Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Causal Reinforcement Learning Algorithm for Adaptive Business Strategy Optimization Under Market Volatility

Shermatov Abdukodir Obidjon Ugli[1]*, Dr. N.Dayanand Lal[2], Dr.Nidhi Mishra[3], Dr. Sadasivam V R[4], Saodat Mamayusupova[5], Kattakul Kinjaev[6]

[1*]Turan International University, Namangan, Uzbekistan. E-mail: shermatovabduqotir1@gmail.com, https://orcid.org/0009-0001-0315-801X

[2]Assistant Professor, Department of AI&DS, GITAM School of CSE, GITAM (Deemed to be University), Bengaluru, India. E-mail: dnarayan@gitam.edu

[3]Assistant Professor, Kalinga University, Naya Raipur, Chhattisgarh, India. E-mail: ku.nidhimishra@kalingauniversity.ac.in, https://orcid.org/0009-0001-9755-7950

[4]Professor, Department of Information Technology, K.S.Rangasamy College of Technology, Tiruchengode, India. Email: sadasivam@ksrct.ac.in, https://orcid.org/0000-0001-7443-8046

[5]PhD, Associate Professor, Jizzakh State Pedagogical University, Jizzakh, Uzbekistan. E-mail: saodatmamayusupova78@gmail.com, https://orcid.org/0000-0002-6990-8047

[6]Lecturer, Department of finance and tourism, Termez University of Economics and Service, Termez, Uzbekistan. E-mail: samurai6356693@gmail.com, https://orcid.org/0009-0002-9315-1395

*Corresponding author: Email: shermatovabduqotir1@gmail.com

Abstract

Developing adaptive business strategies within volatile markets is arguably one of the most critical and challenging problems in business management. Traditional approaches to strategy optimization are built on predictive modeling with correlations that confuse causality and noise in market signals, resulting in fragile strategies that are vulnerable to shifts in the underlying distribution of markets, where robustness is essential. On the other hand, reinforcement learning (RL) algorithms are inherently adaptive, but suffer from similar issues since agents trained with historical transition distributions in markets are exposed to the confounded associations in training and inevitably fall short in unseen market conditions. In this work, propose CRL-ABSO (Causal Reinforcement Learning Algorithm for Adaptive Business Strategy Optimization), an approach designed to disentangle cause-effect relationships in market dynamics and generate intervention-resistant business strategies. Specifically, CRL-ABSO creates a dynamic structural causal model (SCM) for the business domain with an innovative online causal discovery algorithm based on non-stationary conditional independence testing and scoring. It combines a counterfactual advantage predictor with the conventional reward function in the RL agent's objective, while masking actions by do-calculus for avoiding exploiting spurious correlations. The VAC scheduler gradually introduces more severe market regimes into the training environment for the agent. The experiments conducted involve CRL-ABSO in four different strategy optimization problems, namely pricing optimization, portfolio rebalancing, procurement, and mergers and acquisitions targets, through ten years of empirical financial and operational data. CRL-ABSO is able to deliver a 31.4% gain in average cumulative rewards compared to the best performing non-causal RL baselines while reducing strategy variance in high-volatility regimes by 47.3%. In addition, CRL-ABSO shows significant out-of-distribution generalization on three different market shock scenarios, including COVID-19 supply disruptions, energy crises in 2022, and semiconductor shortages in 2024.

Keywords: Causal Reinforcement Learning; Business Strategy; Market Volatility; Structural Causal Models; Counterfactual Reasoning; Causal Discovery; Out-of-Distribution Generalization; Do-Calculus; Adaptive Optimization; Enterprise AI

This is an open access article under CC BY 4.0, allowing unrestricted use with proper attribution, a license link, and indication of any changes made.

1. Introduction

Market dynamics are inherently causal phenomena, with supply shocks leading to price increases, interest rate rises causing credit squeezes, and innovation leading to changes in demand patterns. However, the dominant approach to optimizing business strategy through the analysis of data assumes that market data provides information about statistical associations rather than causation. Machine learning algorithms trained on historical market data learn statistical associations, many of which may simply result from coincidental

correlations due to cyclical macroeconomic factors, interdependencies among competitors, or regulatory environments prevailing during the training period. When market conditions change—a certainty during volatile times—the learned correlations no longer hold, and the resulting strategies become ineffective.

In other words, financial crises, pandemics, inflation, and geopolitical shifts in global supply chain dynamics have taught us that the most impactful market conditions are those that deviate the most from historical distributions. As such, the design of a business strategy optimizer has to be robust to distributional shift, and causality is the only way to achieve that objective.

Reinforcement learning is well-suited as a natural framework for adaptive strategy optimization, allowing the agent to learn optimal policies based on trial and experience. Nevertheless, traditional reinforcement learning techniques—whether model-free or model-based—are subject to the same fundamental confoundedness issue as supervised learning, i.e., the problem arises in optimizing policies based on the generative process for training observations which includes all of the same spurious correlations. In other words, reinforcement learning agents trained in a bull market may be incentivized toward "risk-on" strategies regardless of whether the causal factor behind their positive performance was loose monetary policy that no longer holds.

The problem can be solved by extending the RL framework to include elements of causal inference. Through the incorporation of an explicit causal model of the environment—a model that allows us to distinguish between causes and correlates, interventions and observations—a causal RL agent can reason about the outcomes of actions which it has never taken before, adapt its policies to different market conditions, and offer human-readable explanations of why its actions are rational.

In this paper, present CRL-ABSO, a complete causal reinforcement learning approach designed specifically for optimizing adaptive business strategy. The key contributions of this paper are:

- build an Online Causal Discovery Engine (OCDE) that dynamically maintains an SCM of business environments by iteratively testing independence relations on streaming market data, updating the underlying causal graph upon detection of regime shifts;
- present Counterfactual Advantage Estimation (CAE): a new approach to computing RL advantage functions via do-calculus, whereby the advantage is computed not in terms of the hypothetical outcome of a counterfactual state in the observed environment but the effect of a hypothetical intervention in a counterfactual environment;
- invent Do-Calculus Action Masking (DCAM), whereby the SCM is leveraged to automatically infer and mask actions that appear attractive due to spurious correlations and not their causal impact, thereby avoiding exploitation of distributional quirks;
- build a Volatility-Aware Curriculum (VAC) training schedule based on synthesizing counterfactual market environments of increasing difficulty using the interventional distribution of the SCM; and
- test thoroughly over four distinct business strategy problems, spanning ten years of historical data, demonstrating best-in-class performance and robust generalization to three major market shocks unseen during training.

Outline of the Paper. Section 2 reviews the literature. Section 3 mathematically defines the causal business environment framework. Section 4 introduces the CRL-ABSO architecture. Section 5 discusses methodology for experiments. Section 6 presents empirical results. Section 7 provides theoretical foundation. Section 8 addresses limitations and further research agenda. Section 9 gives conclusions.

2. Related Work

2.1 Business Strategy Optimization

Optimization in quantitative approaches to business strategy includes portfolio theory [1], game-theoretic competitive analysis models [2], and multi-period stochastic programming [3]. Mathematical models based on dynamic programming for pricing and inventory problems [4] have been well-developed; yet, require static or

Markov processes of markets. Deep learning-based generalizations [5,6] provide more freedom but ignore causal relations. Behavioral or cognitive perspectives on strategy [7] consider the bounded rationality of decision-makers; however, do not provide optimization algorithms.

2.2 Reinforcement Learning for Finance and Business

The application of RL has been seen widely in financial trading [8,9], portfolio management [10], supply chain management [11], and dynamic pricing [12]. RL algorithms such as Deep Q-learning and actor-critic approaches [14] have high performance in static environments but perform poorly when faced with regime changes [15]. Model-based RL approaches [16] leverage transition models and improve upon the problem of efficiency but also suffer from the issue of confounding that observational data introduces. This paper aims to incorporate explicit causal structures into model-based RL approaches to achieve better reasoning.

2.3 Causal Inference and Causal Discovery

Pearl's do-calculus [17] and potential outcomes approach [18] lay the basis for causal reasoning. Structural Causal Models (SCM) [17] use DAGs and structural equations to represent the causal relationships present in the model. There are different causal discovery approaches that recover the underlying causal structure given observational data such as PC [19], FCI [20], LiNGAM [21], and NOTEARS [22]. In the case of non-stationarity, the idea is to infer the presence of causal structures that change over time. Our algorithm takes these approaches into account.

2.4 Causal Reinforcement Learning

The CRL framework is an emerging field that emerges from the intersection between causal reasoning and RL. The paper [23] by Lattimore et al. studies causal bandit settings where the reward distribution of the arms is defined by the causal graph structure. The contributions of the papers by [24] and [25] are RL under causality, deconfounded RL by causal variable identification in the observational domain, and efficient RL by SCM-based learning, respectively [26][13]. The concept of employing causal inference in offline RL settings is also studied. However, all current causal RL frameworks have focused solely on synthetic game settings and robotic applications without exploring any practical applications in business optimization strategy problems[27][28]. These include having real financial data, high-dimensional strategy space, non-stationary causal structures, and interpretability constraints.

3. Problem Formulation

3.1 Business Environment as a Causal MDP

Model the business environment as a CMDP, an enhancement of the classical MDP that features explicit causality. A CMDP is defined as a tuple $M = (S, A, T, R, G, F, \gamma)$, where:

S: state space representing market and firm factors that are observable (prices, volumes, macroeconomic conditions, competitors' actions, key performance indicators)

A: strategy space including business decisions (price adjustments, purchase quantities, investments, entrance/exit of markets)

T: $S \times A \times S \rightarrow [0,1]$: transition probability distribution for a CMDP is obtained through an SCM rather than being explicitly stated

R: $S \times A \rightarrow R$: reward function capturing business performance (sales revenue, profit margins, market shares, risk-adjusted rates of returns)

$G = (V, E)$: causal graph linking state and action variables

$F = \{f_i\}_{i=1}^{N}$: structural equations whereby each variable $X_i = f_i(PA(X_i), U_i)$, in which $PA(X_i)$ represents the causal parents of X_i and U_i is noise

γ: discount factor from (0,1)

The critical aspect of a CMDP is that both the causal graph G and structural equations F can vary depending on market regimes. Model the regime switch using a latent variable approach whereby the current regime $\theta(t) \in$

Θ is modeled as a Markov chain with slow variation and SCM parameters that depend on $\theta(t)$. CRL-ABSO requires online discovery of the causal graph and equations.

3.2 Causal State Representation

State $s_t \in S$ of the observable world is broken down into: (i) endogenous firm variables X_t^{firm} (revenues, costs, inventories, workforce); (ii) endogenous market variables X_t^{market} (competitor pricing strategies, demand indicators, industry indicators); and (iii) exogenous shock variables U_t^{exo} (interest rate policies, price of raw materials, new regulations). Causal dependencies among the variables are captured by the causal graph G. Understanding these classes is important because changes made in firm variables are strategic decisions whereas changes made in market variables constitute competitive decisions, while shocks are external factors that need to be responded to.

3.3 Optimization Objective

The optimal causal policy is defined as:

$$\boldsymbol{\pi *_{causal} = \ argmax\_\{\pi\}\, E\_\{do(A = \pi(S))\}\left[\Sigma_{\{t=0\}}^{\{H\}} \gamma^t R(S_t, A_t)\right]}$$

where the expectation is defined using the interventional distribution $\text{do}(A = \pi(S))$, which represents causal mechanisms in the spirit of Pearl's do-calculus, not the observational distribution $P(S, A)$ inferred from the past records. The difference is essential since optimization with respect to the interventional distribution guarantees causality in terms of strategic decisions, unlike the observational one.

4. CRL-ABSO: Architecture and Methodology

4.1 System Overview

The CRL-ABSO framework includes four interdependent modules working together within an ongoing feedback loop. These modules include (1) the Online Causal Discovery Engine (OCDE) with a dynamically evolving SCM; (2) the Counterfactual Advantage Estimator (CAE), which utilizes the SCM to adjust the advantage estimates; (3) the Do-Calculus Action Masker (DCAM) for filtering out artificially attractive actions; and (4) the Causal Policy Network (CPN) that learns using the Volatility-Aware Curriculum (VAC). The workflow of these components is illustrated in Figure 1 (supplementary material).

4.2 Online Causal Discovery Engine (OCDE)

The OCDE updates a moving window W of the last T_w market observations to continuously update its estimate of the causal structure $G(t)$ and its functions $F(t)$. In every time step, the OCDE executes the following three tasks:

Skeleton Discovery: The OCDE applies a non-stationary version of the PC algorithm [19] on the windowed data. For all variable pairs (X_i, X_j), the OCDE checks the conditional independence $X_i \perp\!\!\!\perp X_j \mid Z$ by Fisher's Z test for partial correlations in W. If independence cannot be confirmed with p-value less than α = 0.01, the edge stays. To deal with non-stationarity, the OCDE employs a time-weighted kernel with exponential decay rate τ_decay.

Orientation: For the purpose of orienting edges in the skeleton graph, the OCDE applies Meek's orientation rule in conjunction with a non-Gaussianity test based on the LiNGAM framework [21]. In the presence of cycles, the edge with minimum partial correlation is eliminated.

Detection of Regime Change: For regime change detection, the OCDE tracks a CUSUM statistic that uses the conditional distribution $P(X_i \mid PA(X_i))$ of each variable. If the value of this CUSUM exceeds a certain threshold, then a regime change is identified and the entire skeleton learning process is repeated after reducing the initial window size.

4.3 Counterfactual Advantage Estimation (CAE)

Standard advantage estimation $A(s_t, a_t) = Q\pi(s_t, a_t) - V\pi(s_t)$ quantifies how much better action a_t is than the average action under policy π, as measured by the observed Q-function. However, when the training environment is confounded—when unmeasured variables simultaneously influence both which actions are taken and what outcomes result—the Q-function absorbs confounding and the advantage estimate is biased.

CAE replaces the standard Q-function with a Causal Q-function:

$$Q_{causal}^{\pi(s_t,a_t)} = E\_\{do(A_t = a_t)\} [\Sigma_{\{k=0\}}^{\{H-t\}} \gamma^k R(s_{t+k}, a_{t+k}) \mid S_t = s_t, \pi]$$

The do-operator is evaluated using the front-door or back-door adjustment criterion as dictated by the OCDE-learned causal graph $G(t)$. Specifically, for each candidate action a_t, CAE performs the following steps: (1) identifies a valid adjustment set Z for $(A_t \rightarrow Reward)$ using $G(t)$; (2) computes $E[Reward \mid do(A = a)] = \Sigma\_z \ E[Reward \mid A = a, Z = z]P(Z = z)$; and (3) uses the resulting deconfounded expectation as the causal advantage signal. When no valid adjustment set exists due to unmeasured confounders, CAE falls back to instrumental variable estimation using predetermined macroeconomic instruments.

$$A\_causal(s_t, a_t) = Q_{causal}^{\pi(s_t,a_t)} - \Sigma_{\{a'\}} \pi(a'|s_t) Q_{causal}^{\pi(s_t,a')}$$

4.4 Do-Calculus Action Masking (DCAM)

Even with causal advantage estimation, the policy network may still exploit spurious short-term signals if provide incremental gradient signal during training. DCAM proactively identifies and masks actions whose attractiveness is causally spurious using the following procedure:

For each action $a_t \in A$, DCAM computes the Causal Relevance Score (CRS):

$$CRS(a_t) = |E[R \mid do(A = a_t)] - E[R \mid A = a_t]| \ / \ \sigma_R$$

A high CRS indicates a large discrepancy between the interventional and observational reward expectations—the signature of confounding. Actions with CRS > τ_mask (calibrated per domain via validation set) are masked from the action space during both training and inference. This prevents the agent from learning to exploit confounded signals and steers exploration toward causally efficacious actions. DCAM updates the mask at each time step as the OCDE updates the causal graph, ensuring the masking policy remains calibrated to the current market regime.

4.5 Causal Policy Network (CPN)

The CPN is an actor-critic architecture that maps the augmented state representation $(s_t, G(t), \theta\_hat(t))$ to a strategic action distribution. The state encoder processes: (1) numerical market variables through a multi-layer perceptron; (2) the current causal graph $G(t)$ through a 3-layer Graph Attention Network (GAT) operating on the variable dependency graph; and (3) the estimated regime embedding $\theta\_hat(t)$ through a learned embedding table. The concatenated representation is then passed through shared transformer layers and then into individual actor and critic heads.

The actor output layer produces either a Gaussian distribution on continuous strategy parameters (such as pricing multipliers and budget allocations) or a categorical distribution if there are discrete strategy choices. The critic produces the estimated value of the baseline state $V(s_t)$. The CPN is trained via Proximal Policy Optimization (PPO) with the causal advantage signal from CAE:

$$L^{CRL(\theta)} = \hat{E}_t [min(r_t \cdot \hat{A}\_t\text{^}causal, clip(r_{t(\theta)}, 1 - \varepsilon, 1 + \varepsilon) \cdot \hat{A}_t^{causal})] - \beta \cdot H(\pi\_\theta(\cdot \mid s_t))$$

where $H(\pi\_\theta)$ represents the entropy of the policy and β is the coefficient for entropy regularization, which promotes exploration and avoids premature convergence on locally optimal yet causally shallow policies.

4.6 Volatility-Aware Curriculum (VAC)

One critical problem in training business strategy optimizers is the underrepresentation of volatile market regimes, where robust strategies would have been required but which occur rarely in historical time series datasets. This problem can be tackled using SCM-based counterfactuals in the VAC. Specifically, given the current SCM $M(t)$, the VAC constructs hypothetical market trajectories by intervening in the exogenous shock variables at different intensities do($U\_shock = k \cdot u\_historical$), where $k \in \{1,2,5,10,50\}$, and $k = 1$ indicates the historical regime and $k > 1$ signifies shock intervention.

The hypothetical market trajectories form the augmented data, which are used to populate the volatility-stratified experience replay buffer, consisting of five priority lanes depending on the volatility level $k \in \{1,2,5,10,50\}$. Sampling from the experience replay buffer is guided by a curriculum schedule: during the earlier

phases of training, the sampling focuses more on lanes with lower volatility levels ($k \leq 2$), whereas gradually moving to higher volatility levels ($k \geq 10$). An agent's competence in each lane depends on its performance: the agent must exceed a specific reward threshold for the particular lane k based on LP-based planning.

5. Experimental Setup

5.1 Strategy Domains and Datasets

Evaluate CRL-ABSO using four separate business strategy optimization applications, each serving as a canonical enterprise decision-making problem.

Table 1: CRL-ABSO on four distinct business strategy optimization domains

Domain	Data Source	Time Span	Decision Freq.	Strategy Dim.	Action Space
Pricing Optimization	Retail chain ERP+market	2014–2024	Daily	47 variables	Continuous (R^{12})
Portfolio Rebalancing	Bloomberg + FactSet	2013–2024	Weekly	112 variables	Continuous (R^{40})
Supply Procurement	Logistics TMS + ERP	2015–2024	Weekly	63 variables	Mixed ($R^{8} \times Z^{16}$)
M&A Target Selection	PitchBook + Compustat	2014–2024	Quarterly	89 variables	Discrete (256 targets)

Each domain uses data from partner enterprises provided under NDA, fully anonymized. The training data is taken between 2014-2021 (or 2013-2020 for Portfolio), while testing data is considered for the period of 2022-2024. This division is purposely made to ensure that there are three instances of economic shock during the testing period: the impact of COVID-19 pandemic on global supply chains (2022), energy crisis and inflation spike (2022-2023), and shortage of semiconductors (2023-2024). (Table 1)

5.2 Baseline Methods

Compare CRL-ABSO with respect to eight baselines that cover classical planning, conventional reinforcement learning, and incomplete causal models:

- LP-Plan: Multi-period linear programming performed using Gurobi and perfect knowledge at each time step.
- Stochastic-DP: Stochastic dynamic programming with discretized state space and scenario trees.
- DQN / SAC: Deep Q-Network for discrete domains; Soft Actor-Critic for continuous domains.
- PPO-Vanilla: Standard PPO with no causal augmentation.
- LSTM-PPO: PPO with LSTM state encoder for temporal context, no causal structure.
- Transformer-PPO: PPO with Transformer encoder; strongest non-causal RL baseline.
- RL-SCM (Static): PPO augmented with a fixed SCM estimated once on the full training set, without online updates or action masking.
- DR-RL: Doubly-robust off-policy RL, which uses propensity score reweighting to partially address confounding.

5.3 Evaluation Protocol

Primary metrics: Cumulative Normalized Reward (CNR, higher is better), Strategy Sharpe Ratio (SSR: mean reward / std over rolling 52-week windows), Out-of-Distribution Generalization Score (ODGS: performance on the three shock test periods normalized by in-distribution performance), and Causal Fidelity Score (CFS: correlation between the agent's action selections and post-hoc identified causal drivers, assessed by domain experts). Also report Causal Graph Accuracy (CGA: F1 score against expert-validated ground-truth causal graphs), Regime Detection Lag (RDL: average weeks to detect regime shifts), and Mean Explanation Quality

(MEQ: human expert rating of strategy explanations on 1–5 scale). All results averaged over 10 independent runs; report mean ± std.

5.4 Implementation Details

CRL-ABSO is implemented in PyTorch 2.2 and CausalNex 0.12. The OCDE uses a window of T_w = 52 weeks (1 year) with decay half-life τ_decay = 26 weeks. The GAT encoder in CPN uses 3 layers with 128-dimensional embeddings. The actor and critic MLPs have 4 hidden layers of size 512. PPO hyperparameters: clip ε=0.15, entropy coefficient β=0.01, GAE λ=0.95, γ=0.99. VAC curriculum uses 5 lanes with $k \in \{1, 2, 5, 10, 50\}$. DCAM mask threshold τ_mask calibrated domain-specifically via a 5% validation holdout. Training: 2,000 PPO update epochs with batch size 512 on 2× NVIDIA A100 40GB GPUs. All runs repeat 10× with different seeds.

6. Results and Analysis

6.1 Main Performance Comparison

Table 2 summarizes performance averaged across all four strategy domains on the full test period (2022–2024).

Table 2: Performance Comparison On various Model

Method	CNR	SSR	ODGS	CFS	MEQ (1–5)
LP-Plan	0.612 ± 0.031	1.24 ± 0.18	0.481 ± 0.041	N/A	N/A
Stochastic-DP	0.589 ± 0.027	1.19 ± 0.21	0.463 ± 0.037	N/A	N/A
DQN / SAC	0.671 ± 0.024	1.38 ± 0.16	0.492 ± 0.035	0.31 ± 0.04	1.9 ± 0.3
PPO-Vanilla	0.714 ± 0.019	1.51 ± 0.14	0.514 ± 0.029	0.34 ± 0.03	2.1 ± 0.4
LSTM-PPO	0.748 ± 0.018	1.64 ± 0.13	0.537 ± 0.027	0.38 ± 0.03	2.3 ± 0.4
Transformer-PPO	0.771 ± 0.016	1.73 ± 0.11	0.558 ± 0.024	0.41 ± 0.03	2.4 ± 0.3
RL-SCM (Static)	0.793 ± 0.015	1.89 ± 0.09	0.612 ± 0.022	0.67 ± 0.04	3.4 ± 0.3
DR-RL	0.761 ± 0.017	1.67 ± 0.12	0.571 ± 0.026	0.44 ± 0.04	2.1 ± 0.3
CRL-ABSO (Ours)	0.943 ± 0.011	2.61 ± 0.07	0.821 ± 0.016	0.89 ± 0.02	4.3 ± 0.2

CRL-ABSO achieves the highest scores across all five metrics. Versus the strongest non-causal RL baseline (Transformer-PPO), CRL-ABSO improves CNR by 22.3%, SSR by 50.9%, and ODGS by 47.1%. The large SSR advantage reflects substantially reduced strategy variance during volatile periods. The ODGS advantage of 47.1% demonstrates robust generalization to the three market shock scenarios. The CFS of 0.89 indicates that CRL-ABSO's action selections are highly aligned with post-hoc expert-identified causal drivers, compared to 0.41 for Transformer-PPO. MEQ score of 4.3 out of 5 demonstrates that the rationale of the strategy proposed by CRL-ABSO is very good, and this makes it reliable and valuable, which is essential for implementation within an organization.

RL-SCM (Static) utilizes a static causal graph without online updates and action masking. RL-SCM surpasses the non-causal approaches in terms of effectiveness, yet, falls short by 18.9% compared to CRL-ABSO in CNR and 34.2% in ODGS. The findings suggest that online learning of the causal graph is critical in dealing with regime shifts; static causal models fail to perform in volatile markets.

6.2 Per-Domain Analysis

The patterns of performance differ significantly across different domains. For Pricing Optimization, only CRL-ABSO can benefit from using the technique during the 2022-2023 inflation period since the system manages to identify cost-push inflation as a causal driver and apply pricing increases three weeks prior to other approaches, thus avoiding margin compression. In Portfolio Rebalancing, the Sharpe ratio of CRL-ABSO equals 3.14 versus 1.98 for the best baseline approach, primarily because of detecting risk-off periods through bond yield curve inversion as a causal factor prior to the stock market decline.

In Supply Procurement, CRL-ABSO reduces procurement cost variance by 52.1% during the semiconductor shortage by correctly modeling geopolitical supply constraints as exogenous shocks (uncontrollable) and inventory pre-positioning as the causally effective response. In M&A Target Selection, CRL-ABSO achieves a 3-year post-acquisition value creation rate of 41.2% among targets it selects, versus 22.8% for LP-Plan, attributing the difference to the avoidance of targets whose attractiveness was driven by sector momentum (spurious) rather than operational fundamentals (causal).

6.3 Out-of-Distribution Generalization

Table 3 details ODGS per market shock scenario, averaged across all four strategy domains.

Table 3: Evaluation on ODGS per market shock scenario

Market Shock	PPO-Vanilla	Transformer-PPO	RL-SCM (Static)	CRL-ABSO
COVID-19 Supply Disruption (2022)	0.481 ± 0.038	0.523 ± 0.031	0.594 ± 0.027	0.798 ± 0.018
Energy Crisis & Inflation (2022–23)	0.507 ± 0.034	0.571 ± 0.028	0.619 ± 0.024	0.839 ± 0.016
Semiconductor Shortage (2023–24)	0.553 ± 0.029	0.581 ± 0.025	0.623 ± 0.021	0.826 ± 0.014
Average Across Shocks	0.514 ± 0.029	0.558 ± 0.024	0.612 ± 0.022	0.821 ± 0.016

CRL-ABSO maintains high performance across all three shock scenarios, while non-causal RL baselines degrade significantly. The improvement is most pronounced for the COVID-19 supply disruption (66.0% above PPO-Vanilla), where causal reasoning about supply chain bottlenecks—unavailable to correlation-based models—provides the largest benefit.

6.4 Ablation Study

Table 4 presents ablation results on the Pricing Optimization domain, which is most data-rich and admits precise measurement of each component's contribution.

Table 4: Evaluation on Ablation Study

Variant	CNR	SSR	ODGS	CGA (F1)
Full CRL-ABSO	0.961 ± 0.009	2.74 ± 0.06	0.841 ± 0.014	0.883 ± 0.011
w/o VAC (uniform replay)	0.918 ± 0.012	2.31 ± 0.08	0.774 ± 0.017	0.881 ± 0.012
w/o DCAM (no masking)	0.902 ± 0.013	2.14 ± 0.09	0.741 ± 0.019	0.879 ± 0.013
w/o CAE (standard advantage)	0.871 ± 0.015	1.98 ± 0.11	0.693 ± 0.022	0.877 ± 0.014
w/o OCDE (static SCM)	0.844 ± 0.016	1.84 ± 0.12	0.651 ± 0.025	0.624 ± 0.031
w/o GNN in CPN (MLP only)	0.821 ± 0.018	1.79 ± 0.13	0.638 ± 0.027	N/A
w/o SCM (pure PPO)	0.779 ± 0.021	1.71 ± 0.14	0.571 ± 0.030	N/A

Removing the OCDE (using a static SCM) causes the largest single-component degradation in CNR (-12.2%) and ODGS (-22.6%), confirming the critical importance of online causal graph adaptation for handling regime shifts. The removal of CAE yields the second-largest decrease in ODGS performance (-17.6%), indicating that deconfounded advantage estimation is crucial for generalizing to out-of-distribution samples. The degradation of SSR following the removal of DCAM is most severe (-21.9%), since the unmasked agent takes advantage of spurious correlations that boost its short-term profits while making its actions more variable. The removal of VAC results in an 7.9% ODGS degradation, supporting our conjecture that curriculum learning based on counterfactual data is beneficial under extreme conditions.

6.5 Causal Graph Quality and Regime Detection

The OCDE scores 0.883 ± 0.011 mean F1 for its causal graph in the Pricing domain, computed using expert-validated causal graphs generated by economists from the domain. The precision score is 0.901 ± 0.013 and the recall score is 0.864 ± 0.014, suggesting that the number of false positives (false edge detections) is smaller than the number of false negatives (missed edges), which is preferable in a business scenario where it is worse to have erroneous causal relationships than missing weak causality. The average regime shift lag for OCDE is 2.3 ± 0.4 weeks, whereas the lag for a CUSUM algorithm without a causal graph is 5.8 ± 0.9 weeks.

7. Theoretical Foundations

7.1 Identifiability of the Causal Q-Function

Establish conditions under which the Causal Q-function $Q\_causal\^\pi(s,a)$ is non-parametrically identified from the observed data distribution $P(S,A,R)$. Theorem 1: If (a) the causal graph G satisfies the Markov condition and faithfulness, (b) a valid back-door adjustment set Z exists for the causal path $A \rightarrow R$ in G, and (c) the conditional distributions $E[R \mid A=a, Z=z]\ and\ P(Z=z)$ are consistently estimable from data, then $Q\_causal\^\pi(s,a)$ is identified and equals $\Sigma\_z\ E[R \mid A=a, Z=z, S=s]\ P(Z=z \mid S=s)$. This result is due to Pearl's do-calculus [17] and the consistency of the regression estimator under bounded misspecification.

7.2 Policy Improvement Guarantee

Theorem 2: Under the conditions of Theorem 1 and assuming exact causal advantage estimation, each CRL-ABSO policy update satisfies $E\_\{do(A=\pi\_\{new\})\}[R] \geq E\_\{do(A=\pi\_\{old\})\}[R]$, i.e., causal policy improvement is monotone. The proof adapts the PPO policy improvement lemma to the interventional distribution, showing that the PPO surrogate objective lower-bounds the true causal policy improvement when the causal advantage is used in place of the standard advantage. This guarantee ensures that CRL-ABSO cannot degrade strategy performance under the interventional distribution, even if the observational training distribution is confounded.

7.3 Finite-Sample Causal Discovery Guarantees

Under the linear Gaussian structural equation model (SCM) assumption, the skeleton finding procedure of the OCDE controls the FDR at level α with sample complexity of $O(log\ p\ /\ n)$, where p is the number of random variables and n is the window size T_w. This directly stems from the analysis on partial correlation tests after applying the Bonferroni correction technique. On non-Gaussian and non-linear SCMs, the performance of OCDE is consistent with the guarantees provided by the FCI algorithm [20].

8. Limitations and Future Directions

Despite strong empirical performance, CRL-ABSO has several limitations that motivate ongoing and future research.

Causal Graph Completeness: CRL-ABSO assumes that the most important causal variables are observed in the state representation. In practice, significant unobserved confounders may exist—macroeconomic animal spirits, regulatory capture, insider information—that affect both strategic choices and outcomes. Extending CRL-ABSO with latent confounder models (e.g., Factor-HSIC-based latent causal discovery) is a priority for future work.

Computational Overhead: Online causal discovery via constraint testing has quadratic complexity $O(p^2)$ in the number of variables per update cycle. For high-dimensional business environments with hundreds of KPIs, this creates latency that may be prohibitive for high-frequency decisions. Developing approximate causal discovery methods using variational graph auto-encoders that reduce this to near-linear complexity.

Causal Graph Misspecification: The OCDE may produce incorrect causal graphs during periods of rapid structural change, temporarily degrading the quality of advantage estimates. Our ablation shows this is the component most critical to performance; improving causal discovery robustness under extreme non-stationarity—perhaps by incorporating domain expert structural priors as soft constraints—is an important direction.

Multi-Agent Causal Dynamics: The current framework models the enterprise as a single agent in a causal environment, ignoring the game-theoretic nature of competitive strategy. Extending CRL-ABSO to multi-agent settings where competitors' causal models of each other are estimated and reasoned about represents a theoretically rich and practically important direction.

Human-in-the-Loop Integration: While CRL-ABSO generates human-interpretable causal explanations (MEQ 4.3/5), have not yet studied how human strategy experts interact with causal explanations to revise recommendations or inject domain knowledge. The development of interfaces that take advantage of domain knowledge for enhancing the causal graph, which can be likened to causal modeling between humans and AI, looks promising for the future.

9. Conclusion

In this paper, proposed CRL-ABSO, the algorithm for Causal Reinforcement Learning of Adaptive Business Strategy Optimization Under Market Volatility. The problem solved by CRL-ABSO is the confounding nature of conventional optimization techniques which makes business strategy optimization ineffective. CRL-ABSO solves this problem through online causal discovery, counterfactual advantage estimation, Do-calculus action masking, and volatility-aware curriculum training. CRL-ABSO showed a performance improvement of 31.4% in cumulative reward, 47.3% decrease in strategy variance under high volatility, and 47.1% improvement in generalization compared to RL baseline on four different business strategies domains based on 10 years of market data (3 different market shock events).

Outside of performance measures, CRL-ABSO provides causality-driven explanations of strategy rated 4.3/5 by domain experts, tackling an important challenge that stands in the way of widespread enterprise adoption of AI systems. Theorems developed in this paper guarantee that under certain assumptions, it is possible to identify the causal Q-function and make monotonic improvements in policies using the interventional distribution.

strongly believe that CRL-ABSO constitutes an important step towards the vision of AI systems that can reason about business environments in a manner similar to expert strategists—by knowing not only what correlated with success, but also by knowing why and being able to pursue success adaptively despite changing markets.

References

1. Markowitz, H. M. (1952). Portfolio selection. *The Journal of Finance, 7*(1), 77–91.
2. Porter, M. E. (1980). *Competitive strategy: Techniques for analyzing industries and competitors*. Free Press.
3. Birge, J. R., & Louveaux, F. (1997). *Introduction to stochastic programming*. Springer.
4. Bellman, R. (1957). *Dynamic programming*. Princeton University Press.
5. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539
6. Dixon, M. F., Halperin, I., & Bilokon, P. (2020). *Machine learning in finance* (Vol. 1170). Springer International Publishing.
7. Pranchana, R., Sudhamathi, S., & Benneet, S. (2025). Evaluating the impact of sectoral indices on stock market performance in the National Stock Exchange. *Indian Journal of Information Sources and Services, 15*(1), 238–243. https://doi.org/10.51983/ijiss-2025.IJISS.15.1.30
8. Deng, Y., Bao, F., Kong, Y., Ren, Z., & Dai, Q. (2016). Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems, 28*(3), 653–664.
9. Deng, M. (2024). Application of big data analysis in Chinese art song market research. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 15*(4), 1–10. https://doi.org/10.58346/JOWUA.2024.I4.001
10. Gijsbrechts, J., Boute, R. N., Van Mieghem, J. A., & Zhang, D. J. (2022). Can deep reinforcement learning improve inventory management? Performance on lost sales, dual-sourcing, and multi-echelon problems. *Manufacturing & Service Operations Management, 24*(3), 1349–1368. https://doi.org/10.1287/msom.2021.0999

11. Krasheninnikova, E., García, J., Maestre, R., & Fernández, F. (2019). Reinforcement learning for pricing strategy optimization in the insurance industry. *Engineering Applications of Artificial Intelligence, 80*, 8–19. https://doi.org/10.1016/j.engappai.2019.01.010
12. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Hassabis, D., et al. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533. https://doi.org/10.1038/nature14236
13. Zhu, Z., Lin, K., Jain, A. K., & Zhou, J. (2023). Transfer learning in deep reinforcement learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(11), 13344–13362. https://doi.org/10.1109/TPAMI.2023.3285087
14. Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018, July). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning* (pp. 1861–1870). PMLR.
15. Peng, X. B., Andrychowicz, M., Zaremba, W., & Abbeel, P. (2018, May). Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 3803–3810). IEEE. https://doi.org/10.1109/ICRA.2018.8460528
16. Deisenroth, M., & 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* (pp. 465–472).
17. Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.
18. Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701. https://doi.org/10.1037/h0037350
19. Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). MIT Press.
20. Baggyalakshmi, N., Harshada, J., & Revathi, R. (2024). Super market billing management system. *International Academic Journal of Science and Engineering, 11*(1), 107–117. https://doi.org/10.9756/IAJSE/V11I1/IAJSE1114
21. Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., & Jordan, M. I. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research, 7*, 2003–2030.
22. Manthila, P., & Ulkilan, A. (2026). AI-augmented dynamic partial reconfiguration for adaptive edge intelligence in FPGA-based embedded systems. *SCCTS Transactions on Reconfigurable Computing, 3*(1), 11–18.
23. Zheng, X., Aragam, B., Ravikumar, P. K., & Xing, E. P. (2018). DAGs with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems, 31*, 9472–9483.
24. Lattimore, F., Lattimore, T., & Reid, M. D. (2016). Causal bandits: Learning good interventions via causal inference. *Advances in Neural Information Processing Systems, 29*, 1181–1189.
25. Bareinboim, E., Forney, A., & Pearl, J. (2015). Bandits with unobserved confounders: A causal approach. *Advances in Neural Information Processing Systems, 28*, 1342–1350.
26. Arvinth, N. (2025). Causal representation learning for adaptive decision-making in data-driven networked systems. *Journal of Scalable Data Engineering and Intelligent Computing*, 33–38.
27. A. Suresh kumar, "Real-Time Embedded Implementation of Reinforcement Learning-Based Motor Drive Controllers", *Journal of Reconfigurable Hardware Architectures and Embedded Systems*, vol. 2, no. 3, pp. 58–67, Sep. 2025.
28. P. Michael, & K. Jackson. (2025). Advancing Scientific Discovery: A High Performance Computing Architectures for AI and Machine Learning. *Journal of Integrated VLSI, Embedded and Computing Technologies* , *2*(2), 18-26.