# CausalStock: Deep End-to-end Causal Discovery for News-driven Stock Movement Prediction

**Shuqi Li**[1][*]   **Yuebo Sun**[1][*]   **Yuxin Lin**[2]   **Xin Gao**[3]   **Shuo Shang**[4]   **Rui Yan**[1][†]

[1] Gaoling School of Artificial Intelligence, Renmin University of China
[2] Peking University [3] King Abdullah University of Science and Technology
[4] University of Electronic Science and Technology of China
{shuqili, sunyuebo0418, ruiyan}@ruc.edu.cn
linyuxin@stu.pku.edu.cn, xin.gao@kaust.edu.sa, jedi.shang@gmail.com

## Abstract

There are two issues in news-driven multi-stock movement prediction tasks that are not well solved in the existing works. On the one hand, "relation discovery" is a pivotal part when leveraging the price information of other stocks to achieve accurate stock movement prediction. Given that stock relations are often unidirectional, such as the "supplier-consumer" relationship, causal relations are more appropriate to capture the impact between stocks. On the other hand, there is substantial noise existing in the news data leading to extracting effective information with difficulty. With these two issues in mind, we propose a novel framework called CausalStock for news-driven multi-stock movement prediction, which discovers the temporal causal relations between stocks. We design a lag-dependent temporal causal discovery mechanism to model the temporal causal graph distribution. Then a Functional Causal Model is employed to encapsulate the discovered causal relations and predict the stock movements. Additionally, we propose a Denoised News Encoder by taking advantage of the excellent text evaluation ability of large language models (LLMs) to extract useful information from massive news data. The experiment results show that CausalStock outperforms the strong baselines for both news-driven multi-stock movement prediction and multi-stock movement prediction tasks on six real-world datasets collected from the US, China, Japan, and UK markets. Moreover, getting benefit from the causal relations, CausalStock could offer a clear prediction mechanism with good explainability.

## 1 Introduction

The financial services industry has maintained a leading position in embracing data science methodologies to inform investment determinations. Within this domain, quantitative trading has garnered substantial attention from both academia and industry. Researchers have consistently worked on exploring different approaches to predict the stock movement (rise or fall of stock price) for many years, such as uni-stock movement prediction [21], multi-stock movement prediction [44, 23], news-driven stock movement prediction [42, 19] and so on, which have shown significant success. These methods usually model the stock movement prediction task as a time series classification problem.

In this paper, we focus on the news-driven multi-stock movement prediction task. A prevalent model paradigm for this task often takes the historical price features and the stock-related news of multiple stocks as inputs and then leverages the well-designed neural networks to make stock movement predictions. There are two key modeling points for tackling this task: modeling the stock

---

relations to enhance the prediction accuracy, and building the text mining module to extract effective information from news data that benefits stock movement prediction. Although previous work has made significant progress, there are still some issues that require further attention. We will elaborate on them in the following.

For stock relation modeling, many existing works are commonly attention-based [15, 19, 23] or graph-based [34, 23]. These methods aim to model the correlation relation between stocks. However, the company relations are often unidirectional, such as the "investing" and "member of," leading to the unidirectional relations of their stocks. Thus, causal relations are more appropriate for depicting the impact between stocks, as they identify the direction of information flow and are more informative than correlations. With the development of causal science, many researchers have started to use deep end-to-end networks for causal relations discovery of panel data or temporal data [9, 14], in which the causal relations are defined as directed acyclic graphs, i.e., causal graphs, and the Functional Causal Models (FCMs) are often utilized to optimize the causal graph by simulating the data generation mechanism. This provides a solid theoretical foundation for causal discovery for stocks.

In recent years, an extrinsic text mining module has emerged as a plausible avenue through the alignment of financial news and social media posts, thereby elucidating intricate market insights that extend well beyond mere considerations of price dynamics, trading volumes, or financial indicators [41, 17, 35, 33]. Conventional text representations obtained by using GRU [42] or LSTM [15] exhibit many limitations. Specifically, news text data are often characterized by substantial noise because of the presence of irrelevant or ambiguous information [38, 7, 37]. The effective information for stock movement prediction gets intertwined with this noise, presenting a considerable challenge for these modules to discern meaningful signals accurately. In contrast, Large Language Models (LLMs) have unique advantages in this situation due to their advanced knowledge and reasoning abilities. Besides, LLMs can identify meaningful information within noisy environments [29, 4].

Motivated by these requirements, we propose an innovative news-driven multi-stock movement prediction model named CausalStock. In CausalStock, we design a Denoised News Encoder, which leverages LLMs to score every news text from multiple perspectives. Then the evaluation scores are taken as denoised text representations. To discover the causal relations between stocks, we propose a Lag-dependent temporal causal discovery module, from which we obtain the causal graph distribution. Based on the input market information and learned causal graph distribution, CausalStock employs an FCM [14] to make predictions. We summarize the contributions of our paper as follows:

- We propose a novel news-driven multi-stock movement prediction method named Causal-Stock, which could discover the causal relations among stocks and make accurate movement predictions simultaneously.

- Different from the past lag-independent causal discovery method [9], CausalStock involves a lag-dependent temporal causal discovery module, which intuitively links the temporal causal relations according to the time lag, making it more suitable for temporal stock data.

- To extract useful information from the massive noisy news text data, an LLM-based Denoised News Encoder is proposed by taking advantage of the evaluation ability of LLM, which outputs the denoised news representation for better information utilization.

Experiments on 6 public benchmarks show the performance of CausalStock as a news-driven multi-stock movement prediction method. Moreover, we conduct extensive analytical experiments to show the explainability of our key modules.

## 2 Related work

**Stock prices prediction**    In traditional trading practices, there are two analysis paradigms commonly used to make stock movement predictions: technical analysis and fundamental analysis. With technical analysis, investors and traders tend to forecast stock prices relying on historical price patterns. Fundamental analysis aims to assess the intrinsic value of a stock by considering other factors besides historical prices, such as financial statements, industry trends, and economic conditions.

Since stock movement prediction involves sequential data, RNN-based networks are applied in many works. ALSTM [28] integrated a dual-stage attention mechanism with LSTM. Adv-ALSTM [8] further employed adversarial training by adding perturbations to simulate the stochastic and unstable

nature of the price variable. In recent years, researchers also exploited attention-based mechanisms to model complex interactions. DTML [44] is proposed to predict by using a transformer and LSTM to capture the asymmetric and dynamic correlations between stocks. With the development and prosperity of NLP technology, text from social media and online news has become a new popular source of fundamental analysis. HAN [15] designed two attention networks to recognize both the influential time periods of a sequence and the important news at a given time. Stocknet [42] proposed a deep generative model with recurrent, continuous latent variables. MSHAN [13] exploited a multi-stage TCN-LSTM hybrid model. PEN [21] proposed a Shared Representation Learning module to capture interactions between price data and text data. Additionally, many works modeled the correlation between stocks to enhance stock price prediction. MAN-SF [32] constructed a graph attention network with price features, social media, and inter-stock relationships based on the interrelationship between price and tweets. CMIN [23] was proposed to model the asymmetric correlations between stocks by computing transfer entropy. In addition, Co-CPC [40] modeled the dependence between a certain stock industry and relevant macroeconomic variables. All the aforementioned methods aim to discover the correlation relations among stocks, as elaborated before, the causal relations are more appropriate to depict the information flow of stocks. In this work, we aim to model the causal relations for better stock movement prediction performance.

**Causal discovery**  The conventional approach to discovering causal relations typically involves conducting randomized experiments [27, 12]. However, conducting randomized experiments can often be excessively expensive, overly time-consuming, or impossible to execute. Consequently, causal discovery, which aims to infer causal relationships from purely observational data, has attracted considerable attention within the machine learning community over the last decade [5, 12]. Causal discovery can be classified into three groups: constraint-based [10, 30, 31], score-based [2, 26, 46], and functional causal models (FCMs) [12, 26]. FCMs define the causal relations by directed acyclic graphs (DAGs) and identify causal links through nonlinear functions, such as neural networks [14, 9, 47, 18]. Specifically, DECI [9] is a deep end-to-end framework to discover causal relations based on additive noise FCM. After that, Rhino [14] was proposed to tackle the temporal causal discovery problem, which incorporates non-linear relations, instantaneous effects, and flexible history-dependent noise. In this work, we focus on utilizing the FCM to discover stock relations.

# 3    Preliminary & problem formulation

## 3.1    Preliminary

In CausalStock, we integrate the model inputs with causal relations into FCM for prediction. In this section, we introduce the fundamental concepts of FCM and the temporal causal graph.

**Temporal causal graph**  Consider a multivariate time series $\{X_t^i\}_{i=1}^D$ with $D$ variables, the temporal causal graph $\boldsymbol{G}$ [46] is commonly defined as a series of directed acyclic graph $\boldsymbol{G} = [G_1, G_2, \ldots, G_L] = \{G_l\}_{l=1}^L \in \mathbb{R}^{L \times D \times D}$ with maximum time lag $L$. Each $G_l \in \mathbb{R}^{D \times D}$ specifies the lagged causal relationships between $X_{t-l}$ and $X_t$, the element $G_{l,ji} = 1$ if there exists a causal link $X_{t-l}^j \rightarrow X_t^i$ and $G_{l,ji} = 0$ otherwise.

**Functional causal model (FCM)**  FCM represents a set of generative functions that incorporate the input features based on causal knowledge (structured as a causal graph) to produce a final prediction. Optimizing the prediction accuracy concurrently refines the underlying causal graph. The theoretical demonstration presented in [14, 9] indicates that if the prediction is accurate, the causal graph can be considered a reliable approximation of real causal relations. Given the temporal causal graph $\boldsymbol{G}$ defined as before, a temporal FCM is defined as follows:

$$X_t^i = F_i \left( \mathbf{Pa}_{\boldsymbol{G}}^i \left( < t \right), z_t^i \right), \tag{1}$$

where $\mathbf{Pa}_{\boldsymbol{G}}^i \left( < t \right)$ indicates the time-lagged parent nodes of variable $X_t^i$ following the temporal causal graph $\boldsymbol{G}$ and $z_t^i$ represents mutually and serially independent exogenous noise. Here $F_i$ is a function which implies how variable $X_t^i$ depends on its parents and the noise $z_t^i$. Given the distribution of noises for different variables $\{z_t^i\}_{i=1}^D$ and causal graph, this FCM induces a joint distribution of the multivariate time series process $\{X_t^i\}_{i=1}^D$.

## 3.2  Problem formulation

In this paper, we focus on tackling the news-driven multi-stock movement prediction task. For the target trading day $T$, we denote the model inputs as the past $L$ time lag information of $D$ stocks as $\boldsymbol{X}_{<T} = \{X_t^i\}_{t=T-L:T-1}^{i=1:D} = [\boldsymbol{C}_{<T}, \boldsymbol{P}_{<T}] = \{[C_t^i, P_t^i]\}_{t=T-L:T-1}^{i=1:D}$, where $C_t^i$ and $P_t^i$ represent the news corpora representation and the historical price features representation of $i$-th stock at time step $t$ respectively. The objective is to predict the movement of adjusted close prices $\boldsymbol{y}_T = \{y_T^i\}_{i=1}^D \in \mathbb{R}^{D \times 1}$ on $T$-th trading day of all stocks simultaneously, where $y_T^i \in \{0, 1\}$ representing the $i$-th stock price will fall or rise at trading day $T$, i.e., stock movement. In a theoretical way, this task could be trained by maximizing the log-likelihood of conditional probability distribution $p(\boldsymbol{y}_T \mid \boldsymbol{X}_{<T})$, so that the most likely $\boldsymbol{y}_T$ are generated.
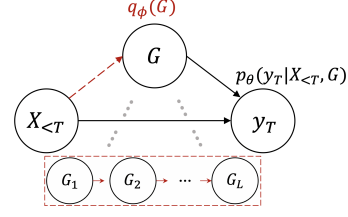


Figure 1: Illustration of the process of stock movement $\boldsymbol{y}_T$ forecasting. The forecasting process is denoted by solid lines with parameters $\theta$ and the causal discovery process is denoted by dashed lines with variational approximation parameters $\phi$, $q_\phi$ is the posterior distribution of the causal graph.

# 4  CausalStock

## 4.1  Model overview

The conditional probability distribution could be further factorized as follows:

$$p(\boldsymbol{y}_T \mid \boldsymbol{X}_{<T}) = \int_{\boldsymbol{G}} p(\boldsymbol{y}_T, \boldsymbol{G} \mid \boldsymbol{X}_{<T}) \, d\boldsymbol{G} = \int_{\boldsymbol{G}} p(\boldsymbol{y}_T \mid \boldsymbol{X}_{<T}, \boldsymbol{G}) \, p(\boldsymbol{G} \mid \boldsymbol{X}_{<T}) \, d\boldsymbol{G}. \qquad (2)$$

The overall process is taken as two joint training parts: temporal causal graph discovery $p(\boldsymbol{G} \mid \boldsymbol{X}_{<T})$ and the prediction process given the causal relations $p(\boldsymbol{y}_T \mid \boldsymbol{X}_{<T}, \boldsymbol{G})$. The probabilistic graphic representation of this modeling process is shown in Figure 1. In CausalStock, we develop a lag-dependent causal discovery module, according to which we could take another step by modeling $p(\boldsymbol{G} \mid \boldsymbol{X}_{<T})$ as a lag-dependent format:

$$p(\boldsymbol{G} \mid \boldsymbol{X}_{<T}) = p(G_1 \mid X_{T-1}) \prod_{l=2}^{L} p(G_l \mid G_{l-1}, X_{T-l}). \qquad (3)$$

For the prediction part $p(\boldsymbol{y}_T \mid \boldsymbol{X}_{<T}, \boldsymbol{G})$, we design an FCM as shown in Equation 9 to predict the future movement based on the past information $\boldsymbol{X}_{<T}$ and the discovered temporal causal graph $\boldsymbol{G}$.

In a nutshell, CausalStock comprises three primary components as shown in Figure 2:

1. Market Information Encoder (MIE) encodes the news text and price features. In this part, an LLM-based Denoised News Encoder is proposed;

2. Lag-dependent Temporal Causal Discovery (Lag-dependent TCD) module leverages variational inference to mine the causal relationship based on the given market information of stocks, i.e., modeling $p(\boldsymbol{G} \mid \boldsymbol{X}_{<T})$;

3. Functional Causal Model (FCM) generates the prediction of future price movements according to the discovered causal graph, i.e., modeling $p(\boldsymbol{y}_T \mid \boldsymbol{X}_{<T}, \boldsymbol{G})$.

## 4.2  Market information encoder (MIE)

Market Information Encoder (MIE) takes news corpora and numerical stock price features as inputs, and outputs the historical market information representations $\boldsymbol{X}_{<T} = [\boldsymbol{C}_{<T}, \boldsymbol{P}_{<T}] = \{[C_t^i, P_t^i]\}_{t=T-L:T-1}^{i=1:D} = \{X_t^i\}_{t=T-L:T-1}^{i=1:D}$ for $D$ stocks with time lag $L$. For $i$-th stock, each time step representation $X_t^i$ is the combination of the text representation $C_t^i$ generated by the news encoder and the historical price features representation $P_t^i$ generated by the price encoder.

**Price encoder**  For $i$-th stock, we denote the raw adjusted closing, highest, lowest, open, closing prices and trading volume on trading day $t$ as $\hat{P}_t^i = \left[\hat{P}_t^{i,a}, \hat{P}_t^{i,h}, \hat{P}_t^{i,l}, \hat{P}_t^{i,o}, \hat{P}_t^{i,c}, V_t\right]$. By feeding $\hat{P}_t^i$ into the embedding layer, the historical prices could be represented as $P_t^i \in \mathbb{R}^{d_p \times 1}$, where $d_p$ is the price embedding size.
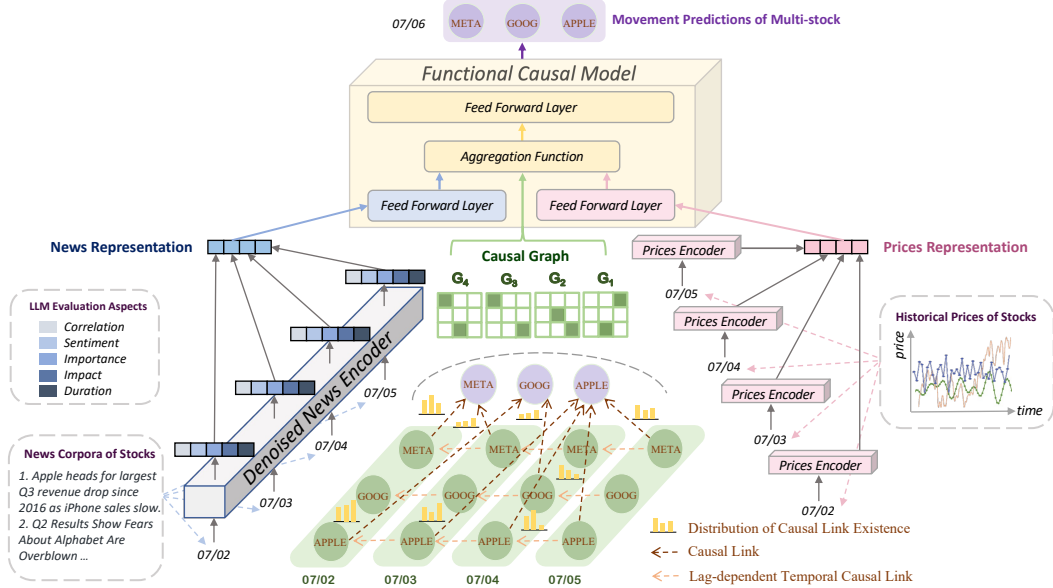
Figure 2: The structure of CausalStock. For illustration, we use market information during 07/02 - 07/05 of three stocks (AAPL, GOOG, META) to predict the movements of 07/05.

**LLM-based denoised news encoder (DNE)** News Encoder aims to embed stock-related news text, which evolves from the small sequential module, e.g., GRU [23], to pre-trained models, e.g., Bert and Roberta [6, 22], offering greater performance and scalability. However, news text data often contains massive noise due to the following factors. Firstly, news comes from a wide range of sources with varying degrees of reliability and editorial standards. This variability contributes to inconsistencies and inaccuracies in the information presented. Secondly, the sheer volume of news content generated daily can lead to information overload, where significant information is buried under less relevant or redundant information. Thirdly, the use of complex or ambiguous language can also add noise, making it difficult to extract precise information relevant to specific needs, such as stock movement prediction. Addressing these challenges requires sophisticated text mining and natural language processing techniques to filter out noise and extract useful, accurate information from news text data. With the development of large language models, current large language models can accurately capture the meaning of text and have a strong capability to evaluate text. Therefore, here we propose an LLM-based Denoised News Encoder to tackle these standing challenges.

LLM-based Denoised News Encoder is an innovative textual representation approach that not only proficiently captures salient information from extensive news texts but also assimilates external knowledge derived from LLMs to enrich the representations. Specifically, we employ an LLM and devise a series of prompts (see Appendix A for the whole designed prompts) to analyze the relationship between a news text and a specific stock from five dimensions: correlation between the news and the stock, sentiment polarity of the news, significance of the news event, potential impact of the news on stock prices, and duration of the news impact. Each dimension is scored, with Correlation, Importance, Impact, and Duration ranging from $0$ to $10$, while Sentiment varies from $-1$ to $1$. Thus the $i$-th text at day $t$ is represented as a five-dimensional representation $\hat{C}_t^i \in \mathbb{R}^{l \times 5}$. After the embedding layer, we obtain the final denoised news embedding $C_t^i \in \mathbb{R}^{l \times d_m}$. This novel encoding method amalgamates information derived from the primary text, the external knowledge embedded and the evaluation ability within the LLM. Besides, this method effectively reduces the significant noise present in the original text data.

### 4.3 Lag-dependent temporal causal discovery (Lag-dependent TCD)

In this section, we propose Lag-dependent Temporal Causal Discovery module. Inspired by [14], our model takes a Bayesian view for modeling the distribution of temporal causal graph, which aims to learn the posterior distribution $p\left(\boldsymbol{G} \mid \boldsymbol{X}_{<T}\right)$. Unfortunately, the exact graph posterior is intractable

because of the large combination space of $\boldsymbol{G}$. Here we adopt the variational inference [3] to get the approximator $q_\phi\left(\boldsymbol{G}\right)$, where $\phi$ indicates the parameter set of variational inference.

**Graph prior**   The prior $p\left(\boldsymbol{G}\right)$ consists of two parts: the graph sparseness prior and the domain-specific knowledge prior. The unnormalised graph prior is as follows,

$$p(\boldsymbol{G}) \propto \exp\left(-\lambda_s \|\boldsymbol{G}_{1:L}\|_F^2 - \lambda_d \|\boldsymbol{G}_{1:L} - \boldsymbol{G}_{1:L}^p\|_F^2\right), \tag{4}$$

where $\lambda_s$ and $\lambda_d$ are scalar weights of graph sparseness and domain-specific knowledge constraint; $\boldsymbol{G}^p$ is an optional domain-specific knowledge graph, which allows users to incorporate pre-defined knowledge for guiding CausalStock, turning it into a knowledge and data-driven framework. Suppose a sudden event affects the causal relationships between stocks, such as a company ending a partnership. By incorporating this new pre-defined knowledge into $\boldsymbol{G}^p$, the causal graph is dynamically updated to reflect the latest market structure and relationship changes. $\|\cdot\|_F$ denotes Frobenius norm. It should be noted that there is no need to give a DAG constraint for the temporal causal graph defined in our paper, it is DAG naturally for the irreversibility of time.

**Variational approximating graph posterior**   According to Equation 3, we factorize the approximator $q_\phi\left(\boldsymbol{G}\right)$ in the same way. For each underlying causal link $G_{l,ji}$ in $\boldsymbol{G}$, we let the posterior $q_\phi\left(G_{l,ji} \mid G_{l-1,ji}\right)$ subject to a Bernoulli distribution $\boldsymbol{B}$. So that the probability distribution of $q_\phi\left(\boldsymbol{G}\right)$ could be a product of Bernoulli distributions as follows,

$$q_\phi\left(\boldsymbol{G}\right) = q_\phi\left(G_1\right)\prod_{l=2}^{L}q_\phi\left(G_l \mid G_{l-1}\right) = \prod_{i=1}^{D}\prod_{j=1}^{D}q_\phi\left(G_{1,ji}\right)\prod_{l=2}^{L}\prod_{i=1}^{D}\prod_{j=1}^{D}q_\phi\left(G_{l,ji} \mid G_{l-1,ji}\right). \tag{5}$$

The existence and non-existence likelihood tensors of causal links are parameterized as $\boldsymbol{U} = \{U_l\}_{l=1}^{L} = \{u_{l,ji}\}_{l=1:L}^{j,i=1:D} \in \mathbb{R}^{L \times D \times D}$ and $\boldsymbol{V} = \{V_l\}_{l=1}^{L} = \{v_{l,ji}\}_{l=1:L}^{j,i=1:D} \in \mathbb{R}^{L \times D \times D}$ separately, where $u_{l,ji}$ indicates the likelihood for edge existence from $X_{T-l}^j$ to $y_T^i$ and $v_{l,ji}$ is the likelihood for no-edge, which are all learnable parameters. To model the dependency between $G_{l,ij}$ with $G_{l-1,ij}$, we propose the following transformation:

$$u'_{l,ji} = h_u\left(u_{l,ji}, u_{l-1,ji}\right), v'_{l,ji} = h_v\left(v_{l,ji}, v_{l-1,ji}\right), \tag{6}$$

where $h_u$ and $h_v$ are trainable 3-layer MLPs. After normalization, the link existence probability tensor is denoted as $\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_l\}_{l=1}^{L} = \{\sigma_{l,ji}\}_{l=1:L}^{j,i=1:D}$,

$$\sigma_{l,ji} = \exp\left(u'_{l,ji}\right) / \left(\exp\left(u'_{l,ji}\right) + \exp\left(v'_{l,ji}\right)\right), \tag{7}$$

where $\sigma_{l,ji}$ represents the link probability from $X_{T-l}^j$ to $y_T^i$. Thus, we could derive the variational posterior:

$$q_\phi\left(G_{1,ji}\right) \sim \mathbf{B}\left(1, \sigma_{1,ji}\right), q_\phi\left(G_{l,ji}|G_{l-1,ji}\right) \sim \mathbf{B}\left(1, \sigma_{l,ji}\right), q_\phi\left(\mathbf{G}\right) \sim \prod_{l=1}^{L}\prod_{i=1}^{D}\prod_{i=1}^{D}\mathbf{B}\left(1, \sigma_{l,ji}\right). \tag{8}$$

In the training stage, we employ the Gumbel-softmax reparameterization [24, 16] to stochastically estimate the gradients with respect to $\phi$. Besides, we design another parameterized learnable causal weight graph $\hat{\boldsymbol{G}} = \{\hat{G}_l\}_{l=1}^{L} \in \mathbb{R}^{L \times D \times D}$ to measure the causal degree. The separate design of the causal existence graph and the causal weight graph allows for more comprehensive modeling of causality. Once our model is fitted, the time series causal graph $\boldsymbol{G}$ can be sampled by $\boldsymbol{G} \sim q_\phi\left(\boldsymbol{G}\right)$ to represent the relation network and information flow of the stock market.

### 4.4  Functional causal model (FCM)

In this section, we design an FCM to model $p_\theta\left(\boldsymbol{y}_T \mid \boldsymbol{X}_{<T}, \boldsymbol{G}\right)$, where $\theta$ denotes the parameter set of FCM. We focus on additive noise FCM [18] to generate $\boldsymbol{y}_T = \{\boldsymbol{y}_T^i\}_{i=1}^{D} \in \mathbb{R}^{D \times 1}$:

$$\boldsymbol{y}_T^i = F_i\left(\mathbf{Pa}_{\boldsymbol{G}}^i\left(<T\right), z_T^i\right) = f_i\left(\mathbf{Pa}_{\boldsymbol{G}}^i\left(<T\right)\right) + z_T^i, \tag{9}$$

where $z_t^i$ represents mutually and serially independent dynamical noise, and $f_i : \mathbb{R}^{D \times L} \to \mathbb{R}^1$ are general differentiable non-linear function that satisfies the relations specified by the temporal causal graph $\boldsymbol{G}$ strictly, namely, if $X_t^j \notin \mathbf{Pa}_{\boldsymbol{G}}^i\left(<T\right)$, then $\partial f_i/\partial X_t^j = 0$.

We design a novel FCM to aggregate market information including news and prices based on the discovered causal graph $G$ and causal weight graph $\hat{G}$:

$$f_i\left(\mathbf{Pa}_G^i\left(<T\right)\right) = \text{Sigmoid}\left(\zeta_i\left(\sum_{l=1}^{L}\sum_{j=1}^{D}G_{l,ji}\hat{G}_{l,ji}\left[\ell\left(P_{T-l}^j\right),\psi\left(C_{T-l}^j\right)\right]\right)\right), \tag{10}$$

where $\zeta_i$, $\ell$ and $\psi$ are all neural networks. $\ell$ and $\psi$ are shared weights across nodes and lags for efficient computation. $[\cdot,\cdot]$ denotes the concatenate operation. We apply the logistic Sigmoid function to output the movement probability of $y_T$ and use it directly as the output of CausalStock.

For the exogenous noise $z_T^i$ modeling, we adopt Gaussian distribution, i.e., $z_T^i \sim \mathcal{N}\left(0,\left(\sigma^i\right)^2\right)$, where per-variable variances $\left(\sigma^i\right)^2, i \in [1,D]$ are trainable parameters to represent the uncertainty part. According to Change of variables formula [18], the conditional distribution $p_\theta\left(y_T^i \mid \mathbf{Pa}_G^i\left(<t\right)\right)$ could be represented as:

$$p_\theta\left(y_T^i \mid \mathbf{Pa}_G^i\left(<t\right)\right) = p_{z_i}\left(z_T^i\right)\left|\frac{\partial F_i}{\partial z_T^i}\right|^{-1} = p_{z_i}\left(z_T^i\right), \tag{11}$$

where $p_{z_i}$ is the aforementioned Gaussian distribution for stock $i$. $\left|\frac{\partial F_i}{\partial z_T^i}\right|$ indicates the absolute value of the Jacobian-determinant for $F_i$, $\left|\frac{\partial F_i}{\partial z_T^i}\right|^{-1} = 1$ is derived according to Equation 9. Now the log likelihood $\log p_\theta\left(y_T \mid X_{<T}, G\right)$ could be further represented as:

$$\log p_\theta\left(y_T \mid X_{<T}, G\right) = \sum_{i=1}^{D}\log p_\theta\left(y_T^i \mid \mathbf{Pa}_G^i(<T)\right) = \sum_{i=1}^{D}\log p_{z_i}\left(z_T^i\right). \tag{12}$$

### 4.5 Training objective

We train our model by maximizing the conditional log-likelihood $\log p_\theta\left(y_T \mid X_{<T}\right)$. The variational evidence lower bound (*ELBO*) of the model objective is derived as follows:

$$
\begin{aligned}
\log p_\theta\left(y_T \mid X_{<T}\right) &= \log \int_G \frac{q_\phi\left(G\right)}{q_\phi\left(G\right)}p_\theta\left(y_T \mid X_{<T}, G\right)p\left(G\right)dG \\
&\geq \int_G q_\phi\left(G\right)\log p_\theta\left(y_T \mid X_{<T}, G\right)p\left(G\right)dG + H\left(q_\phi\left(G\right)\right) \\
&\geq E_{q_\phi(G)}\left[\log p_\theta\left(y_T \mid X_{<T}, G\right) + \log p\left(G\right)\right] + H\left(q_\phi\left(G\right)\right) \\
&\geq E_{q_\phi(G)}\left[\sum_{i=1}^{D}\log p_{z_i}\left(z_T^i\right) + \log p\left(G\right)\right] + H\left(q_\phi\left(G\right)\right).
\end{aligned}
\tag{13}
$$

Here, $p\left(G\right)$ represents the prior of causal graph, and $H\left(q_\phi\left(G\right)\right)$ is the entropy of the posterior approximator. $\log p_\theta\left(y_T \mid X_{<T}, G\right) = \log p_z\left(z_T\right)$ is the log-likelihood of the target distribution, in which $z_T$ is calculated by Equation 9 at training stage.

Besides, we further adopt the binary cross entropy loss as another objective $BCE\left(g_T, y_T\right)$ to improve the learning performance, where $g_T$ is the ground truth movement at target trading day $T$. Overall, the final training loss $\mathcal{L}$ is as follows,

$$BCE\left(g_T, y_T\right) = -\sum_{i=1}^{D}\left(g_T^i\log\left(y_T^i\right) + \left(1-g_T^i\right)\log\left(1-y_T^i\right)\right)$$

$$\mathcal{L} = \frac{1}{D}\left(-ELBO + \lambda BCE(g_T, y_T)\right) \tag{14}$$

where $\lambda$ is the scalar weight to balance loss terms. We note that the required assumptions and the theoretical guarantees are summarized in Appendix B.

## 5 Experiments

### 5.1 Experimental setup

Except for the news-driven multi-stock movement prediction task, our model could also handle the multi-stock movement prediction task without news by removing the Denoised News Encoder. Thus, we do the experiments for both two tasks.

Table 1: Main results of CausalStock and baselines for two stock movement prediction tasks on multiple datasets. Following the setting of baselines, the standard deviations are calculated across 10 runs for the news-driven task and 5 runs for the task without news.

| Models | News-driven Multi-stock movement prediction task | | | | | |
| | ACL18 (US) | | CMIN-US (US) | | CMIN-CN (CN) | |
| | ACC | MCC | ACC | MCC | ACC | MCC |
| --- | --- | --- | --- | --- | --- | --- |
| HAN | 57.64 ± 0.0040 | 0.0518 ± 0.0050 | 53.72 ± 0.0020 | 0.0103 ± 0.0015 | 53.59 ± 0.0037 | 0.0159 ± 0.0026 |
| StockNet | 58.23 ± 0.0030 | 0.0808 ± 0.0071 | 52.46 ± 0.0041 | 0.0220 ± 0.0025 | 54.53 ± 0.0062 | 0.0450 ± 0.0043 |
| PEN | 59.89 ± 0.0090 | 0.1556 ± 0.0018 | 53.20 ± 0.0051 | 0.0267 ± 0.0023 | 54.83 ± 0.0086 | 0.0857 ± 0.0065 |
| CMIN | 62.69 ± 0.0029 | 0.2090 ± 0.0016 | 53.43 ± 0.0085 | 0.0460 ± 0.0055 | 55.28 ± 0.0094 | 0.1110 ± 0.0990 |
| CausalStock | **63.42 ± 0.0039** | **0.2172 ± 0.0017** | **54.64 ± 0.0083** | **0.0481 ± 0.0057** | **56.19 ± 0.0084** | **0.1417 ± 0.0813** |

| Models | Multi-stock movement prediction task | | | | | |
| | KDD17 (US) | | NI225 (JP) | | FTSE100 (UK) | |
| | ACC | MCC | ACC | MCC | ACC | MCC |
| --- | --- | --- | --- | --- | --- | --- |
| LSTM | 51.18 ± 0.0066 | 0.0187 ± 0.0110 | 50.79 ± 0.0079 | 0.0148 ± 0.0162 | 50.96 ± 0.0065 | 0.0187 ± 0.0129 |
| ALSTM | 51.66 ± 0.0041 | 0.0316 ± 0.0119 | 50.60 ± 0.0066 | 0.0125 ± 0.0139 | 51.06 ± 0.0038 | 0.0231 ± 0.0077 |
| StockNet | 51.93 ± 0.0001 | 0.0335 ± 0.0050 | 50.15 ± 0.0054 | 0.0050 ± 0.0118 | 50.36 ± 0.0095 | 0.0134 ± 0.0135 |
| Adv-ALSTM | 51.69 ± 0.0058 | 0.0333 ± 0.0137 | 51.60 ± 0.0103 | 0.0340 ± 0.0201 | 50.66 ± 0.0067 | 0.0155 ± 0.0140 |
| DTML | 53.53 ± 0.0075 | 0.0733 ± 0.0195 | 52.76 ± 0.0103 | 0.0626 ± 0.0230 | 52.08 ± 0.0121 | 0.0502 ± 0.0214 |
| CausalStock | **56.09 ± 0.0069** | **0.1235 ± 0.0189** | **53.01 ± 0.0150** | **0.0640 ± 0.0310** | **52.88 ± 0.0009** | **0.0534 ± 0.0210** |

**Dataset** (Appendix C.1): We train and evaluate our model and baselines on six datasets: ACL18 [42], CMIN-US [23], CMIN-CN [23], KDD17 [45], NI225 [44], and FTSE100 [44]. The first three of them including both historical prices and text data are used for news-driven multi-stock movement prediction task evaluation, while the last three are for multi-stock movement prediction task evaluation without news data. **Evaluation metrics** (Appendix C.2): We evaluate the prediction performance of models by Accuracy (ACC) and Matthews Correlation Coefficients (MCC). **Baselines** (Appendix C.3): HAN [15], Stocknet [42], PEN [21], CMIN [23] for news-driven multi-stock movement prediction task. LSTM [25], ALSTM [28], Adv-ALSTM [8], DTML [44] for multi-stock movement prediction task. **Parameter setup** (Appendix C.4): Our model is implemented with Pytorch on 4 NVIDIA Tesla V100 and optimized by Adam [20]. The parameter sensitivity study can be found in Appendix C.4.

## 5.2 Results of prediction accuracy

As shown in the top half of Table 1, CausalStock outperforms all baselines on ACC as well as MCC across three datasets demonstrating robustness performance for news-driven multi-stock movement prediction task. For the multi-stock movement prediction task, the results are reported in the bottom half of Table 1. As can be seen, CausalStock exceeds all baselines across three datasets with stable performance. Overall, the results demonstrate that the proposed CausalStock can indeed improve the performance for two stock movement prediction tasks, showing the strong capabilities in handling financial texts and discovering causal relations among stocks.

## 5.3 Ablation study

For the ablation study, we conduct several model variants on ACL18, CMIN-CN and CMIN-US to explore the contributions of different settings in CausalStock. For the main framework, we have the following five variants. **CausalStock w/o TCD**: removing the causal discovery module from CausalStock; **CausalStock w/o News**: removing the news encoder from CausalStock and just taking prices data as input; **CausalStock w/o link non-existence modeling**: only model the causal link existence likelihood and leverage Sigmoid function to obtain the link existence probability; **CausalStock w/o Lag-dependent TCD**: replacing the Lag-dependent Temporal Causal Discovery module with the Lag-independent Temporal Causal Discovery module; **CausalStock with Variable-dependent TCD**: we add a variable-dependent causal mechanism that explicitly captures the dependencies among different stock edges. Specifically, each edge's probability is conditioned on the states of all other edges at the same time step, and the conditional function is the same as the function in the lag-dependent mechanism (Equation 6). Furthermore, we explore the performance of six different Traditional News Encoders by replacing the denoised news encoder, which outputs the news embeddings as representations. **CausalStock with Glove + Bi-GRU**: leveraging the Glove word embedding

Table 2: Ablation study results on different datasets.

| Ablation Type | Ablation Variants | ACL18 | | CMIN-US | | CMIN-CN | |
|---|---|---|---|---|---|---|---|
| | | ACC | MCC | ACC | MCC | ACC | MCC |
| Main Framework | CausalStock w/o TCD | 51.08 | 0.0102 | 51.48 | 0.0106 | 51.37 | 0.0102 |
| | CausalStock w/o news | 58.10 | 0.1421 | 53.16 | 0.0375 | 54.16 | 0.1264 |
| | CausalStock w/o link non-existence | 58.21 | 0.1652 | 52.32 | 0.0241 | 53.96 | 0.0670 |
| | CausalStock w/o Lag-dependent TCD | 59.19 | 0.1757 | 52.93 | 0.0312 | 54.97 | 0.1298 |
| | CausalStock with Variable-dependent TCD | **63.50** | **0.2175** | 54.60 | 0.0479 | **56.25** | **0.1419** |
| Traditional News Encoder | CausalStock with Glove+Bi-GRU | 60.78 | 0.1952 | 53.87 | 0.0467 | 55.13 | 0.1326 |
| | CausalStock with Bert | 61.74 | 0.2067 | 53.92 | 0.0472 | 55.43 | 0.1352 |
| | CausalStock with Roberta | 61.81 | 0.2071 | 54.06 | 0.0477 | 55.58 | 0.1364 |
| | CausalStock with FinBert | 61.72 | 0.2062 | 54.01 | 0.0471 | 55.61 | 0.1362 |
| | CausalStock with FinGPT | 61.69 | 0.2060 | 54.00 | 0.0470 | 55.60 | 0.1360 |
| | CausalStock with Llama | 62.20 | 0.2130 | 54.40 | 0.0480 | 55.85 | 0.1390 |
| Denoised News Encoder | CausalStock with FinGPT | 61.92 | 0.2105 | 54.30 | 0.0475 | 55.67 | 0.1386 |
| | CausalStock with Llama | 62.82 | 0.2164 | 54.52 | **0.0483** | 55.97 | 0.1406 |
| | CausalStock (with GPT-3.5) | 63.42 | 0.2172 | **54.64** | 0.0481 | 56.19 | 0.1417 |

and the Bi-GRU as news encoder [23]; **CausalStock with Bert**: leveraging the pre-trained Bert (Bert-base-multilingual-cased [6]) as news encoder; **CausalStock with Roberta**: leveraging the pre-trained Roberta (Roberta-base [22]) as news encoder; **CausalStock with FinBert**: leveraging the pre-trained FinBert [1] as news encoder; **CausalStock with FinGPT**: leveraging the pre-trained FinGPT (FinGPT-v3.3 [43]) as news encoder; **CausalStock with Llama**: leveraging the pre-trained Llama ( Llama-7b-chat-hf [39]) as news encoder to output news embeddings. Moreover, we explore the performance of three different LLMs for the denoised news encoder. **CausalStock with FinGPT**: leveraging a financial LLM FinGPT (FinGPT-v3.3 [43]) as denoised news encoder; **CausalStock with Llama**: leveraging Llama (Llama-7b-chat-hf [39]) as denoised news encoder. The ablation study results are summarized in Table 2.

We have the following observations: (1) CausalStock with news encoders all perform better than CausalStock without news, suggesting news data is particularly helpful for stock movement prediction. (2) Compared to CausalStock w/o Lag-independent TCD, CausalStock with Lag-dependent TCD has a better performance, demonstrating the value of the lag-dependent mechanism. (3) By comparing the CausalStock and CausalStock with Variable-dependent TCD, the results show that incorporating a variable-dependent causal mechanism has the potential to enhance model performance. However, the improvements are not uniform and vary depending on the dataset, which emphasizes that further validation is needed. While the above results show a promising performance of the variable-dependent causal mechanism, it significantly increases the computational complexity (from $O(L \times D^2)$ to $O(L \times D^4)$), making it challenging to apply the model to markets with large numbers of stocks. (3) By using FinGPT and Llama as the news encoder and denoised news encoder respectively, we can observe that denoised news encoders have a relatively higher ACC and MCC than their as the traditional news encoders, suggesting the value of denoised news encoders. Overall, the ablation studies show that every component contributes to CausalStock.

## 5.4 Results of explainability

Here, we present many cases detailing the interpretability of CausalStock from two perspectives: the news representation from the Denoised News Encoder, and the causal graph discovered by the Lag-dependent TCD module.

Firstly, regarding the Denoised News Encoder module, three cases are selected as shown in Figure 3(c). A piece of news about APPL suggests a potential delay in its 5G iPhone launch, with Denoised News Encoder giving it a negative sentiment score of $-0.7$ and an impact score of $9$. Similarly, a news about TSLA hints at surpassing a significant delivery milestone, receiving a positive sentiment score of $0.7$. In contrast, a news piece showing no discernible connection to GOOG is scored with negligible impact. These cases indicate the Denoised News Encoder's efficacy in discerning and quantifying the potential influence of news on respective stock prices.

Secondly, concerning the causal graph discovered by Lag-dependent TCD, we denote the causal strength graph as the dot product of the causal graph $G$ and the causal weight graph $\hat{G}$. Every item of causal strength graph indicates not only the causality of two stocks but also the degree of causality. The visualized causal strength matrix for ACL18 is shown in Figure 3(b) with a heatmap. From
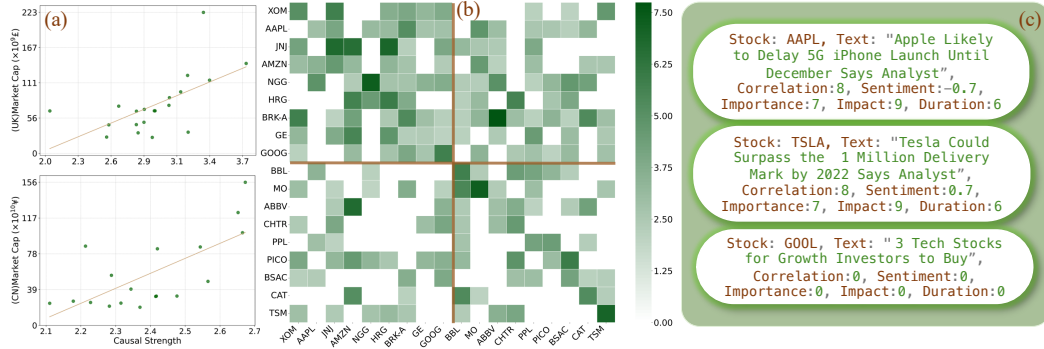
Figure 3: (a) Correlation visualization between market value and causal strength for the top 20 companies of UK and Chinese markets. (b) Partial causal strength matrix visualization for ACL18, encompassing the companies with the highest and lowest market values across various industries. Each matrix entry indicates causal strength between stocks, with darker shades signifying stronger causality. (c) Examples of denoised news encoder module output.

various industries, we select companies with the highest and lowest market value. The top half of Figure 3(b) represents stocks corresponding to the nine companies with the largest market value, while the bottom half illustrates stocks from companies with the smallest. The causal strength of stocks is determined based on the average overall lags. In this heatmap, we could observe that distinct patterns emerge according to different market values. Stocks of low-market-value companies appear to have less pronounced causal relationships. We could also observe causal connections between certain high and low-market-value stocks. This is attributable to the dominant roles of large-value companies with their significant impact on those small-value firms and the stock prices.

Based on these observations, we compute the Spearman's rank correlation coefficient [36] between the aforementioned company's market value and their stock's causal strength on ACL18, CMIN-CN, NI225, and FSTE100 datasets, representing the US, Chinese, Japanese, and UK stock markets respectively. The correlation results are shown in Appendix D and we also visualize some results in Figure 3(a). These results show a strong positive correlation between the market value and causal influence. This aligns with the intuition that not only do large-value companies hold pivotal economic positions, but also play crucial roles in influencing other companies. Our findings demonstrate that CausalStock does well in uncovering the causal relations within the stock market.

## 5.5 Investment simulation

Following prior works [44, 23], we evaluate Causal-Stock's applicability to the real world trading scenario. We conduct a portfolio strategy by choosing the top three stocks (based on predicted probabilities) with equal weight on each day of the test set and calculate the Accumulated Portfolio Value (APV) and Sharpe Ratio (SR) for evaluation. See Appendix C.2 for a metrics details. The results on three datasets are shown in Table 4, which indicates that CausalStock achieves higher profits, and the excellent capabilities of CausalStock to balance risk with returns.

Figure 4: Investment simulation results.

| Model | ACL18 | | KDD17 | | NI225 | |
|---|---|---|---|---|---|---|
| | SR | APV | SR | APV | SR | APV |
| Market Index | 0.107 | 1.07 | 0.056 | 1.10 | 0.080 | 1.18 |
| PEN | 0.293 | 1.12 | 0.132 | 1.39 | 0.171 | 1.43 |
| DTML | 0.304 | 1.11 | 0.157 | 1.39 | 0.184 | 1.42 |
| CMIN | 0.357 | 1.24 | 0.169 | 1.46 | 0.201 | 1.51 |
| CausalStock | **0.369** | **1.32** | **0.192** | **1.49** | **0.259** | **1.52** |

## 6 Conclusion

In this paper, we propose a novel news-driven multi-stock movement prediction framework called CausalStock. We design a lag-dependent temporal causal discovery mechanism to uncover the causal relations among the stocks. Then the functional causal model is employed to encapsulate causal relations and predict future movements. The effectiveness of CausalStock is demonstrated by experiments on multiple real-world datasets. Moreover, CausalStock could offer a clear prediction process with explainability.

## Acknowledgments and Disclosure of Funding

## References

[1] D Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

[2] Alexis Bellot, Kim Branson, and Mihaela van der Schaar. Neural graphical modelling in continuous-time: consistency guarantees and algorithms. *arXiv preprint arXiv:2105.02522*, 2021.

[3] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[4] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[5] Yuxiao Cheng, Ziqian Wang, Tingxiong Xiao, Qin Zhong, Jinli Suo, and Kunlun He. Causaltime: Realistically generated time-series for benchmarking of causal discovery. *arXiv preprint arXiv:2310.01753*, 2023.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Ilia Dichev. News or noise? *Research in Higher Education*, 42(3):237–266, 2001.

[8] Fuli Feng, Huimin Chen, Xiangnan He, Ji Ding, Maosong Sun, and Tat-Seng Chua. Enhancing stock movement prediction with adversarial training. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5843–5849. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[9] Tomas Geffner, Javier Antoran, Adam Foster, Wenbo Gong, Chao Ma, Emre Kiciman, Amit Sharma, Angus Lamb, Martin Kukla, Nick Pawlowski, et al. Deep end-to-end causal inference. *arXiv preprint arXiv:2202.02195*, 2022.

[10] Andreas Gerhardus and Jakob Runge. High-recall causal discovery for autocorrelated time series with latent confounders. *Advances in Neural Information Processing Systems*, 33:12615–12625, 2020.

[11] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[12] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

[13] Jiaying Gong and Hoda Eldardiry. Multi-stage hybrid attentive networks for knowledge-driven stock movement prediction. In Teddy Mantoro, Minho Lee, Media Anugerah Ayu, Kok Wai Wong, and Achmad Nizar Hidayanto, editors, *Neural Information Processing - 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8-12, 2021, Proceedings, Part IV*, volume 13111 of *Lecture Notes in Computer Science*, pages 501–513. Springer, 2021.

[14] Wenbo Gong, Joel Jennings, Cheng Zhang, and Nick Pawlowski. Rhino: Deep causal temporal relationship learning with history-dependent noise. *arXiv preprint arXiv:2210.14706*, 2022.

[15] Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 261–269, 2018.

[16] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[17] Weiwei Jiang. Applications of deep learning in stock market prediction: recent progress. *Expert Systems with Applications*, 184:115537, 2021.

[18] Ilyes Khemakhem, Ricardo Monti, Robert Leech, and Aapo Hyvarinen. Causal autoregressive flows. In *International conference on artificial intelligence and statistics*, pages 3520–3528. PMLR, 2021.

[19] Raehyun Kim, Chan Ho So, Minbyul Jeong, Sanghoon Lee, Jinkyu Kim, and Jaewoo Kang. Hats: A hierarchical graph attention network for stock movement prediction. *arXiv preprint arXiv:1908.07999*, 2019.

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[21] Shuqi Li, Weiheng Liao, Yuhan Chen, and Rui Yan. Pen: Prediction-explanation network to forecast stock price movement with better explainability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5187–5194, 2023.

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[23] Di Luo, Weiheng Liao, Shuqi Li, Xin Cheng, and Rui Yan. Causality-guided multi-memory interaction network for multivariate stock price movement prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12164–12176, 2023.

[24] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

[25] David MQ Nelson, Adriano CM Pereira, and Renato A De Oliveira. Stock market's price movement prediction with lstm neural networks. In *2017 International joint conference on neural networks (IJCNN)*, pages 1419–1426. Ieee, 2017.

[26] Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pages 1595–1605. PMLR, 2020.

[27] Judea Pearl. Causal inference in statistics: An overview. 2009.

[28] Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison Cottrell. A dual-stage attention-based recurrent neural network for time series prediction. *arXiv preprint arXiv:1704.02971*, 2017.

[29] Subhro Roy and Dan Roth. Solving general arithmetic word problems. *arXiv preprint arXiv:1608.01413*, 2016.

[30] Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1388–1397. PMLR, 2020.

[31] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5(11):eaau4996, 2019.

[32] Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Ratn Shah. Deep attentive learning for stock movement prediction from social media text and company correlations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8415–8426, Online, November 2020. Association for Computational Linguistics.

[33] Robert P Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2):1–19, 2009.

[34] Shun-Yao Shih, Fan-Keng Sun, and Hung-yi Lee. Temporal pattern attention for multivariate time series forecasting. *Machine Learning*, 108:1421–1441, 2019.

[35] Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. Exploiting topic based twitter sentiment for stock prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 24–29, 2013.

[36] Charles Spearman. The proof and measurement of association between two things. 1961.

[37] Timm O. Sprenger, Philipp G. Sandner, Andranik Tumasjan, and Isabell M. Welpe. News or noise? using twitter to identify and understand company-specific news flow. *Social Science Electronic Publishing*, 41(7-8):791–830, 2014.

[38] Timm O. Sprenger and Isabell M. Welpe. News or noise? the stock market reaction to different types of company-specific news events. *SSRN Electronic Journal*, 2001.

[39] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timo-thée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[40] Guifeng Wang, Longbing Cao, Hongke Zhao, Qi Liu, and Enhong Chen. Coupling macro-sector-micro financial indicators for learning stock representations with less uncertainty. *AAAI21*, pages 1–9, 2021.

[41] Frank Z Xing, Erik Cambria, and Roy E Welsch. Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1):49–73, 2018.

[42] Yumo Xu and Shay B. Cohen. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[43] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*, 2023.

[44] Jaemin Yoo, Yejun Soun, Yong-chan Park, and U Kang. Accurate multivariate stock movement prediction via data-axis transformer with multi-level contexts. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2037–2045, 2021.

[45] Liheng Zhang, Charu Aggarwal, and Guo-Jun Qi. Stock price prediction via discovering multi-frequency trading patterns. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2141–2149, 2017.

[46] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

[47] Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. PMLR, 2020.

# A  Prompt design

This structured prompt encompasses three fundamental components:

System: This section defines the role of the AI. It acts as a preliminary introduction to set the tone and context for the AI. It informs the AI that its primary role is to analyze stock-related news in various dimensions such as correlation, sentiment, importance, impact on prices, and duration of impact.

Default Prompt: This segment provides detailed instructions to the AI on how to carry out its analysis. It outlines the specific criteria and the scales on which the news should be evaluated. It also provides guidance on how to handle ambiguous or non-analyzable content and finally, it prescribes the desired output format.

Input: The final section is where the user provides the specific details about the stock, the news content, and the time of publication. It acts as the data point based on which the AI will perform its analysis as instructed in the Default Prompt.

[System]

{As a stock trading news analyst, you are a helpful and precise assistant. Your task is to analyze the correlation between news and the given stock, sentiment polarity of the news, importance of the news, the impact of the news on stock prices, and the duration of the news impact.}

[Default Prompt]

I need you to analyze the provided stock-related news from five dimensions:

1. Correlation between the news and the given stock: Rate the correlation on a scale of 0 to 10, where a higher score indicates a stronger correlation between the news and the given stock.

2. Sentiment polarity of the news: Rate the sentiment polarity on a scale of -1 to 1, where a value closer to -1 indicates stronger negative sentiment and a value closer to 1 indicates stronger positive sentiment.

3. Importance of the news event: Rate the importance on a scale of 0 to 10, where a higher score indicates higher importance of the news event.

4. Impact of the news on stock prices: Rate the impact on a scale of 0 to 10, where a higher score indicates a greater impact of the news on stock prices.

5. Duration of the news impact: Rate the duration on a scale of 0 to 10, where a higher score indicates a longer potential duration of the news impact.

(When you encounter a situation where analysis is not possible, please try to avoid assigning all-zero scores and instead make an effort to analyze the text content and derive scores accordingly. Only when analysis is truly impossible should you assign a score of 0 to all factors.)

(Please refrain from providing analysis and simply provide the answer according to the following format.)

Output format:

Correlation: <Correlation score between the news and the stock>

Sentiment: <Sentiment polarity score of the news>

Importance: <Importance score of the news event>

Impact: <Impact score of the news on stock prices>

Duration: <Duration score of the news impact>

[Input]

[Stock Name]: {*stock name*}

[News Content]:{*news content*}

[Publish Time]:{*publish time*}

## B  Assumptions and theoretical guarantees

There are some common assumptions in causal discovery. In this paper, we assume our model satisfies the *Causal Markov Property*, *Minimality and Structural Identifiability*, *Correct Specification*, *Causal Sufficiency* and *Regularity of log likelihood*. A detailed explanation can be found in [9], which explains how our model satisfies these assumptions. These assumptions guarantee the validity of the causal relations discovered by CausalStock. Considering the instability of news data, we only leverage price data $P_{<T}$ to discover causal graph $G$ to meet the *Causal Stationary* assumption. Then we use the discovered causal graph $G$ for aggregating both news information and price information. Technically, this could be realized by detaching the gradient from $C_{<T}$ to $G$.

## C  Datasets & metrics & baselines & parameter setting

### C.1  Dataset

Six datasets from various countries' stock markets are employed for conducting the experiments. The first three are used for models of fundamental analysis, which include both historical prices and text data. ACL18 [42] is a collection of data from 88 stocks in 9 industries in the US market over two years. Specifically, the price vectors after preprocessing are made up of 7 entries: date, movement percent, open price, high price, low price, close price, and volume, and the text data from Twitter are treated with tokenization and cleaning. Two CMIN datasets [23] are published subsequently following a similar format as ACL18. CMIN-US is collected from the US market, whereas CMIN-CN comes from 300 CSI300 stocks in the Chinese market.

The other three datasets contain historical prices only and are applied to methods of technical analysis. KDD17 [45] collects prices of 50 US stocks. [8] proposed to transfer the raw price vectors of KDD17 into 11 temporal features for normalizing prices and capturing the interaction between different raw price entries. With this transfer calculation, NI225, and FTSE100 [44] record 11-feature stock prices from the US, China, Japan, and UK market respectively over the different time periods. For all datasets, the train-test set split is chronological. More detailed statistics about the datasets are presented in Table 3 below.

Table 3: Dataset Description

| Dataset | Country | Stock | Data Range | | | Data Resource | | Price Dim |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | | | Train | Valid | Test | Price | Text | |
| ACL18 [1] | US | 88 | 2014/01/02-2015/08/02 | 2015/08/03-2015/09/30 | 2015/10/01-2016/01/01 | Yahoo Finance | Twitter | 7 |
| CMIN-US [2] | US | 110 | 2018/01/01-2021/04/30 | 2021/05/01-2021/08/31 | 2021/09/01-2021/12/31 | Yahoo Finance | Yahoo | 7 |
| CMIN-CN [2] | CN | 300 | 2018/01/01-2021/04/30 | 2021/05/01-2021/08/31 | 2021/09/01-2021/12/31 | Yahoo Finance | Wind | 7 |
| KDD17 [3] | US | 50 | 2007/01/03-2015/01/01 | 2015/01/02-2016/01/03 | 2016/01/04-2017/01/01 | Yahoo Finance | - | 11 |
| NI225 [4] | JP | 51 | 2016/07/01-2018/03/01 | 2018/03/02-2019/01/06 | 2019/01/07-2019/12/31 | Yahoo Finance | - | 11 |
| FTSE100 [4] | UK | 24 | 2014/01/06-2017/01/03 | 2017/01/04-2017/07/03 | 2017/07/04-2018/06/30 | - | - | 11 |

### C.2  Metrics

Given the confusion matrix ( $\begin{smallmatrix} tp & fn \\ fp & tn \end{smallmatrix}$ ), where $tp, fp, tn, fn$ represent the true positives, false positives, true negatives and false negatives, we calculate ACC and MCC as follows:

$$ACC = \frac{tp + tn}{tp + tn + fp + gn}, \tag{15}$$

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(fn + tp)(fn + tn)(fp + tn)}}. \tag{16}$$

---

[1] https://github.com/yumoxu/stocknet-dataset

[2] https://github.com/BigRoddy/CMIN-Dataset

[3] https://github.com/fulifeng/Adv-ALSTM

[4] https://datalab.snu.ac.kr/dtml

Accumulated investment portfolio value (APV) shows the accumulation of wealth over time in an intuitive form and the Sharpe Ratio (SR) is probably the most widely used metric to measure a trading strategy's return compared to its risk. The Sharpe ratio is calculated as follows,

$$\text{APV}^t = \prod_{i=1}^{t}(1 + r^i), \tag{17}$$

$$\text{SR} = \frac{\mathbb{E}\left[\text{APV}^t - R_f\right]}{\mathbb{S}\left[\text{APV}^t - R_f\right]}, \tag{18}$$

where $r^i$ is the daily return ratio on $i$-th trading day and $R_f$ is the risk-free return.

### C.3 Baselines

**For multi-stock movement prediction**

- **LSTM** [25] is an LSTM-based network that is trained with a rolling window of the last 10 months. 175 technical indicators on the characteristic of stocks and 5 features on normalized historical prices jointly form the input layer and are fed into the model.

- **ALSTM** [28] uses attentive LSTM in both encoder and decoder. The input attention mechanism in the encoder could extract the relevant features of stock price, whereas the temporal attention mechanism in the decoder could help learn the long-term dependencies.

- **Adv-LSTM** [8] tries to improve ALSTM through adversarial training to capture the stochastic nature of stock price and ameliorate the over-fitting. During the training process, adversarial examples are generated from latent representation and integrated with clean samples to serve as input.

- **DTML** [44] exploits the correlations between stocks in three parts: compressing the multivariate historical prices of a stock into a context vector with attentive LSTM, generating multi-level context vectors by aggregating local and global context, and capturing the correlations between stocks via transformer encoder and self-attention.

**For news-driven multi-stock movement prediction**

- **HAN** [15]: uses attention mechanism to select useful news for stock movement prediction from chaotic online resources. The framework first applies news-level attention to find out more significant news in a date and encodes the output corpus vectors with Bi-GRU. Then, another temporal attention is applied to focus on more impactful time periods.

- **Stocknet** [42]: predicts stock trend based on text and price with recurrent, continuous latent variables. The model has 3 modules, which are Market Information Encoder (MIE), Variational Movement Decoder (VMD), and Attentive Temporal Auxiliary (ATA) in sequence.

- **PEN** [21]: a model that fuses the Bi-GRU text embedding and price inputs into Shared Representation Learning (SRL) to study their interaction. SRL also yields a Vector of Salient (VOS) that could display the importance of a piece of news and display the explainability of the model.

- **CMIN** [23]: integrates causality-enhanced stock correlations and text for stock movement prediction. The approach aims to cover not only the asymmetric correlations between stocks via a newly proposed causal attention mechanism but also the multi-directional interactions between text and stock correlations. In addition, two memory networks are used for selecting the relevant information in text and stock correlation.

### C.4 Parameter setting

Our model is implemented with Pytorch on 4 NVIDIA Tesla V100 and optimized by Adam [20]. All parameters of our model are initialized with Xavier Initialization [11].To better explore the model's performance, we use grid search to decide on many key hyper-parameters. The learning rate is set as $1e-5$ selected from $[1e-3, 1e-4, 1e-5, 1e-6]$. The time lag $L$ is set as $5$ selected from

$[3, 5, 7, 9]$. We select the price encoder hidden size from $[4, 8, 16]$ and get the best performance with size 4. The batch size is set as 32. The scalar weight $\lambda$ is set to 0.01. For the traditional news encoder, the maximum word number in one piece of news and news number in one day are set to $w = 20, l = 10$, respectively. The embedding size of word and news are set to $d_w = 50, d_m = 64$, respectively. For the Lag-dependent temporal causal discovery module, $\lambda_s = 1$, $h_v$ and $h_u$ are all 1-layer MLPs. For the FCM part, the neural modules $\zeta_i$, $\ell$ and $\psi$ are all 3-layer MLPs with hidden size 332.

**Hyper-parameter sensitivity study**   We take a further step to analyze the main parameter sensitivity of CausalStock. We tune the key hyper-parameters learning rate $lr$, maximum time lag $L$ and loss weight $\lambda$ by grid search from this combination $lr = 1e - 5, L = 5, \lambda = 0.01$ while controlling other parameters. Table 4 presents the results of metric ACC with different parameter settings on two tasks.

Table 4: Hyper-parameter sensitivity study results.

| Parameters | Learning rate $lr$ | | | | Time lag $L$ | | | | Loss weight $\lambda$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1e-3 | 1e-4 | **1e-5** | 1e-6 | 3 | **5** | 7 | 9 | 0 | 0.1 | **0.01** | 0.001 |
| **ACL18 (with news)** | 62.56 | 62.34 | **63.42** | 61.58 | 61.04 | **63.42** | 63.29 | 63.15 | 58.26 | 62.35 | **63.42** | 63.45 |
| **KDD17 (w/o news)** | 55.45 | 55.69 | **56.09** | 55.13 | 54.94 | **56.09** | 55.95 | 55.94 | 53.19 | 55.57 | **56.09** | 55.45 |

## D   The correlation results

Table 5: The correlation of the causal strength and the market value of companies on four datasets.

| Statistics | ACL 18 | NI225 | CMIN-CN | FTSE100 |
|---|---|---|---|---|
| **Spearman Corr.** | 0.7939 | 0.7212 | 0.6491 | 0.8909 |
| **P-Value** | 0.006 | 0.0185 | 0.0036 | 0.0005 |

## E   Limitations and future works

This paper explores a method that discovers causal relations based on theoretical considerations. In the future, we could try to adopt meta-learning or incremental learning training methods to update the causal graph iteratively, i.e. explore the time-varied causal graph. While the Bernoulli distribution is suitable for determining whether a causal link exists, if we want to further explore the multi-level nature of causal relationships, more complex distributions might be needed. In the future, we could improve the model in this way.

## F   Broader impacts and safety issues

In this paper, we designed an LLM-based Denoised News Encoder to evaluate the news from multiple perspectives by LLMs. There exists a risk that the evaluation results of LLMs may violate human values. This safety issue needs careful consideration.

# NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We propose a novel stock movement prediction framework based on the causal discovery method and LLMs, which are all introduced in detail in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We create a separate "Limitations" section, see Appendix E.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide the full proof and the assumptions in Section B and Section 4.5.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We include all the details of the implementing process.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We release code and data in GitHub.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: See Section 5 and Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report the standard error in the result tables 1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Section 5.

   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have conformed the NeurIPS Code of Ethics.

   Guidelines:
   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Section F.

    Guidelines:
    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: See Section F.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make the best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have correctly cited all the data, scripts, and models we used.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We have a READEME Document for our code.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowdsourcing experiments.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing experiments.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.