normalized and then linearly combined with the corresponding TD errors at the same position. The combined values serve as the leaf nodes of the causal replay weight tree for bottom-up overall updates. The comparison results are listed in rows 6 to 9 of Table 1.

Model	AS ↑	BS ↑	SAS ↓	ACS ↑
DDPG-PERCI	24.52%	26.93%	-0.45%	22.78%
TD3-PERCI	10.88%	6.52%	-1.83%	9.54%

Table 3: Performance enhancements of PERCI.

Similarly, Table 3 shows a comprehensive comparison of the PERCI with baseline, revealing significant improvements in most metrics. However, the optimization of SAS was less pronounced than that of ERCI. Upon analysis, this is fully reasonable and still demonstrates the superiority of PERCI. As the causal weights are diluted during updates with TD-Error, its impact on the experience sampling process is not as significant as with the ERCI. While SAS, which indicates convergence speed shows less improvement, other metrics still exhibit substantial enhancement. In summary, the extended PERCI approach enhances the explainability and performance of experience replay, outperforming the baseline on key metrics, which fully demonstrates the scalability of our proposed approach.

Conclusion

In this work, we propose a novel experience replay approach, ERCI, to enhance explainability and sample efficiency in DRL training. By employing a two-stage time series representation approach, we extract TSCFs to capture complex internal causality within multivariate time series, simplifying long sequences and addressing non-uniform time steps. We leverage causal inference to quantify the causal strength between TSCFs and DRL learning targets and refine the experience replay strategy. Our approach integrates ERCI with experience replay approaches across various complex DRL environments, proving its explainability, scalability and overall performance.

Acknowledgments

This work was supported in part by National Key R&D Program of China (No. 2022ZD0120302).

References

Andrychowicz, M.; Wolski, F.; Ray, A.; Schneider, J.; Fong, R.; Welinder, P.; McGrew, B.; Tobin, J.; Pieter Abbeel, O.; and Zaremba, W. 2017. Hindsight experience replay. *Advances in neural information processing systems*, 30.

Arrieta, A. B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58: 82–115.

Berndt, D. J.; and Clifford, J. 1994. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, 359–370.

Beyret, B.; Shafti, A.; and Faisal, A. A. 2019. Dot-to-dot: Explainable hierarchical reinforcement learning for robotic manipulation. In *2019 IEEE/RSJ International Conference on intelligent robots and systems (IROS)*, 5014–5019. IEEE.

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. Openai gym. *arXiv preprint arXiv:1606.01540*.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Cideron, G.; Seurin, M.; Strub, F.; and Pietquin, O. 2019. Self-educated language agent with hindsight experience replay for instruction following. *arXiv preprint arXiv:1910.09451*.

Fedus, W.; Ramachandran, P.; Agarwal, R.; Bengio, Y.; Larochelle, H.; Rowland, M.; and Dabney, W. 2020. Revisiting fundamentals of experience replay. In *International Conference on Machine Learning*, 3061–3071. PMLR.

Fujimoto, S.; Hoof, H.; and Meger, D. 2018. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, 1587–1596. PMLR.

Granger, C. W. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, 424–438.

Hallac, D.; Vare, S.; Boyd, S.; and Leskovec, J. 2017. Toeplitz inverse covariance-based clustering of multivariate time series data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 215–223.

Heuillet, A.; Couthouis, F.; and Díaz-Rodríguez, N. 2021. Explainability in deep reinforcement learning. *Knowledge-Based Systems*, 214: 106685.

Hyvärinen, A.; and Oja, E. 2000. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5): 411–430.

Juozapaitis, Z.; Koul, A.; Fern, A.; Erwig, M.; and Doshi-Velez, F. 2019. Explainable reinforcement learning via reward decomposition. In *IJCAI/ECAI Workshop on explainable artificial intelligence*.

Konda, V.; and Tsitsiklis, J. 1999. Actor-critic algorithms. *Advances in neural information processing systems*, 12.

Kuang, K.; Li, L.; Geng, Z.; Xu, L.; Zhang, K.; Liao, B.; Huang, H.; Ding, P.; Miao, W.; and Jiang, Z. 2020. Causal inference. *Engineering*, 6(3): 253–263.

Leurent, E. 2018. An Environment for Autonomous Driving Decision-Making. https://github.com/eleurent/highway-env.

Lillicrap, T. P.; Hunt, J. J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; and Wierstra, D. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Lin, L.-J. 1992. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine learning*, 8: 293–321.

Madumal, P.; Miller, T.; Sonenberg, L.; and Vetere, F. 2020. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2493–2500.

Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, 1928–1937. PMLR.

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.; et al. 2015. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533.

Ng, A. Y.; Harada, D.; and Russell, S. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, 278–287. Citeseer.

Ogarrio, J. M.; Spirtes, P.; and Ramsey, J. 2016. A Hybrid Causal Search Algorithm for Latent Variable Models. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, 368–379.

Pearl, J.; and Mackenzie, D. 2018. *The book of why: the new science of cause and effect*. Basic books.

Raffin, A.; Hill, A.; Traoré, R.; Lesort, T.; Díaz-Rodríguez, N.; and Filliat, D. 2019. Decoupling feature extraction from policy learning: assessing benefits of state representation learning in goal based robotics. *arXiv preprint arXiv:1901.08651*.

Schaul, T.; Quan, J.; Antonoglou, I.; and Silver, D. 2016. Prioritized experience replay. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Silver, D.; Lever, G.; Heess, N.; Degris, T.; Wierstra, D.; and Riedmiller, M. 2014. Deterministic policy gradient algorithms. In *International conference on machine learning*, 387–395. Pmlr.

Spirtes, P.; and Glymour, C. 1991. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1): 62–72.

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

Sullivan, R. S.; and Longo, L. 2023. Explaining Deep Q-Learning Experience Replay with SHapley Additive exPlanations. *Machine Learning and Knowledge Extraction*, 5(4): 1433–1455.

Vecerik, M.; Hester, T.; Scholz, J.; Wang, F.; Pietquin, O.; Piot, B.; Heess, N.; Rothörl, T.; Lampe, T.; and Riedmiller, M. 2017. Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. *arXiv preprint arXiv:1707.08817*.

Volodin, S.; Wichers, N.; and Nixon, J. 2020. Resolving spurious correlations in causal models of environments via interventions. *arXiv preprint arXiv:2002.05217*.

Wang, J.; Zhang, Y.; Kim, T.-K.; and Gu, Y. 2020. Shapley Q-value: A local reward approach to solve global reward games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 7285–7292.

Wang, J. X.; Kurth-Nelson, Z.; Tirumala, D.; Soyer, H.; Leibo, J. Z.; Munos, R.; Blundell, C.; Kumaran, D.; and Botvinick, M. 2016. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.

Wilson, S. J. 2017. Data representation for time series data mining: time domain approaches. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(1): e1392.

Yeh, C.-C. M.; Kavantzas, N.; and Keogh, E. 2017. Matrix profile VI: Meaningful multidimensional motif discovery. In *2017 IEEE international conference on data mining (ICDM)*, 565–574. IEEE.

Young, G.; et al. 2016. Unifying causality and psychology. *Cham: Springer*.

Zeng, Y.; Cai, R.; Sun, F.; Huang, L.; and Hao, Z. 2023. A survey on causal reinforcement learning. *arXiv preprint arXiv:2302.05209*.

Zhang, J. 2008. Causal Reasoning with Ancestral Graphs. *Journal of Machine Learning Research*, 9(7).

Zhang, S.; and Sutton, R. S. 2017. A deeper look at experience replay. *arXiv preprint arXiv:1712.01275*.