VI. Appendix

A. Proof of Theorem 1

As discussed in Sec. III-C The loss function is written as

$$\begin{aligned} L &= \sum_{j=1}^{N} \mathcal{L}_2(\hat{x}_{1:L}, \tilde{x}_{1:L}) + \lambda ||G||_1 \\ &= \sum_{j=1}^{N} \sum_{t=1}^{L} (x_t^j - f_{\phi_j}(X \odot \sigma(g_\theta(x_{t-\tau:t-1})))) + \lambda ||\sigma(g_\theta(x_{t-\tau:t-1}))||_1 \end{aligned} \tag{5}$$

Theorem:

Given a non-stationary time-series dataset $X = \{x_{1:L}^i\}_{i=1}^N$ generated with time-varying causal mechanisms, we have:

1. $\exists \lambda, \forall \tau \in \{1, ..., \tau_{\max}\}$, $\sigma(g_\theta((x_{t-\tau:t-1}))_{\tau,ij})$ converges to 0 if time-series i does not Granger cause j in any time period.

2. $\exists \tau \in \{1, ..., \tau_{\max}\}, \exists t \in \{1, ..., L\}$, $\sigma(g_\theta((x_{t-\tau:t-1}))_{\tau,ij})$ converges to 1 if time-series i Granger causes j at time t,

if the following conditions hold:

1. The predictor network f_{ϕ_j} in the second stage models the time-varying generative function $f_{\phi_j}(t, \cdot)$ with an error smaller than an arbitrarily small value $e_{NN,j}$;

2. $\exists \lambda_0, \forall i, j = 1, ..., N, \forall t \in \{1, ..., L\}$, $||f_{\phi_j}(X \odot \sigma(g_\theta((x_{t-\tau:t-1}))_{\tau,ij=1}) - f_{\phi_j}(X \odot \sigma(g_\theta((x_{t-\tau:t-1}))_{\tau,ij=0})||_2^2 > \lambda_0$.

The Proof of the theorem 1 can be own in the following steps:

Step 1: Set up the loss function

The loss function is defined as:

$$L = \sum_{j=1}^{N} \sum_{t=1}^{L} (x_t^j - f_{\phi_j}(X \odot (\sigma(\theta)^T)))^2 + \lambda ||\sigma(\theta)^T||_1 \tag{6}$$

Step 2: Calculate the gradient

Using the REINFORCE trick [21] θ, we can calculate the gradient:

For $\theta_{\tau,ij}$:

$$\begin{aligned} \frac{\partial}{\partial \theta_{\tau,ij}} \mathbb{E}_S[L] &= \mathbb{E}_S[(x_t^j - f_{\phi_j}(X \odot (\sigma(\theta)^T))^2 \frac{\partial}{\partial \theta_{\tau,ij}} \log p_{S_{\tau,ij}}] + \lambda \theta'(\theta_{\tau,ij}) \\ &= \sigma(\theta_{\tau,ij})((x_t^j - f_{\phi_j}(X \odot (\sigma(\theta_{\tau,ij=1})^T)))^2 \frac{1}{\sigma(\theta_{\tau,ij})} \sigma'(\theta_{\tau,ij}) + \\ &(1 - \sigma(\theta_{\tau,ij}))((x_t^j - f_{\phi_j}(X \odot (\sigma(\theta_{\tau,ij=0})))^2 \frac{1}{\sigma(\theta_{\tau,ij}) - 1} \sigma'(\theta_{\tau,ij}) + \lambda \sigma'(\theta_{\tau,ij}) \\ &= \sigma'(\theta\tau, ij)[(x_t^j - f_{\phi_j}(X \odot (\sigma(\theta_{\tau,ij=1})))^2 - \\ &(x_t^j - f_{\phi_j}(X \odot (\sigma(\theta_{\tau,ij=0})))^2] \end{aligned} \tag{7}$$

This approach using the REINFORCE trick allows us to compute gradients through the discrete sampling operations for both S and B. The key idea is that we're not directly differentiating the loss with respect to S, but rather with respect to the parameters of their distributions (θ respectively).

Step 3: Analyze the non-causal case

If time-series i does not Granger cause j at time t, then

$$f_{\phi_j}(X \odot \sigma(\theta_{\tau,ij=1}) \approx f_{\phi_j}(X \odot (\sigma(\theta_{\tau,ij=0})) \tag{8}$$

This leads to:

$$\frac{\partial}{\partial \theta_{\tau,ij}} \mathbb{E}_S[L] \approx \sigma'(\theta\tau, ij) > 0 \tag{9}$$

The positive gradient for $\theta_{\tau,ij}$ will push it towards $-\infty$, causing $(\sigma(\theta))_{\tau,ij}$ to converge to 0.

Step 4: Analyze the causal case

If time-series i Granger causes j at time t, then $\exists \tau, \exists t$ such that

$$f_{\phi_j}(X \odot (S_{\tau,ij=1}^T) \neq f_{\phi_j}(X \odot (S_{\tau,ij=0})) \tag{10}$$

Let $\Delta f_{i,j}(t) = f_{\phi_j}(X \odot (S_{\tau,ij=1}) - f_{\phi_j}(X \odot (S_{\tau,ij=0})$. The gradient becomes:

$$\begin{aligned}\frac{\partial}{\partial \theta_{\tau,ij}} \mathbb{E}_S[L] =& \sigma'(\theta\tau, ij)[(x_t^j - f_{\phi_j}(X \odot (\sigma(\theta_{\tau,ij=1})))^2 \\ & - (x_t^j - f_{\phi_j}(X \odot (\sigma(\theta_{\tau,ij=0})^T))^2]\end{aligned} \tag{11}$$

The negative gradient for $\theta_{\tau,ij}$ will push it towards $+\infty$, causing $(\sigma(\theta))_{\tau,ij}$ to converge to 1.

Step 5: Establish conditions for convergence

The gradient for $\theta_{\tau,ij}$ is expected to be negative when:

$$(x_t^j - f_{\phi_j}(X \odot \text{sq}(\sigma(\theta_{\tau,ij=1})))^2 - (x_t^j - f_{\phi_j}(X \odot (\sigma(\theta_{\tau,ij=0})))^2 < -\lambda \tag{12}$$

This condition is more likely to be met when including the causal link $(\sigma(\theta_{\tau,ij=1}))$ significantly improves prediction compared to excluding it $(\sigma(\theta_{\tau,ij=0}))$.

Step 6: Conclude the proof

With a properly chosen λ satisfying the above inequality, $\theta_{\tau,ij}$ will go towards $+\infty$ for causal relationships, causing $((\sigma(\theta))_{\tau,ij}$ to converge to 1 when the causal relationship is active.

B. Detailed Discussion about the attention matrix

The attention matrix from the last attention layer of a large language model (LLM), [17] has been proposed to interpret a given sentence in a causal view [22]. This attention matrix is claimed to be viewed as a coefficient correlation matrix, which can be used to conduct conditional independence tests [23]. However, in our experiment, we found out the attention matrices can neither be directly used as representation of causal mechanism nor be utilized to extract causal graph via conditional independence test.

Our study employed t-SNE (t-distributed stochastic neighbor embedding) to project high-dimensional attention data onto a 2D space, allowing for visual analysis of potential causal classes (Fig. 3). We defined five distinct classes (0-4), each representing a unique causal mechanism, as illustrated by the directed graphs in Fig. 3.

The t-SNE visualization reveals clear clustering of data points, suggesting that the attention matrices contain some structure related to causal mechanisms. However, the overlap between classes, particularly evident in the central region of the plot, indicates that these representations are not perfectly separable based on causal structure alone.

Further, we show conditional independecne test method utilized in [22] can not be straightforwardly applied on attention matrices to discover Granger causal relations for time series data. As shown in Figure Fig. 4, while some predictions (e.g., the purple circle) align closely with the ground truth, others (e.g., the black circle) deviate significantly. This inconsistency suggests that attention matrices, in their raw form, do not provide a reliable basis for causal mechanism identification or graph extraction through conditional independence testing.

Several factors may contribute to these limitations. Firstly, the contextual ambiguity in attention may capture linguistic patterns that do not directly correspond to causal relationships. Secondly, the complexity of causal structures in time series data often extends beyond the simplified representations used in our classification scheme. Lastly, while attention mechanisms have proven powerful for language understanding tasks, they may not be inherently optimized for causal inference in time series. In conclusion, while the attention matrices of LLMs show some promise in capturing aspects of causal structure in time

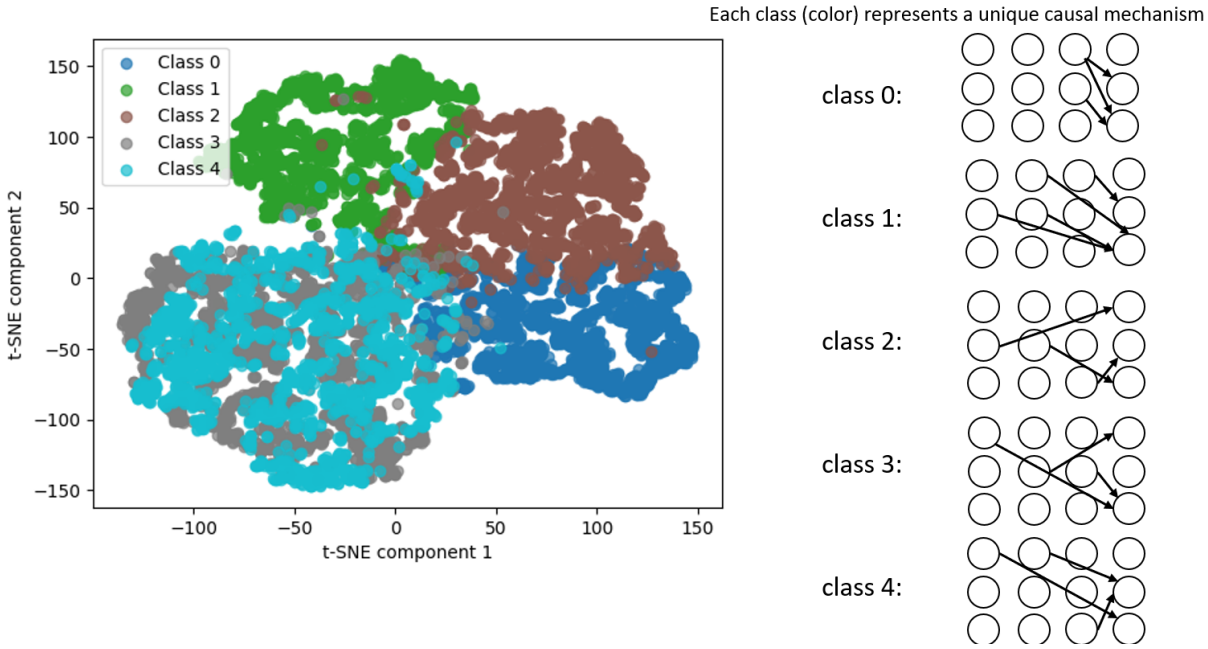

Fig. 3: Attention Matrix can be used to represent causal relations for Time Series.Each class (color) represents a unique causal mechanism

series data, our findings indicate that they cannot be directly utilized as representations of causal mechanisms or for extracting causal graphs via conditional independence tests. Future research should explore more sophisticated methods for bridging the gap between linguistic attention and causal inference, potentially incorporating domain knowledge or developing specialized architectures that explicitly model temporal causal relationships.

C. Implementation Detail

The study employs two types of simulated datasets to evaluate the proposed method: a linear Vector Autoregressive (VAR) model and a nonlinear VAR model. These datasets are designed to mimic real-world scenarios where causal relationships may be linear or nonlinear in nature. For each type of dataset, the time series structure consists of multivariate data where each

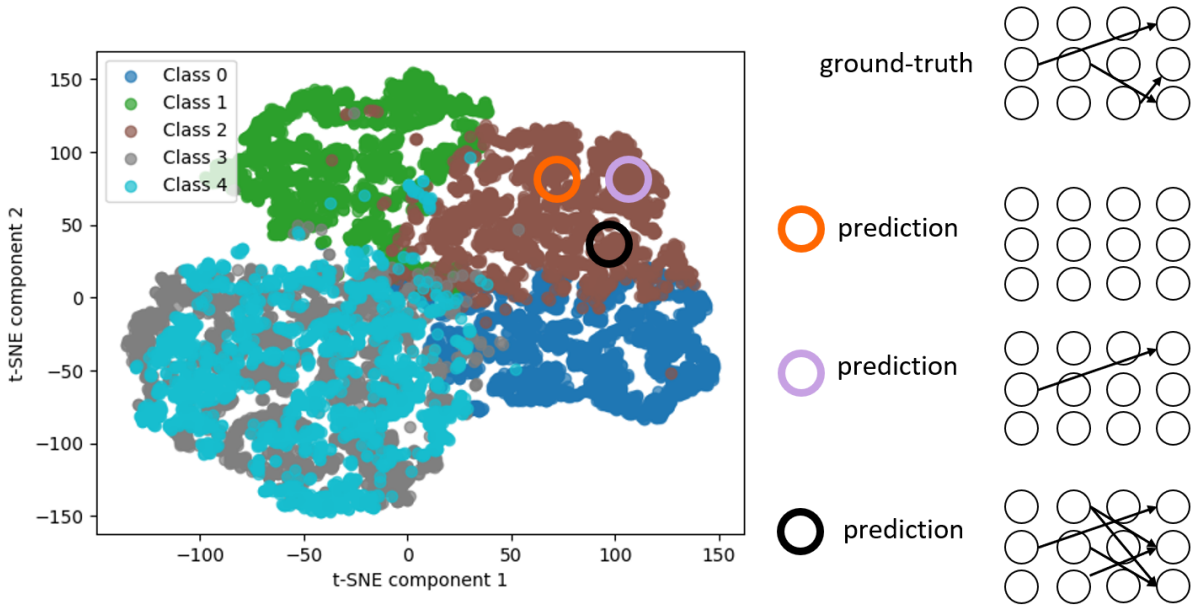

Fig. 4: The causal graphs generated from different samples in the same TSNE cluster are different.

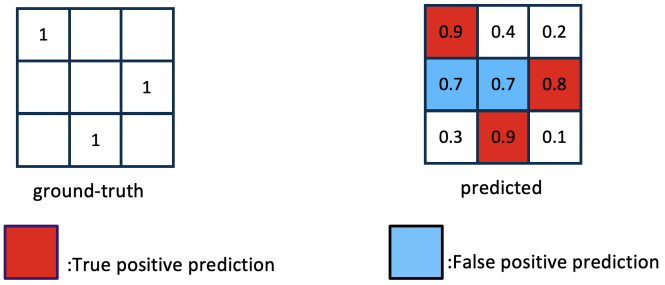

Fig. 5: True positive rate (TPR), False positive rate (FPR) calculation.

variable's current value depends on past values of itself and other variables. To simulate non-stationarity, the entire time series is divided into five distinct segments, each governed by a different causal mechanism. This design reflects real-world scenarios where underlying causal relationships may evolve over time. In the linear VAR model, each variable is a linear combination of lagged values of all variables, plus some noise. The coefficients of these linear combinations change across the five segments to represent different causal mechanisms. The nonlinear VAR model introduces more complex relationships between variables, potentially involving quadratic function ax^2, exponential function$a \cdot exp(x)$, sinusoidal $asin(x)$ function and cubic interaction effects anx^3, where a is coefficient that is randomly sampled from a normal distribution. As with the linear model, these nonlinear relationships also change across the five segments.

D. Details for Metrics Calculation

As shown in Fig. 5, to evaluate method performances, we report true positive rate (TPR), false positive rate (FPR), and Area Under the Receiver Operating Characteristic (AUROC). TPR measures the proportion of correctly identified true causal connections, calculated as the number of correctly predicted edges (represented by the red squares in the 'predicted' matrix that align with the '1's in the 'ground-truth' matrix) divided by the total number of '1's in the ground-truth matrix. FPR measures the proportion of incorrectly included non-existent causal connections, calculated as the number of incorrectly predicted edges (represented by the blue squares in the 'predicted' matrix that align with empty cells in the 'ground-truth' matrix) divided by the total number of empty cells in the ground-truth matrix. AUROC represents the overall ability to distinguish between true causal connections and non-connections across various certainty thresholds. It is calculated from the ROC curve plotted using TPR and FPR at different prediction strength thresholds. In the context of the figure, this would involve considering how the distribution of red (true positive) and blue (false positive) squares in the 'predicted' matrix changes as we vary the threshold for what prediction strength (the numbers in the 'predicted' matrix) we consider a positive prediction.