Proceedings of the 23rd International Conference on Principles of Knowledge Representation and Reasoning
KR meets Machine Learning and Explanation

Neuro-Symbolic Causal Boosting: A Framework for Interpretable Attribution of Business Fluctuations

Yonghe Zhao[1,2], **Yuezhu Wang**[2], **Chao Ma**[1,*], **Huiyan Sun**[2,3,*]
[1]Artificial Intelligence Research Institute, Mashang Consumer Finance Co., Ltd., Chongqing, China
[2]School of Artificial Intelligence, Jilin University, Changchun, Jilin, China
[3]International Center of Future Science, Jilin University, Changchun, Jilin, China
yhzhao26@gmail.com, yuezhuw23@mails.jlu.edu.cn, macmacmac@yeah.net huiyansun@jlu.edu.cn

Abstract

Attributing business fluctuations to actionable drivers is a critical component of decision-making in high-stakes domains. However, prevailing predictive models, predominantly driven by correlations, often yield inconsistent explanations that degrade under distribution shifts or latent confounding. Empirical causal inference to address this limitation remains challenge due to the identifiability gap in purely data-driven discovery and the complexity of encoding domain knowledge into differentiable learning pipelines. To bridge this gap, we propose Neuro-Symbolic Causal Boosting, a unified framework that integrates semantic domain priors with gradient-based causal estimation. First, we introduce the Complete Cause Identification Algorithm (CCIA). Unlike global search methods, CCIA recursively reconstructs the ancestral causal graph of the target variable by employing Kolmogorov-Arnold Networks (KANs) as high-precision filters for low-order independence testing, coupled with a neuro-symbolic adjudication based on Large Language Model to resolve directionality. Subsequently, the identified structure scaffolds the Causal Additive Boosting Network (CABN). Grounded in the theory of structural identification, CABN enforces a reverse topological learning process. It utilizes weighted KANs to sequentially estimate downstream effects and adjust for confounding, thereby isolating invariant causal mechanisms. Empirical evaluation on a real-world telesales dataset and synthetic benchmarks demonstrates that our framework achieves a 57.5% reduction in Out-of-Distribution prediction error compared to strong correlation-based baselines. Additionally, it identifies and quantifies the impact of actionable drivers, providing a structured approach from observational data to trustworthy and interpretable business strategies.

1 Introduction

In high-stakes domains such as finance, marketing, and industrial control, the ability to understand why an event occurs is as critical as the accuracy of predicting if it will occur (Rudin 2019). Decisions in these fields carry significant economic consequences, yet the attribution methods predominantly used are rooted in statistical correlation rather than causal reasoning (Hernán and Robins 2020). This fundamental mismatch creates a unreliable basis for decision-making.

The prevailing machine learning paradigm relies on identifying robust correlations to minimize predictive error. However, feature importance rankings produced by popular models like XGBoost are well-documented to suffer from correlation-induced biases and exhibit significant instability (Lundberg et al. 2020; Lipton 2018). Furthermore, post-hoc explanation methods such as SHAP or LIME, while widely adopted, are often approximations that rest on unrealistic assumptions (Molnar 2020) and can be adversarially manipulated (Slack et al. 2020). These tools explain the model's predictions rather than the phenomenon's causal mechanisms, creating a false sense of understanding that can justify flawed strategies (Glymour, Zhang, and Spirtes 2019). Addressing these limitations necessitates a shift from purely data-driven correlation to knowledge-guided causality (Gong 2020).

However, operationalizing causality in complex business environments faces two profound methodological challenges. The first is the identifiability gap in causal discovery. Reconstructing the causal graph from observational data is an NP-hard problem (Spirtes, Glymour, and Scheines 2000). Existing algorithms confront an inherent dilemma between scalability and reliability (Geiger, Verma, and Pearl 1990). Traditional constraint-based approaches often struggle to navigate the super-exponential search space or remain brittle to statistical errors in finite samples (Li et al. 2019). Conversely, emerging functional-based methods typically achieve identifiability only by imposing strong assumptions about the data distribution (Salehkaleybar et al. 2020), restricting their applicability in real-world scenarios.

The second challenge is the neuro-symbolic gap with causal learning. The formal language of causal inference, rooted in the Structural Causal Model (SCM) (Pearl 2009) or Potential Outcome Framework (Imbens and Rubin 2015), is optimized for estimating the effect of predefined interventions (Bareinboim et al. 2022; Chernozhukov et al. 2018). This is distinct from the machine learning paradigm, which excels at building high-accuracy predictive representations from high-dimensional features (Lecca 2021). Consequently, there is a lack of frameworks that effectively unify symbolic domain knowledge with the expressive power of neural networks to solve attribution tasks (d'Garcez and Lamb 2023; Kıcıman et al. 2023).

To address these bottlenecks, this study introduces a unified framework, Neuro-Symbolic Causal Boosting. Recent trends in neuro-symbolic AI have explored causal model-

ing agents that leverage Large Language Models (LLMs) to deduce graph structures via text-based reasoning (Abdulaal et al. 2024; Vashishtha et al. 2025). Unlike these purely prompt-driven approaches, our framework utilizes LLMs as localized adjudicators for statistical equivalence classes, maintaining a foundation in rigorous conditional independence testing while resolving directionality through semantic priors. We posit that the ambiguity of statistical discovery can be resolved by treating domain knowledge as a structural prior and leveraging the spectral efficiency of novel neural architectures. First, we automate causal discovery using the Complete Cause Identification Algorithm (CCIA). CCIA recursively traces the ancestral lineage of the target variable by employing Kolmogorov-Arnold Networks (KANs) (Liu et al. 2024) to capture non-linear dependencies in low-dimensional conditional sets. Furthermore, we introduce a cascaded neuro-symbolic adjudication module, which integrates LLM-based semantic reasoning with statistical tests of Additive Noise Models (ANM) (Hoyer et al. 2008) to resolve Markov equivalence classes. Subsequently, the discovered symbolic graph serves as the structural scaffold for the Causal Additive Boosting Network (CABN). Unlike traditional boosting algorithms, CABN enforces a reverse topological learning sequence. It integrates graph-based de-confounding with the inherent transparency of KANs to quantify actionable drivers precisely. The contributions of this work are threefold:

- We propose an end-to-end attribution framework that formalizes domain business logic as structural constraints, bridging the gap between observational data and verifiable causal models.
- We introduce CCIA, a recursive discovery algorithm that synergizes the low-dimensional approximation efficiency of KANs with a neuro-symbolic adjudication mechanism, overcoming the identifiability gap in finite-sample regimes where traditional tests fail.
- We develop CABN, a graph-constrained boosting architecture. By enforcing reverse topological learning and performing variable-wise de-confounding, CABN enables precise and quantifiable causal attribution, moving beyond static feature importance.

2 Methodology

2.1 Preliminaries

To rigorously attribute business fluctuations to actionable drivers, we formalize the operational environment using the Structural Causal Model (SCM) framework. Unlike predictive paradigms that map features to outcomes based on conditional expectations, our objective is to model the invariant mechanisms governing the system.

Structural Causal Models We assume the data generating process follows an SCM, defined as a tuple $\mathcal{M} = (\mathcal{X}, \mathcal{N}, \mathcal{F}, P_{\mathcal{N}})$. Here, $\mathcal{X} = \{X^1, \dots, X^d\}$ represents the set of endogenous variables (e.g., observed metrics like *Talk_Hrs*, *Sales*), and $\mathcal{N} = \{N^1, \dots, N^d\}$ represents the set of independent exogenous noise variables distributed according to $P_{\mathcal{N}}$. The causal relationships are governed by a set of structural functions $\mathcal{F} = \{f_1, \dots, f_d\}$, where each variable X^i is determined by its causal parents $Pa^i \subseteq \mathcal{X} \setminus \{X^i\}$ and its specific noise term N^i:

$$X^i := f_i(Pa^i, N^i), \quad i = 1, \dots, d. \tag{1}$$

These structural equations induce a Directed Acyclic Graph (DAG), $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where vertices $\mathcal{V}$ correspond to variables in $\mathcal{X}$, and directed edges represent the input-output dependencies in $\mathcal{F}$.

Interventional Attribution The primary challenge in business analysis is to estimate the effect of interventions. We denote the post-intervention distribution of the target variable $Y \in \mathcal{X}$ under a hard intervention $do(X^i = x)$ as $P(Y \mid do(X^i = x))$. Reliable attribution requires learning the true causal graph $\mathcal{G}^*$ to identify the correct adjustment set $\mathbf{Z} \subseteq \mathcal{X}$ that satisfies the backdoor criterion. Specifically, we aim to estimate the Marginal Treatment Effect Function (MTEF), which describes how changes in a driver X^i propagate to Y while keeping other mechanisms invariant.

Neuro-Symbolic Constraints In finite-sample regimes, identifying $\mathcal{G}^*$ solely from observational data $\mathcal{D}$ is often impossible due to Markov equivalence classes. We posit that valid causal graphs must satisfy a set of semantic constraints $\mathcal{K}$ derived from domain knowledge. The learning problem is thus to identify the optimal graph $\hat{\mathcal{G}}$ that maximizes the data likelihood while satisfying $\mathcal{K}$:

$$\hat{\mathcal{G}} = \arg\max_{\mathcal{G} \in \mathbb{G}} P(\mathcal{D} \mid \mathcal{G}) \quad \text{s.t.} \quad \mathcal{G} \models \mathcal{K}, \tag{2}$$

where $\mathcal{K}$ serves as a structural prior reducing the search space $\mathbb{G}$.

2.2 The Neuro-Symbolic Causal Boosting Framework

Addressing the interpretable attribution of business fluctuations, we introduce the Neuro-Symbolic Causal Boosting framework. As illustrated in Fig. 1, the framework operates as a closed-loop dual-stage system that explicitly bridges the gap between semantic domain priors and gradient-based learning.

The first stage focuses on structure learning via the CCIA. This module reconstructs the causal graph $\mathcal{G}$ by integrating statistical independence testing with a semantic consistency constraint. A LLM acts as a reasoning agent to enforce the constraints $\mathcal{K}$, adjudicating causal directions where statistical signals are ambiguous. This ensures the discovered graph aligns with operational logic (e.g., temporal precedence).

The second stage executes parameter estimation and attribution via the CABN. Utilizing the discovered graph $\mathcal{G}$ as a structural scaffold, this module performs graph-based de-confounding. It estimates the causal mechanisms using a weighted KAN architecture embedded within a reverse topological boosting framework. This design ensures that

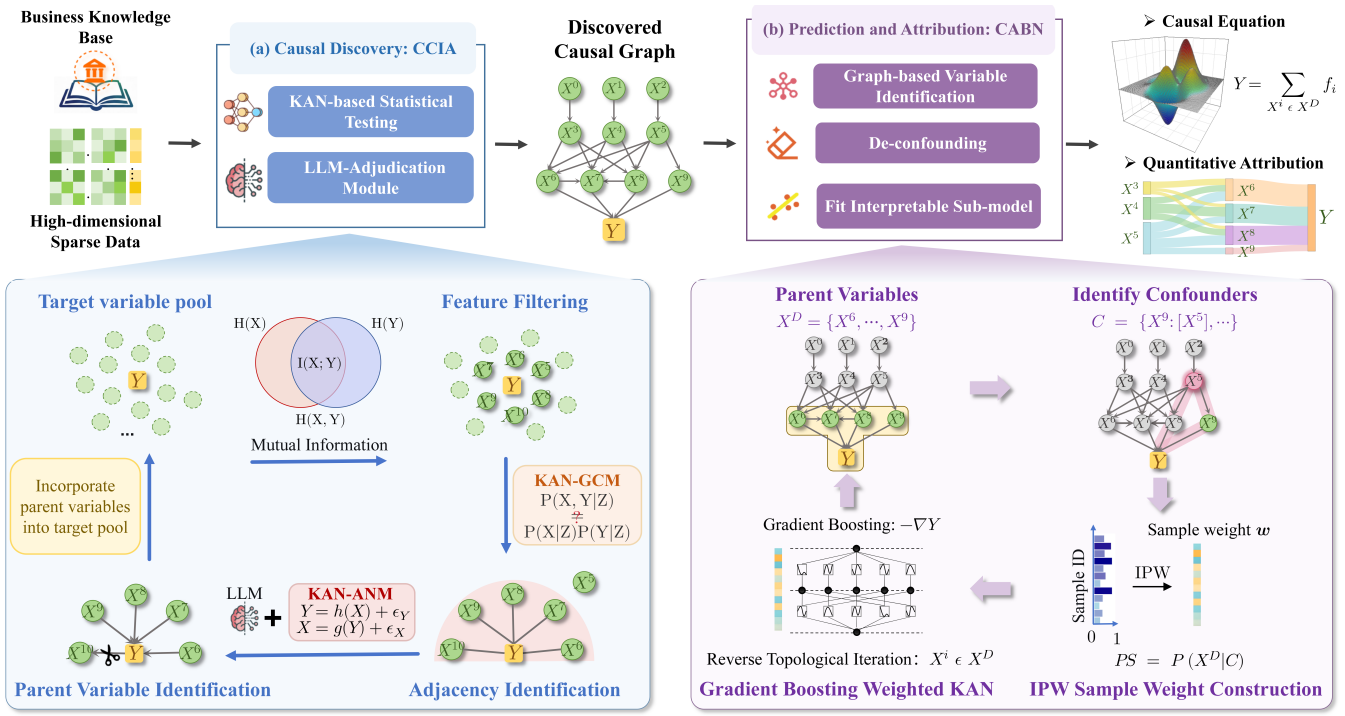


Figure 1: A neuro-symbolic framework for interpretable attribution of business fluctuations. (a) Causal Discovery: The CCIA algorithm processes observational data and business knowledge base to produce an ancestral causal graph. (b) Prediction and Attribution: The CABN model uses this graph as a structural scaffold for de-confounding and interpretable sub-modeling. The final outputs are the causal graph, a robust predictive model and quantified causal contributions.

the final model captures the interventional distribution necessary for robust attribution, rather than merely fitting the observational distribution.

2.3 Complete Cause Identification Algorithm

The first stage of our framework aims to recover the ancestral causal structure $\mathcal{G}^*$ surrounding the target variable, illustrated in Fig. 1(a). Unlike traditional algorithms (e.g., PC, GES) that attempt to reconstruct the entire graph, an computationally expensive and often unnecessary feat for attribution tasks, we propose the CCIA with a target-centric recursive ancestral identification strategy. This approach focuses solely on recursively identifying the causal ancestors of the target, ensuring efficiency and relevance.

Recursive Ancestral Identification via KAN-GCM The identification process of CCIA operates recursively. Starting with the target variable Y, we maintain a candidate set of ancestors to be identified. In each iteration, we identify the direct parents of the current variable and add them to the candidate set for the next iteration. Crucially, to accelerate convergence, variables identified as descendants are explicitly excluded from future search spaces, leveraging the acyclicity of the causal graph to prune the search tree dynamically.

Low-Order Conditional Independence Testing. To determine the adjacency between a variable X^i and a candidate parent X^j, we employ the Generalized Covariance Measure (GCM) (Shah and Peters 2020) to test the null hypothesis of conditional independence $X^i \perp X^j | Z$. A critical challenge in GCM is selecting the conditioning set Z. Following the theoretical insight (Yan and Zhou 2020), which asserts that "if a conditioning set of more than three variables does not render two variables independent, then it is highly probable that these two variables are directly related," we restrict the cardinality of the conditioning set to $|Z| \leq 3$. This constraint not only significantly reduces the combinatorial search space but also aligns perfectly with the strengths of our underlying estimator.

Why KAN is Essential for Independence Testing. The restriction to low-dimensional conditioning sets ($|Z| \leq 3$) creates a unique modeling requirement: the regressor must capture potentially complex, high-frequency non-linear relationships (e.g., saturation points, thresholds) using limited input features. Standard Multi-Layer Perceptrons often struggle in this regime due to *spectral bias*, the tendency to learn low-frequency functions first, and may require excessive depth to approximate simple non-monotonic functions, leading to overfitting or optimization difficulties.

In contrast, KANs are mathematically ideal for this low-dimensional setting. By parameterizing activation functions as learnable B-splines on edges rather than fixed functions on nodes, KANs eliminate the need for deep stacking to achieve non-linearity. This spline-based architecture allows KANs to locally adjust to data irregularities with high parameter efficiency. In our framework, we model the conditional expectations $E[X|Z]$ using KANs to compute the

residuals:

$$\epsilon^i = X^i - \phi_{KAN}(Z), \quad \epsilon^j = X^j - \psi_{KAN}(Z) \quad (3)$$

The edge $X^i - X^j$ is removed only if the correlation between residuals $\rho(\epsilon^i, \epsilon^j)$ vanishes. The superior expressiveness of KANs in low dimensions acts as a rigorous non-linear filter, ensuring that non-vanishing residuals represent true causal connections rather than model misspecification.

Cascaded Neuro-Symbolic Adjudication Resolving the directionality of edges (e.g., distinguishing $X^i \to X^j$ vs. $X^j \to X^i$) within the identified skeleton is the most challenging step. We propose a cascaded neuro-symbolic adjudication mechanism that effectively operationalizes the Bayesian intuition of combining prior knowledge with data likelihood.

We formulate the edge orientation as a hierarchical decision process:

- **Semantic Prior (LLM).** We first query a LLM to evaluate the semantic plausibility of the causal link $P(\mathcal{G}|\mathcal{K})$. The LLM analyzes metadata (e.g., variable names, descriptions) based on logical criteria such as temporal precedence and operational logic. If the LLM provides a high-confidence judgment (consistency score $> \tau_{high}$), we accept this direction as the structural prior.
- **Statistical Likelihood (KAN-ANM).** If the semantic signal is ambiguous (low confidence), we fallback to data-driven discovery using ANM. We fit KANs in both directions ($Y = f(X) + \epsilon$ vs. $X = g(Y) + \epsilon$) and select the direction that yields independent residuals.

This cascaded approach ensures that the final graph $\hat{\mathcal{G}}$ respects domain common sense while remaining statistically grounded where intuition falls short.

2.4 Causal Additive Boosting Network

Having established the structural prior $\mathcal{G}$, the second stage transitions from qualitative discovery to quantitative estimation. We propose the CABN illustrated in Fig. 1(b) that eschews traditional greedy boosting in favor of a graph-constrained sequential causal modoling process. As formalized in Algorithm 1, this approach is grounded in the necessity of recovering the invariant structural functions f_j in the additive model $Y = \sum f_j(X^j) + \epsilon$, effectively simulating a sequence of randomized controlled trials from observational data.

Theoretical Foundation: Reverse Topological Boosting Standard regression or boosting treats features symmetrically, which is problematic in causal chains (e.g., $X \to M \to Y$). Regressing Y on upstream drivers X before (or simultaneously with) downstream mediators M risks capturing the total effect rather than the direct structural mechanism, as X implicitly contains information about M. To disentangle these effects, we introduce the logic of Reverse Topological Boosting, which serves as the theoretical basis for the sorting phase of CABN.

Proposition 1 (Reverse Topological Boosting)**.** *Assume the data generating process follows an additive Structural Causal Model. Under the assumptions of Causal Sufficiency and Faithfulness, the structural functions $\{f_j\}$ can be consistently identified via sequential residual learning if and only if the estimation order follows the Reverse Topological Sort (i.e., peeling off descendants before ancestors).*

Proof. Consider a chain $X^u \to X^d \to Y$.

- **Downstream First (Correct):** If we estimate X^d first, we condition on its parents. The residual $R = Y - \hat{f}_d(X^d)$ removes the mechanism of X^d, leaving $R \approx f_u(X^u) + \epsilon$. This effectively "blocks" the path from X^u through X^d.
- **Upstream First (Incorrect):** If we estimate X^u first, the projection captures the total effect of X^u on Y (both direct and indirect). Subtracting this contaminates the subsequent estimation of X^d, as the variance attributed to X^d has been partially explained away by X^u.

Thus, CABN enforces a strict learning sequence $\mathcal{S}$ where proximal causes are peeled off before distal causes.

Algorithm 1 Causal Additive Boosting Network (CABN)

Input: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, Causal Graph $\mathcal{G}$
Output: Additive Structural Model $F(x) = \sum \hat{f}_j(x^j)$

Phase 1: Hierarchy Sorting
Compute topological depth of all ancestors of Y via BFS on $\mathcal{G}$.
Sort parents into sequence $\mathcal{S} = (X^{(1)}, \ldots, X^{(k)})$ by ascending depth (Proximal $\to$ Distal).
Phase 2: Sequential Causal Peeling
Initialize residuals $r_i \leftarrow y_i$ for all $i = 1 \ldots N$.
Initialize model ensemble $\mathcal{M} \leftarrow \emptyset$.
for each driver $X^{(t)}$ in sequence $\mathcal{S}$ **do**
  // Step A: Variable-wise De-confounding
  Identify confounder set $Z^{(t)}$ for $X^{(t)}$ from $\mathcal{G}$.
  if $Z^{(t)} \neq \emptyset$ **then**
    Estimate $P(X^{(t)}|Z^{(t)})$ using **KAN Density Est**.
    Compute IPW weights w_i (clipped for stability).
  else
    Set $w_i \leftarrow 1$.
  end if
  // Step B: Weighted Structural Fitting
  Fit KAN component $\hat{f}_t$ to residuals r minimizing:
    $\mathcal{L} = \sum_{i=1}^N w_i \cdot (r_i - \hat{f}_t(x_i^{(t)}))^2$
  // Step C: Peeling Update
  Update residuals: $r_i \leftarrow r_i - \hat{f}_t(x_i^{(t)})$
  Add $\hat{f}_t$ to ensemble $\mathcal{M}$.
end for
return Final Model $F(x) = \sum_{\hat{f} \in \mathcal{M}} \hat{f}(x)$

Algorithm Implementation Guided by this proposition, CABN executes a structured learning loop detailed in Algorithm 1. The process consists of three critical phases:

Hierarchy Sorting (Lines 1-3). We first compute the topological depth of each parent relative to the target using a Breadth-First Search (BFS) on $\mathcal{G}$. Variables are sorted into a sequence $\mathcal{S}$ by increasing depth. This ensures that when the algorithm addresses upstream strategic drivers

(e.g., *Work_Hrs*), the effects of downstream operational metrics have already been explicitly removed.

Variable-wise De-confounding (Lines 7-15). For each driver $X^{(t)}$ in the sequence, we identify its specific confounder set $Z^{(t)}$ from the graph. A crucial innovation here is the use of KANs for density estimation. As shown in Line 11, we estimate the Generalized Propensity Score $P(X^{(t)}|Z^{(t)})$ (Hirano and Imbens 2004) using a KAN. The spline-based flexibility of KANs ensures accurate probability estimation even for continuous variables with irregular distributions, producing stable Inverse Probability Weights (IPW) w_i that create a pseudo-population where $X^{(t)}$ is independent of its confounders.

Weighted Structural Fitting (Lines 16-22). Finally, we isolate the causal mechanism by fitting a weighted KAN to the current residuals r_i. The objective function $\sum w_i(r_i - h(x_i^{(t)}))^2$ forces the network to learn the relationship under the de-confounded distribution. The fitted function $\hat{f}_t$ is then stored as the structural component, and its contribution is peeled from the residuals for the next iteration.

3 Experiments

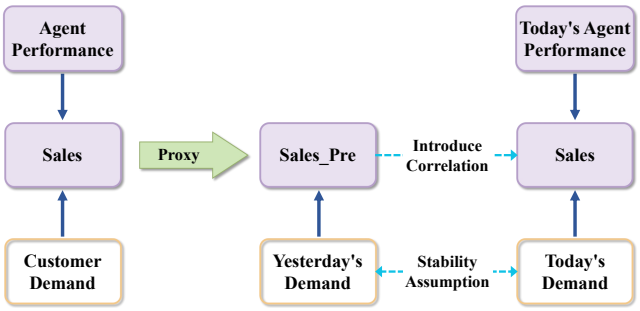
To rigorously validate the proposed framework, we conducted a comprehensive evaluation on both a large-scale real-world financial dataset and controlled synthetic benchmarks. The experiments are designed to assess three critical dimensions: (1) the structural accuracy of the neuro-symbolic causal discovery (CCIA), (2) the predictive robustness of the causal model (CABN) under distribution shifts, and (3) the interpretability of the generated attributions. All source codes and synthetic datasets are available at the anonymous repository[1].

3.1 Data Sources and Processing Strategy

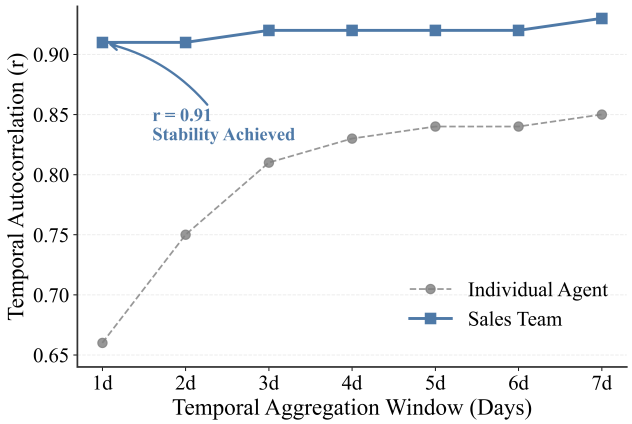
Real-world Telesales Dataset and Causal Mechanisms The primary dataset was sourced from a financial institution's telesales operation, spanning March to June 2025. The initial corpus comprised 121,635 raw samples. To ensure a consistent operational environment, we restricted the scope to the primary product line within a single sales region.

A fundamental challenge in this domain is modeling the unobserved drivers of fluctuation. As illustrated in the causal mechanism diagram (Fig. 2(a)), domain consensus posits that *Sales* volume is a joint function of *Agent Performance* (observable variables like *Talk_Hrs*) and *Customer Demand* (a latent confounder). Ignoring this latent variable would introduce significant confounding bias into any attribution model. To address this, we rely on the *Demand Stability Assumption*: we assume that the distribution of underlying customer demand remains relatively stable over short temporal windows, allowing the previous sales (*Sales_Pre*) to serve as a valid proxy for current demand.

Empirical Validation and Aggregation We rigorously tested the validity of this stability assumption via a multi-scale autocorrelation analysis. As shown in Fig. 2(b),



(a) Causal Mechanism and Proxy Construction



(b) Empirical Validation of Aggregation Stability

Figure 2: Latent Confounder and Knowledge Constraints. (a) The business assumption modeling unobserved customer demand as a latent confounder, addressed by introducing *Sales_Pre* as a proxy variable. (b) Statistical justification for the aggregation strategy. The saturation curve reveals that team-level aggregation achieves high temporal stability ($r = 0.91$) even at a 1-day window.

the data exhibits distinct behaviors at different granularities. At the individual agent level, the temporal autocorrelation is moderate ($r = 0.66$), indicating that individual performance variance overshadows the demand signal. Conversely, when aggregated to the team level, the stability significantly increases ($r = 0.91$). This empirical evidence confirms that team-level aggregation effectively filters out stochastic noise, rendering *Sales_Pre* a reliable proxy for the latent demand.

Consequently, we processed the data into daily team-level observations. After removing statistical outliers via the Interquartile Range (IQR) method, the final dataset comprises 7,592 samples. The feature space consists of 14 variables, categorized into Operational Context, Process Metrics, and Target. The detailed definitions for all variables are provided in Table 1.

Synthetic Benchmarks via Non-linear SEM Since the ground truth causal structure of real-world data is unknown, we generated synthetic datasets to benchmark the discovery accuracy of CCIA. To ensure the evaluation reflects the

[1]https://github.com/ZYH30/NeuSymCauFram

Category	Variable	Description
Target	*Sales*	Daily sales amount (Target)
	Sales_Pre	Previous day's sales (Demand Proxy)
Context	*Staffs*	Number of agents on duty
	Work_Hrs	Daily average work hours
	Initiative	Marketing project identifier
Process	*Calls_Num*	Agent's daily average calls
	Calls_Ans	Total answered calls
	Calls_Per	Daily average connected calls
	Calls_15s	Daily average effective calls ($> 15s$)
	Talk_Hrs	Total talk duration (Hours)
	Talk_Per	Agent's daily average talk duration
	Talk_15s	Daily average talk duration for effective calls ($> 15s$)
	ATT	Average Talk Time per call
	ATT_15s	Average Talk Time per effective call ($> 15s$)
	Attendance	Daily agent attendance count

Table 1: Variable definitions of the team-level daily financial telesales dataset for causal attribution of sales fluctuations. Variables are grouped into target, context, and process categories, with *Sales_Pre* serving as an empirically verified proxy for latent customer demand.

complexity of the actual business environment, we adopted a semi-synthetic strategy: we utilized the causal structure discovered by CCIA from the real-world telesales data (verified by domain experts in Fig. 3) as the ground truth topology, denoted as $\mathcal{G}_{real}$. Based on this structure, we employed a non-linear Structural Equation Model (SEM) to simulate the saturation effects inherent in business dynamics. The data generating process for each variable X^j is defined as:

$$X^j := \sum_{k \in Pa^j} \tanh(w_{kj} X^k) + \epsilon^j, \quad \epsilon^j \sim \mathcal{N}(0, 1), \qquad (4)$$

where edge weights w_{kj} are sampled from $\mathcal{U}(0.5, 2.0)$ to ensure strong causal signals. We generated $N = 10,000$ samples under three scenarios to test robustness:

- **Scenario A (Standard):** Data generated strictly following the topology of the real-world graph $\mathcal{G}_{real}$, serving as the idealized baseline to verify if the algorithm can recover the self-discovered structure under non-linear assumptions.
- **Scenario B (Structural Noise):** To test resistance to spurious correlations, we randomly permuted 20% of the edges while respecting the topological hierarchy.
- **Scenario C (Sparse Signals):** To test sensitivity, we randomly removed 20% of the edges, simulating missing information.

3.2 Baselines and Evaluation

Baselines for Causal Discovery We benchmark CCIA against five established algorithms representing different discovery paradigms:

- **Constraint-based:** The PC algorithm (standard baseline) and FCI (specifically designed to handle latent confounders) (Spirtes, Glymour, and Scheines 2000).
- **Score-based:** GES (Greedy Equivalence Search) (Chickering 2002) and GRaSP (Greedy Relaxed Acyclic Search Pruning) (Lam, Andrews, and Ramsey 2022), a state-of-the-art permutation-based approach known for scalability.
- **Functional-based:** DirectLiNGAM, which assumes linear non-Gaussianity, serving to test the impact of non-linear assumption violations (Shimizu et al. 2011).

Note that continuous optimization methods, such as NOTEARS(Zheng et al. 2018) and DAG-GNN(Yu et al. 2019), were excluded from the primary benchmarking. Preliminary tests indicated that their reliance on global acyclicity penalties often led to unstable local optima in our non-linear, finite-sample regime. Furthermore, the linear assumption of standard NOTEARS resulted in weight decay toward zero when facing the non-linear saturation effects characteristic of our data generation process.

Baselines for OOD Prediction For the downstream regression task, we compare CABN against strong correlational models: LGBM, XGBoost, CatBoost, and a standard Neural Network (NN(MLP)). To ensure a fair comparison, all baselines were trained using the set of causal parents identified by CCIA, isolating the performance gain to the modeling architecture rather than feature selection. All hyperparameters were tuned via Bayesian optimization.

LLM Configuration For the neuro-symbolic reasoning module in CCIA, we utilized a locally deployed and distilled variant of the Qwen-235B Mixture-of-Experts model. Local deployment was chosen to ensure data privacy and inference stability. The detailed prompt is provided in Appendix A. To guarantee reproducibility, we fixed the sampling *temperature* at 0.0 and *top_p* at 0.1. Furthermore, we implemented a consistency voting mechanism to mitigate generative hallucinations: causal directions were confirmed only if they received a majority vote with a confidence threshold > 0.6; otherwise, the edge orientation reverted to pure statistical testing. All experiments were conducted on a cluster equipped with NVIDIA A800 GPUs.

OOD Generalization Setup Unlike standard supervised learning evaluations that assume independent and identically distributed (i.i.d.) data, our primary goal is to assess robustness to distribution shifts. We implemented a rigorous Target Shift splitting strategy. The real-world dataset was partitioned based on the quantile of the target variable *Sales*. Samples above the 80^{th} percentile constituted the Out-of-Distribution (OOD) test set, strictly isolating high-performance regimes from the training phase. This setup mimics the challenge of generalizing from routine operations to peak business conditions, where correlation-based models typically fail.

Metrics For structural learning, we report Structural Hamming Distance (SHD), Precision, Recall, and F1-Score. For prediction task, we report the Mean Squared Error (MSE) on the OOD test set. To quantify stability, we report the mean and standard deviation across 10 independent runs.

3.3 Evaluation of Causal Discovery

To answer the critical question of whether the proposed neuro-symbolic approach effectively bridges the identifiability gap, we first analyze the discovery performance on controlled synthetic benchmarks (Table 2) and then interpret the structure recovered from the real-world dataset (Fig. 3).

Benchmarking on Synthetic Data: The Precision Advantage Table 2 presents the performance metrics across three scenarios. The results reveal a consistent superiority of CCIA, particularly in its ability to reject spurious correlations. Standard score-based methods like GRaSP, while achieving high Recall (1.0 in Scenarios A and B), suffer significantly in Precision ($0.51 - 0.60$). This indicates a tendency to over-connect, mistaking non-linear spurious associations for causal links when faced with saturation effects. Similarly, DirectLiNGAM's moderate performance ($F1 \approx 0.55$) suggests that relying solely on statistical properties is insufficient for complex dynamics.

Method	SHD (↓)	Prec (↑)	Rec (↑)	F1 (↑)
Scenario A: Standard				
PC	20	0.46	0.72	0.57
GES	28	0.37	0.78	0.50
FCI	33	0.30	0.61	0.40
GRaSP	12	0.60	**1.00**	0.75
DirectLiNGAM	18	0.50	0.61	0.55
CCIA (Ours)	**5**	**0.93**	0.78	**0.85**
Scenario B: Permuted				
PC	17	0.52	0.89	0.65
GES	21	0.46	0.89	0.60
FCI	24	0.41	0.78	0.54
GRaSP	17	0.51	**1.00**	0.68
DirectLiNGAM	17	0.52	0.72	0.60
CCIA (Ours)	**7**	**1.00**	0.61	**0.76**
Scenario C: Sparse				
PC	18	0.45	**0.93**	0.61
GES	20	0.42	0.87	0.57
FCI	18	0.43	0.67	0.53
GRaSP	16	0.48	**1.00**	0.65
DirectLiNGAM	16	0.48	0.67	0.56
CCIA (Ours)	**5**	**0.81**	0.87	**0.84**

Table 2: Robustness evaluation on synthetic data. CCIA demonstrates superior structural fidelity, achieving the highest F1-Score across all scenarios. The best results are **underlined and bolded**.

In contrast, CCIA maintains high Precision (> 0.81) across all scenarios. The most striking validation of the neuro-symbolic hypothesis is observed in Scenario B (Structural Noise), where edges were randomly permuted to simulate misleading correlations. Here, purely data-driven baselines falter (Precision ≈ 0.5), unable to distinguish true signals from structural noise. However CCIA achieves a Precision of 1.0, contrasted with the sharp decline observed in purely data-driven baselines, provides a quantified validation of the semantic module's capacity to effectively filter out spurious correlations by enforcing logical priors. This result theoretically validates our framework as a form of Semantic Prior: the LLM acts as a high-pass filter orthogonal to the statistical signal, effectively pruning edges that violate logical priors even when the data suggests a correlation. In high-stakes business attribution, this property is crucial, as a False Positive (identifying a non-causal driver) is significantly costlier than a False Negative.

Real-world Causal Graph Discovery Applying CCIA to the telesales dataset yielded the ancestral graph $\mathcal{G}_{real}$ depicted in Fig. 3. The discovery process demonstrated the necessity of the hybrid approach: while KAN-based independence tests determined the majority of adjacencies, the directionality of ambiguous edges (where statistical residuals were symmetric, $p > 0.05$) was resolved by the Semantic Adjudication module.

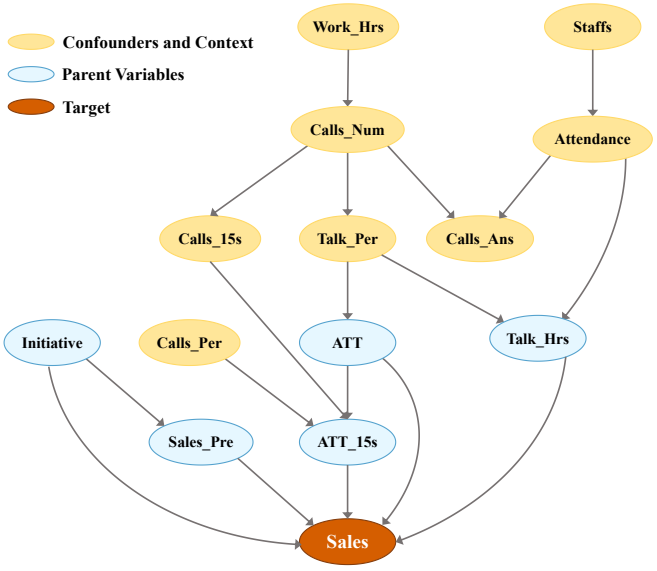
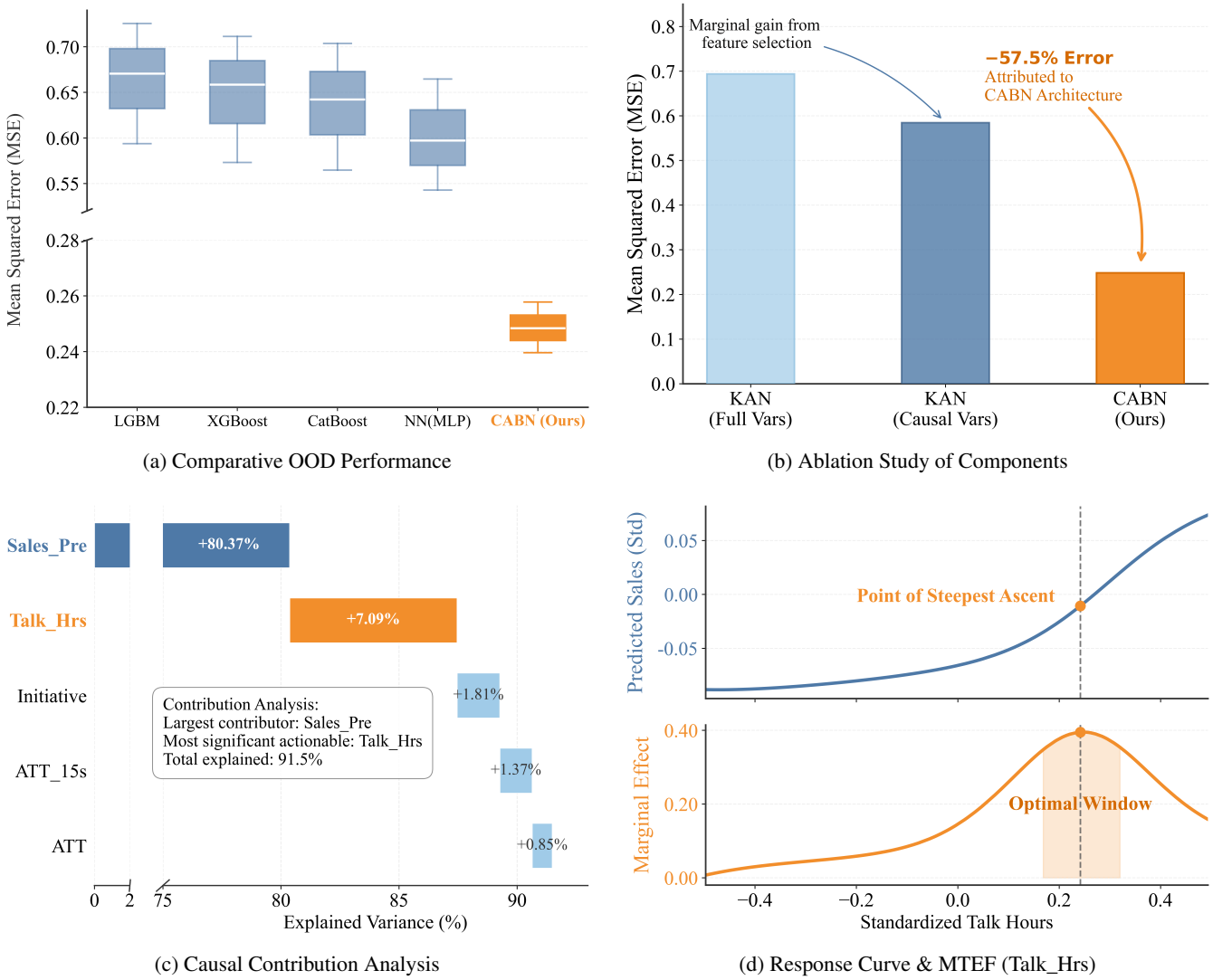


Figure 3: The final causal graph learned automatically by CCIA, which serves as the structural prior for modeling.

The resulting graph reveals a coherent and multi-layered causal hierarchy that aligns with domain intuition. As shown in Fig. 3, the structure traces a clear path from Operational Context (e.g., *Staffs*, *Work_Hrs*) through intermediate Process Metrics (e.g., *Calls_Num*, *Talk_Per*) to the final Target (*Sales*). Crucially, the algorithm correctly identified *Sales_Pre* as a parent of *Sales*, validating its role as a proxy for latent demand. To further verify the fidelity of this structure, we conducted a workshop with senior business ana-



(a) Comparative OOD Performance

(b) Ablation Study of Components

(c) Causal Contribution Analysis

(d) Response Curve & MTEF (Talk_Hrs)

Figure 4: Superior OOD Performance and Quantitative Attribution. (a) CABN (red) achieves significantly lower mean error and higher stability than all baselines on the OOD test set. (b) The marginal improvement from *KAN (Full Vars)* to *KAN (Causal Vars)* shows selection is insufficient; the 57.5% error reduction by CABN confirms the gain is driven by the causal architecture. (c) *Talk_Hrs* is identified as the most significant actionable driver, accounting for 7.09% of variance. (d) The interpretable KAN sub-model reveals a non-linear, sigmoid-like causal relationship and its marginal effect.

lysts. The expert review confirmed the logical plausibility of the identified paths, particularly the temporal precedence where *Talk_Hrs* drives *Sales*, rather than the reverse, providing qualitative validation for the graph's reliability. This verified graph serves as the structural scaffold for the subsequent attribution analysis.

3.4 OOD Robustness and Attribution Analysis

With the causal structure established, we evaluate the downstream performance of the CABN model. This section addresses two pivotal questions: Does respecting causal structure yield superior generalization under distribution shifts? And can the framework provide transparent and actionable insights beyond static feature importance?

Superior Generalization via Invariant Mechanism Learning The comparative results on the OOD Test Set (Fig. 4(a)) demonstrate the critical limitations of traditional correlation-based architectures. Despite being trained with the correct predictors identified by CCIA, strong baselines including LGBM, XGBoost, and CatBoost fail to generalize to the high-sales regime, exhibiting high error (MSE ranging from 0.60 to 0.67) and significant instability (large variance across runs). This failure stems from their reliance on spurious correlations that, while robust in the training distribution, collapse when the distribution shifts.

Conversely, CABN achieves a substantial reduction in error, recording an MSE of 0.25, a 57.5% improvement over the strongest baseline. Furthermore, the minimal stan-

dard deviation (0.03) confirms the stability of the estimate. This performance improvement substantiates our core hypothesis: by enforcing graph-based constraints and de-confounding, CABN learns the invariant causal mechanisms ($P(Y|do(X))$) rather than fragile conditional expectations ($P(Y|X)$). This invariance is the mathematical prerequisite for robustness in volatile business environments.

Ablation Study: Architecture vs. Feature Selection To isolate the source of this robustness, we deconstruct the framework in Fig. 4(b). A naive hypothesis might suggest that the gain comes solely from selecting the correct causal features. However, the *KAN (Causal Vars)* model, which applies a standard KAN to the causal variables without structural constraints, yields only a marginal improvement (MSE 0.58) over the full-variable model (MSE 0.69).

The critical finding is that the transition from *KAN (Causal Vars)* to CABN drives the majority of the performance gain. This confirms that feature selection alone is insufficient. The robustness is intrinsic to the CABN architecture itself, specifically, the Reverse Topological Boosting strategy and IPW De-confounding. These components ensure that the model sequentially peels off downstream effects and adjusts for confounding bias, effectively simulating a randomized controlled trial from observational data.

From Attribution to Action Beyond prediction, CABN provides a transparent window into the data generating process. The causal contribution analysis (Fig. 4(c)) quantifies the variance explained by each driver. *Sales_Pre* accounts for 80.37% of the variance, confirming that latent demand is indeed the dominant factor in sales fluctuations. Among actionable variables, *Talk_Hrs* emerges as the most significant driver (7.09%).

Crucially, our framework transcends static importance scores by deriving the MTEF. Fig. 4(d) visualizes the learned mechanism for *Talk_Hrs*. The top panel (Response Curve) reveals a non-linear sigmoid-like relationship, capturing the saturation effect inherent in human performance. The bottom panel (MTEF), derived analytically as the derivative of the KAN spline, explicitly plots the marginal return on investment. The MTEF curve peaks at a specific interval, identifying the operational Optimal Window. This insight is prescriptive: it signals to managers that extending talk time beyond this window yields diminishing returns. By quantifying where the marginal effect turns negative or negligible, the framework provides a precise verifiable boundary for operational optimization, an actionable capability that black-box models like XGBoost cannot offer.

4 Conclusion

Reliable decision-making in high-stakes domains demands models that incorporate causal reasoning, not just predict. This study establishes a unified framework to transcend the limitations of correlation-based learning by integrating automated neuro-symbolic causal discovery with structurally interpretable modeling. By synergizing the functional approximation power of KANs with the semantic reasoning of LLMs, we effectively overcome the identifiability gap in causal discovery, translating observational data into verifiable symbolic graphs.

Our empirical results, particularly the 57.5% reduction in OOD error compared to strong baselines, provide compelling evidence that respecting the causal structure is not a hindrance to predictive performance but a prerequisite for robustness. The framework's ability to derive marginal treatment effect further bridges the gap between machine learning and decision science, offering actionable insights, such as the optimal operational window for talk duration, rather than scalar importance scores.

However, the current framework of de-confounding relies on the assumption of causal sufficiency; while we utilize proxy variables to mitigate latent factors, the potential influence of unobserved root confounders remains a theoretical limitation that may affect the structural identification. Additionally, while the additive structure of CABN offers transparency, it may underperform in systems dominated by complex high-order interactions. Given that our empirical evidence is currently grounded in a specific telesales context, future research will focus on establishing the cross-domain generalizability of the framework, such as in healthcare or industrial control, while integrating more expressive temporal logics to handle hidden confounders without explicit proxies, further expanding the applicability of knowledge-guided AI.

A Semantic Adjudication Prompts

A.1 System Prompt Template

System Prompt

Role: You are a semantic adjudicator and domain expert in consumer credit telesales. Your task is to determine the most plausible direction of the causal edge when given two variables X and Y
Semantic Consistency Criteria: Evaluate the relationship based on the following dimensions:

- **Temporal Precedence:** Does the occurrence of the cause strictly precede the effect in time? (e.g., *Talk_Hrs* must happen before *Sales*).
- **Operational Logic (Mechanism):** Does business common sense support a direct mechanism? (e.g., $Avg = Total/Count$ implies strict computational causality).
- **Interventionism:** If we strictly intervene on X (do(X)), will Y change? Conversely, does changing Y necessarily change X?
- **Confounding Check:** Is the relationship likely spurious due to a common unobserved confounder?

Output Requirements: Strict JSON format containing "judgement" (A or B) and "reason".

A.2 Input Task Example

User Query

Context: High-stakes telesales scenario in the consumer credit industry. **Pair Description:**

- X (*Talk_Hrs*): Total talk duration of agents (measured in hours).
- Y (*Sales*): Daily total sales amount (business outcome).

Adjudication Task: Analyze the semantic consistency and determine the causal direction (A: $X \to Y$, or B: $Y \to X$).

LLM Output (Semantic Prior)

{ "judgement": "A", "reason": "Based on Temporal Precedence and Operational Logic: The agent's action of talking (X) is a necessary operational precondition for generating sales (Y). Temporally, the effort (duration) accumulates before the sales result is finalized. Logically, manipulating talk time (e.g., extending shifts) can causally increase sales capacity, whereas artificially fixing the sales number cannot physically alter the past talk duration. Thus, the direction $X \to Y$ is semantically consistent." }

Acknowledgments

The authors thank Mashang Consumer Finance Co., Ltd. for providing the data and computational hardware support. This work was supported by the Natural Science Foundation of Jilin Province (Grant No. 20240101025JJ) and the National Natural Science Foundation of China (Grant No. 62372210).

AI Declaration

The authors used AI-assisted tools for language polishing only. No generative AI tools were used for literature review, data analysis, interpretation, or drafting of the scientific content. The authors take full responsibility for the content of the publication.

References

Abdulaal, A.; Hadjivasiliou, A.; Montaña Brown, N.; He, T.; Ijishakin, A.; Drobnjak, I.; Castro, D. C.; and Alexander, D. C. 2024. Causal modelling agents: Causal graph discovery through synergising metadata- and data-driven reasoning. In *The Twelfth International Conference on Learning Representations*.

Bareinboim, E.; Correa, J. D.; Ibeling, D.; and Icard, T. 2022. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*. 507–556.

Chernozhukov, V.; Chetverikov, D.; Demirer, M.; Duflo, E.; Hansen, C.; Newey, W.; and Robins, J. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21(1):C1–C68.

Chickering, D. M. 2002. Optimal structure identification with greedy search. *Journal of Machine Learning Research* 3(Nov):507–554.

d'Garcez, A. S., and Lamb, L. C. 2023. Neurosymbolic ai: The 3rd wave. *Artificial Intelligence Review* 56(11):12387–12422.

Geiger, D.; Verma, T.; and Pearl, J. 1990. d-separation: From theorems to algorithms. In *Machine Intelligence and Pattern Recognition*, volume 10. Amsterdam: North-Holland. 139–148.

Glymour, C.; Zhang, K.; and Spirtes, P. 2019. Review of causal discovery methods based on graphical models. In *Frontiers in genetics*, volume 10, 524. Frontiers Media SA.

Gong, M. 2020. Bridging causality and learning: How do they benefit from each other? In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 5150–5153.

Hernán, M. A., and Robins, J. M. 2020. *Causal Inference: What If*. Boca Raton, FL: CRC Press.

Hirano, K., and Imbens, G. W. 2004. The propensity score with continuous treatments. In *Applied Bayesian modeling and causal inference from incomplete-data perspectives*. Wiley. 73–84.

Hoyer, P.; Janzing, D.; Mooij, J. M.; Peters, J.; and Schölkopf, B. 2008. Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems* 21.

Imbens, G. W., and Rubin, D. B. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences*. New York, NY: Cambridge University Press.

Kıcıman, E.; Ness, R.; Sharma, A.; and Tan, C. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*.

Lam, W. M.; Andrews, B.; and Ramsey, J. 2022. Greedy relaxed acyclic sketch pruning. In *Uncertainty in Artificial Intelligence*, 1096–1106. PMLR.

Lecca, P. 2021. Machine learning for causal inference in biological networks: perspectives of this challenge. *Frontiers in Bioinformatics* 1:746712.

Li, H.; Cabeli, V.; Sella, N.; and Isambert, H. 2019. Constraint-based causal structure learning with consistent separating sets. *Advances in Neural Information Processing Systems* 32.

Lipton, Z. C. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16(3):31–57.

Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T. Y.; and Tegmark, M. 2024. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*.

Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; and

Lee, S.-I. 2020. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence* 2(1):56–67.

Molnar, C. 2020. *Interpretable machine learning*. Morrisville, NC: Lulu.com.

Pearl, J. 2009. *Causality*. New York, NY: Cambridge University Press, second edition.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5):206–215.

Salehkaleybar, S.; Ghassami, A.; Kiyavash, N.; and Zhang, K. 2020. Learning linear non-gaussian causal models in the presence of latent variables. *Journal of Machine Learning Research* 21(39):1–24.

Shah, R. D., and Peters, J. 2020. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics* 48(3):1514–1538.

Shimizu, S.; Inazumi, T.; Sogawa, Y.; Hyvarinen, A.; Kawahara, Y.; Washio, T.; Hoyer, P. O.; and Bollen, K. 2011. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research* 12(Apr):1225–1248.

Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; and Lakkaraju, H. 2020. How can we fool lime and shap? adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 557–563.

Spirtes, P.; Glymour, C. N.; and Scheines, R. 2000. *Causation, Prediction, and Search*. Cambridge, MA: MIT Press, second edition.

Vashishtha, A.; Reddy, A. G.; Kumar, A.; Bachu, S.; Balasubramanian, V. N.; and Sharma, A. 2025. Causal order: The key to leveraging imperfect experts in causal inference. In *The Thirteenth International Conference on Learning Representations*.

Yan, C., and Zhou, S. 2020. Effective and scalable causal partitioning based on low-order conditional independent tests. *Neurocomputing* 389:146–154.

Yu, Y.; Chen, J.; Gao, T.; and Yu, M. 2019. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, 7154–7163. PMLR.

Zheng, X.; Aragam, B.; Ravikumar, P. K.; and Xing, E. P. 2018. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31.