CAMEF: Causal-Augmented Multi-Modality Event-Driven Financial Forecasting by Integrating Time Series Patterns and Salient Macroeconomic Announcements

Yang Zhang[*]
Southwestern University of Finance
and Economics
Chengdu, China
zhang.yang.r54@kyoto-u.jp

Wenbo Yang
Southwestern University of Finance
and Economics
Chengdu, China
223081200030@smail.swufe.edu.cn

Jun Wang[*]
Southwestern University of Finance
and Economics
Chengdu, China
wangjun1987@swufe.edu.cn

Qiang Ma
Kyoto Institute of Technology
Kyoto, Japan
qinag@kit.ac.jp

Jie Xiong
Southwestern University of Finance
and Economics
Chengdu, China
xiongjie@swufe.edu.cn

Abstract

Accurately forecasting the impact of macroeconomic events is critical for investors and policymakers. Salient events like monetary policy decisions and employment reports often trigger market movements by shaping expectations of economic growth and risk, thereby establishing causal relationships between events and market behavior. Existing forecasting methods typically focus either on textual analysis or time-series modeling, but fail to capture the multi-modal nature of financial markets and the causal relationship between events and price movements. To address these gaps, we propose **CAMEF** (Causal-Augmented Multi-Modality Event-Driven Financial Forecasting), a multi-modality framework that effectively integrates textual and time-series data with a causal learning mechanism and an LLM-based counterfactual event augmentation technique for causal-enhanced financial forecasting. Our contributions include: (1) a multi-modal framework that captures causal relationships between policy texts and historical price data; (2) a new financial dataset with six types of macroeconomic releases from 2008 to April 2024, and high-frequency real trading data for five key U.S. financial assets; and (3) an LLM-based counterfactual event augmentation strategy. We compare CAMEF to state-of-the-art transformer-based time-series and multi-modal baselines, and perform ablation studies to validate the effectiveness of the causal learning mechanism and event types.

CCS Concepts

• **Applied computing** → *Economics*; • **Computing methodologies** → **Neural networks**.

Keywords

Multimodal learning, Causal Learning, Financial dataset, Time-series Forecasting

ACM Reference Format:
Yang Zhang, Wenbo Yang, Jun Wang, Qiang Ma, and Jie Xiong. 2025. CAMEF: Causal-Augmented Multi-Modality Event-Driven Financial Forecasting by Integrating Time Series Patterns and Salient Macroeconomic Announcements. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3711896.3736872

1 Introduction

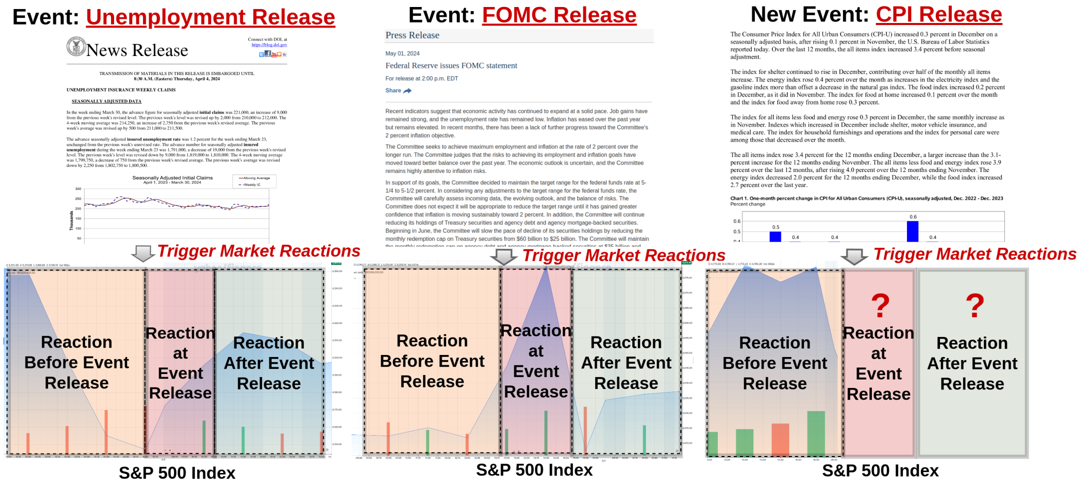
The prices of financial assets reflect all available information, according to Fama's Efficient Market Theory [15, 16]. Major financial releases from government sectors often trigger market movements by shaping investors' expectations and evaluations of economic conditions, asset growth potential, and associated risks. For example, during the FOMC meeting on March 16, 2020, the Fed's emergency rate cut to 0-0.25% sharply altered investors' economic outlook, resulting in a massive sell-off. Major indices, including the S&P 500, NASDAQ, and Dow Jones, dropped by over 10%, marking the steepest single-day decline since 1987 [13]. These salient macroeconomic events cause reactions in financial assets, establishing causal relationships between events and financial assets. Figure 1 illustrates multiple types of events that cause financial market reactions. Therefore, **accurately forecasting the causal consequences of the salient macroeconomic releases on financial market is essential, not only to help investors manage risks and maximize returns, but also to provide policymakers with valuable insights for evaluating and refining future policies.**

Previous studies on event-driven forecasting have primarily adopted three lines of methodologies. The first line of approaches utilizes text feature-based models, where language models, ranging from self-crafted RNN-based architectures [11, 22, 29, 57] to pre-trained transformers [49, 62], embed sentiment information into text vectors, and then stock movements are predicted as a binary classification task (e.g., hawkish vs. dovish). The second line of methodology focuses on historical time-series data, treating

[*] Corresponding authors.

KDD '25, August 3–7, 2025, Toronto, ON, Canada

ACM ISBN 979-8-4007-1454-2/2025/08
https://doi.org/10.1145/3711896.3736872


Event: Unemployment Release
News Release
Event: FOMC Release
Press Release
Federal Reserve issues FOMC statement
New Event: CPI Release
Trigger Market Reactions
Trigger Market Reactions
Trigger Market Reactions
Reaction Before Event Release
Reaction at Event Release
Reaction After Event Release
Reaction Before Event Release
Reaction at Event Release
Reaction After Event Release
Reaction Before Event Release
?
Reaction at Event Release
?
Reaction After Event Release
S&P 500 Index
S&P 500 Index
S&P 500 Index

(a) SP500 Index Reaction to Unemployment Release

(b) SP500 Index Reaction to FOMC Meeting

(c) Prediting the SP500 Index Reaction to new event

Figure 1: Event-Driven Forecasting Examples: (a) Market reaction to employment insurance release; (b) Market reaction during FOMC meeting; (c) Forecasting market reactions to future events.

stock price movements as a regression problem [7, 54]. Recently, transformer-based architectures have been applied for time-series prediction, including Informer [60], FedFormer [61], and AutoFormer [8], etc. However, both of these directions typically focus on a single modality, neglecting multi-dimensional information. The third line of research adopts a multi-modality approach, leveraging multiple types of data sources to enhance forecasting performance. For instance, studies like [35, 37] incorporate textual, video, and audio data from FOMC meetings alongside corresponding market movements. While these approaches show promise for event-driven financial forecasts, they face three major limitations:

- **Data Limitation:** Existing approaches predominantly focus on a single type of event, such as FOMC meetings [35, 37, 49], while neglecting other crucial macroeconomic events like unemployment insurance releases, CPI, PPI, and GDP advance reports. Additionally, many studies rely on daily-based time-series data for financial assets [7, 8, 35, 37, 49, 54, 60, 61], which limits their applicability and precision in real-time trading scenarios where high-frequency data is mostly adopted.
- **Modality Limitation:** Most prior studies rely on single-modality analysis, using either textual models [11, 22, 29, 49, 57, 62] or time-series models [7, 8, 54, 60, 61], which fail to integrate the complementary strengths of both modalities. While some multi-modality approaches have been proposed [35, 37], they often lack advanced mechanisms for feature fusion, effective decoding strategies, and causal learning, which are critical for understanding the complex interplay between event texts and market dynamics.
- **Causality Limitation:** Existing methods [35, 37] fail to incorporate causal reasoning frameworks, overlooking the causal relationships between events and market reactions. Without explicitly modeling these relationships, such approaches cannot fully capture the drivers of financial market behavior, limiting their predictive robustness.

To address the limitations of previous studies, we propose a novel multi-modality framework, **CAMEF**[1] (**c**ausal-**A**ugmented **M**ulti-Modality **E**vent-Driven Financial **F**orecasting). CAMEF integrates time-series and textual features through specially designed multi-feature fusion techniques, time-series decoding mechanisms, and causal learning strategies. By conducting a thorough review of financial literature, we identify six types of salient macroeconomic events for the forecasting analysis. Furthermore, the framework employs causal data augmentation powered by Large Language Models (LLMs) and a causal contrastive learning approach to enhance the causal understanding and forecast accuracy of CAMEF. This paper offers three key contributions:

- **Novel Dataset:** We introduce a novel open-source synthetic dataset comprising 6 types of macroeconomic event scripts (ref to Tab. 1 for details) from 2008 to April 2024 through reviewing from financial literature [3, 10, 17–20, 26, 33, 34, 38, 42–44, 52], alongside intra-day _high-frequency financial data at 5-minute intervals from key U.S. stock indexes and Treasury bonds. To support causal learning, the dataset also includes counterfactual event scripts generated using our LLM-based causal argumentation prompting, making it the first to integrate policy texts, high-frequency trading data, and causally augmented content.
- **Novel Multi-Modality Model:** We propose a novel multi-modality approach, CAMEF, that integrates time-series and textual features, incorporating specifically designed multi-feature fusion and time-series decoding networks, which

[1] *The dataset and code for CAMEF are open-sourced at: https://github.com/lakebodhi/CAMEF*

have been demonstrated to be effective for forecasting. Additionally, the model includes a causal learning mechanism to enhance forecasting capability by capturing the causal relationships between events and market reactions.

- **Counterfactual Generation and Learning:** We introduce a counterfactual data augmentation strategy to generate counterfactual event scripts based on collected macroeconomic releases. This approach leverages LLMs to create scripts with varying sentiment levels by modifying key numerical values and sentiment-relevant phrases, while preserving the original format, writing style, and neutral words of the factual reports. Counterfactual events enable CAMEF to better understand the causal relationships between events and market reactions by learning from hypothetical scenarios, thereby improving its forecasting ability.

2 Related Work

2.1 Event-Drive Financial Forecasting

Event-driven financial forecasting [2] focuses on predicting asset prices [18, 20] and market volatility [10, 33] based on events like macroeconomic releases [18], news [29], corporate announcements [62], and social media activity [57]. Three main approaches exist in this area. The first leverages text analysis to predict asset responses based on event-related text. Early works utilized TF-IDF [28, 39] and topic models [36, 50], progressing to RNN-based models [22, 29] and pre-trained transformers [49, 62], which capture nuanced semantics. Although these models excel at semantic extraction, they often lack integration with historical price data, crucial for holistic forecasting.

The second line of approaches uses statistical and sequential models on numerical data, such as linear regression [5], ARIMA [1], and GARCH [23]. Later, deep learning methods like RNNs [29] and CNNs [14, 48] enhanced nonlinear modeling capabilities. More recently, transformer-based models, such as Informer [60] and FedFormer [61], improved long-range dependency modeling for time series data. However, these models tend to be "case-specific," requiring task-specific training. In contrast, the lastest pre-trained models for time-series data, like MOMENT [21], Timer [32], and TOKEN [53], offer more generalized and adaptable solutions for time-series tasks.

The third line of research adopts multi-modality approaches, combining diverse data types to improve forecasting accuracy. Some studies incorporate text and audio [40, 58] but often overlook time-series dependencies. Recent work has integrated time-series and textual data; for example, [46, 47] employed SVM and GRU models to capture time-series features. However, these models are relatively shallow for extracting complex patterns. More recent studies [24, 27] leverage transformer-based models for time-series analysis, better capturing deeper temporal structures. Building on these advancements, this paper aims to utilize state-of-the-art pre-trained models with enhanced feature fusion and causal learning for multi-modality forecasting.

2.2 Salient Macroeconomic Factors

Which macroeconomic announcements have a greater impact on financial markets than others? This question has been widely studied in the financial literature, with Central Bank Communications standing out as the most-researched factor [10, 17, 33, 34, 42–44, 52]. Beyond central bank communications, various other macroeconomic factors have also been identified as significant drivers of market movements. Among these, Non-farm Payrolls, Unemployment Releases, Initial Unemployment Claims, ISM Manufacturing Index, GDP Advance Releases, Consumer Confidence Index, and Producer Price Index (PPI) Reports have been found to notably influence price movements and market volatility through empirical statistical testings [3, 18–20, 26, 38]. In this paper, we aim to leaverage the most significant factors evidented by the past financial literaure [3, 18–20, 26, 38], which include FOMC Meeting Documents, Non-farm Payrolls, Unemployment Releases, Initial Unemployment Claims, ISM Manufacturing Index, GDP Advance Releases, Consumer Confidence Index, and Producer Price Index (PPI) Reports.

2.3 Counterfactual Data Augmentation by LLMs

Counterfactual Data Augmentation seeks to reduce spurious correlations and enhance model robustness. Kaushik et al. [25] introduced a method that augments training data with counterfactuals written by human annotators, effectively helping to mitigate spurious patterns. Ross et al. [45], Wu et al. [56] later proposed the use of hand-crafted templates and trained text generators to create counterfactual data through predefined perturbation types. However, these methods are limited by their reliance on fixed perturbations. More recently, Chen et al. [9], Wang et al. [55] proposed more flexible, LLM-based approaches that leverage specifically designed in-context learning prompts and generation pipelines for counterfactual and instruction data generation. Following this direction, we present a counterfactual generation framework specifically designed for macroeconomic releases.

3 Problem Formulation

Event Set is defined as $\mathbb{E} := \{\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_{|\mathbb{E}|}\}$, where $\mathbb{E}$ represents a collection of $|\mathbb{E}|$ event scripts. Each event $\mathcal{E}_i$ occurs at a specific timestamp i and belongs to one of the event types shown in Table 1A. Each event script $\mathcal{E}_i$ consists of a sequence of word tokens, represented as $\mathcal{E}_i := \{w_1, w_2, \ldots, w_m\}$.

Time Series Data is defined as $\mathcal{X} := \{X_1, X_2, \ldots, X_{|\mathcal{X}|}\}$, where each X_i represents the numerical data at time step i. An event $\mathcal{E}_i$ is **aligned** with a time series segment $\mathcal{X}_{i-\tau:i+\tau}$, where i denotes the time of the releasement of the event, and τ represents the duration of the time-series segment both preceding and succeeding time step i, denoted as $[\mathcal{X}_{i-\tau:i+\tau} \mapsto \mathcal{E}_i]$. This alignment reflects the time series segment leading up to the event ($\mathcal{X}_{i-\tau}$) and the period during which the event is expected to have an effect ($\mathcal{X}_{i+\tau}$).

Event-Driven Forecasting: Given a dataset $\mathcal{U} = \{[\mathcal{X}_{i-\tau:i+\tau} \mapsto \mathcal{E}_i]\}_{i=1}^{n}$ consisting of n aligned event and time-series pairs, for each data pair in $\mathcal{U}$, the model uses both the event text $\mathcal{E}_i$ and the historical time-series segment $\mathcal{X}_{i-\tau:i}$, which spans τ steps before the event's release at time i, to forecast the future time steps $\mathcal{X}_{i+1:i+\tau}$.

Causal Effect Graph: A Causal Effect Graph represents the causal links among variables: textual modality $\mathcal{E}$ (event scripts), current price trend $\mathcal{X}$, and future time series movements Y. In event-driven financial prediction, events influence the market movements

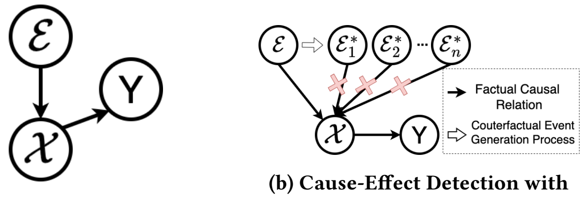
Table 1: Summary of Macroeconomic and Time-Series Data Types, Characteristics and Sources

A. Macroeconomic Event Summary

Event Type	Data Type	Frequency	Period	No. of Events	No. of C.F.s	Source
FOMC	Html	Quarterly	1993.3 ~2024.6	255	2,550	www.federal reserve.gov
Unemployment Insurance Claims	PDF, Txt	Weekly	2002.10 ~2024.6	913	9,130	oui.doleta.gov
Employment Situation	Html, Txt	Monthly	1994.2 ~2024.6	363	3,630	www.bls.gov
GDP Advance Report	Html	Monthly	1996.8 ~2024.6	333	3,330	www.bea.gov
CPI Report	Html, Txt	Monthly	1994.2 ~2024.6	357	3,570	www.bls.gov
PPI Report	Html, Txt	Monthly	1994.2 ~2024.6	348	3,480	www.bls.gov

B. Time-Series Data Summary

Time Series Data	Data Types	Frequency	Range	No. of Data Points
SP500 (SPX)	Open, Close, High, Low	5 Min	2008.01 ~2024.06	331,257
Dow Industrial (INDU)	Open, Close, High, Low	5 Min	2012.07 ~2024.06	263,445
NASDAQ (NDX)	Open, Close, High, Low	5 Min	2008.01 ~2024.06	332,616
US Treasury Bond at 1-Month (USGG1M)	Open, Close, High, Low	5 Min	2013.01 ~2024.06	751,443
US Treasury Bond at 5-Year (USGG5YR)	Open, Close, High, Low	5 Min	2013.01 ~2024.06	734,773



(a) Causal Effect Graph

(b) Cause-Effect Detection with Counterfactual Events

Figure 2: The illustration of causal relationships and counterfactual events.

of financial assets, forming a causal chain denoted as $\mathcal{E} \to \mathcal{X} \to$ Y, as illustrated in Figure 2a.

Counterfactual Event: A Counterfactual Event (CE) represents a modified event script in which key variables (e.g., unemployment rates, GDP values) are altered relative to the factual event, while the surrounding context remains unchanged. These events are denoted as $\{\mathcal{E}_1^*, \mathcal{E}_2^*, \ldots, \mathcal{E}_n^*\}$. CEs are utilized to train CAMEF, enabling it to identify factual cause-effect relationships, represented as $\mathcal{E} \to \mathcal{X} \to \mathcal{Y}$, as illustrated in Figure 2b.

4 Data Collection and Counterfactual Event Augmentation

This section introduces the proposed dataset and the methodology for counterfactual event augmentation. The dataset includes 6 types of key macroeconomic announcements ranging from 2004 to 2024, selected through an extensive review of the financial literature, along with _high-frequency trading data. Unlike the daily-based trading data used in previous studies, this high-frequency data provides more predictive accuracy and better reflects real trading behavior in the industry.

4.1 Dataset Acquisition

The primary question guiding the collection of this dataset is: **"Which macroeconomic releases have the greatest impact on financial markets?"** To address this, we conducted a comprehensive review of the financial literature to identify key macroeconomic factors that influence market behavior. Several dominant factors emerged, including the **FOMC Minutes** [10, 17, 33, 34, 42–44, 52], along with **Unemployment Insurance Claims**, **Employment Situation Reports**, **GDP Advance Releases**, and the **Consumer Price Index (CPI)** and **Producer Price Index (PPI)** reports [3, 18–20, 26, 38], which serve as the textual modality data for our dataset.

To collect these data, we developed web crawlers to extract raw files directly from official sources, including HTML, PDF, and TXT formats. These raw files were then pre-processed and converted into a structured and unified text format, ensuring consistency and ease of subsequent analysis. Table 1 provides a summary of the data types, collection frequencies, time periods, and sources of the events included in our dataset. Further details on the data crawling and pre-processing methodologies can be found in Appendix B.

In addition to the textual data, studies [10, 17, 33, 34, 42–44, 52] have demonstrated that the largest market impacts are typically observed in major U.S. stock indexes and Treasury bonds. Therefore, we focused on collecting high-frequency trading time-series data at 5 minute interval for key stock indexes, including the **S&P 500 (SPX)**, **Dow Industrial (INDU)**, **NASDAQ (NDX)**, as well as **U.S. Treasury Bond at 1-Month (USGG1M)** and **Treasury Bond at 5-Year (USGG5YR)**.

4.2 Counterfactual Events Generation based on LLM

This section describes the process of counterfactual event generation, creating hypothetical scenarios from existing event scripts. The aim is to reflect a target sentiment of a given event script while maintaining logical consistency and coherence of the original script. Our goal is modifying sentiment-relevant elements (such as key facts, sentiment-indicative phrases, or numerical values) without disrupting the sentiment-neutral components of the script.

Formally, for a given event script $\mathcal{E}_i := \{w_1, w_2, \ldots, w_m\}$, the objective is to produce a counterfactual version $\mathcal{E}_i'$ that embodies the desired target sentiment S_i'. Conceptually, the event script can be viewed as comprising sentiment-relevant content ($\mathcal{E}_i^{\text{sentiment}}$) and sentiment-neutral content ($\mathcal{E}_i^{\text{neutral}}$), so that $\mathcal{E}_i = \mathcal{E}_i^{\text{neutral}} \cup \mathcal{E}_i^{\text{sentiment}}$. Instead of explicitly decomposing the script, we guide a language model (LLM) using structured prompts to modify only the sentiment-relevant content. This ensures that neutral content remains intact or is replaced with semantically equivalent expressions. Formally:

$$\mathcal{E}_i' = \mathcal{E}_i^{\text{neutral}} \cup f_{\text{LLM}}(\mathcal{E}_i^{\text{sentiment}} \mid S_i'), \tag{1}$$

where f_{LLM} adjusts $\mathcal{E}_i^{\text{sentiment}}$ to align with S_i', and $\mathcal{E}_i^{\text{neutral}}$ remains unchanged or is replaced by equivalent expressions.

Specifically, we used the LLaMA-3 8B model with a series of carefully designed prompts. These prompts include three key steps, with detailed templates provided in Appendix A:

(1) **Summarization Prompt (Appendix A.1):** Condenses lengthy event scripts into concise summaries, addressing memory

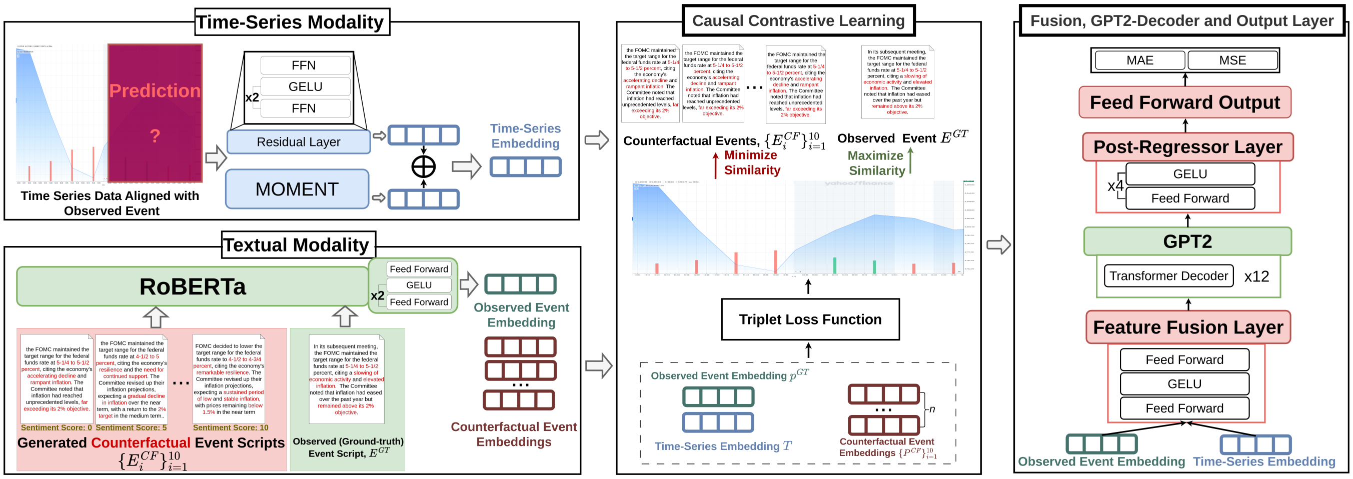


Figure 3: The Pipeline and Neural Architecture of CAMEF

constraints while retaining sentiment-relevant content and key numerical variables.

(2) **Sentiment Analysis Prompt (Appendix A.2):** Assigns a sentiment score (from 1, very negative, to 10, very positive) to the original event script. This score provides a baseline for generating counterfactual versions.

(3) **Counterfactual Generation Prompt (Appendix A.3):** Produces multiple counterfactual scripts, each reflecting a different sentiment level. The prompt modifies sentiment-related phrases and numerical values ($\mathcal{E}_i^{\text{sentiment}}$) while preserving or equivalently substituting neutral content ($\mathcal{E}_i^{\text{neutral}}$). This approach ensures numerical reasonability, sentiment relevance, and structural consistency.

This multi-step prompt strategy facilitates the generation of coherent, contextually relevant counterfactual events, enabling exploration of diverse market scenarios and deeper causal understanding.

5 CAMEF Architecture

The CAMEF model integrates both textual and time-series information through a structured architecture consisting of a textual encoder, a time-series encoder, and a forecasting decoder, as depicted in Fig. 3. Each component is detailed below.

5.1 Textual Modality Encoder (CAMEF$_{\text{Textual}}$)

We encode event scripts using RoBERTa [31]. Given an input script $\mathbf{E}_i = w_1, w_2, \ldots, w_m$, RoBERTa produces contextual token embeddings:

$$\{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_m\} = \text{RoBERTa}(\{w_1, w_2, \ldots, w_m\}), \tag{2}$$

where $\mathbf{h}_j \in \mathbb{R}^{1\times768}$ is the embedding of token w_j. Each embedding is passed through a projection network with three linear layers and GELU activations:

$$\mathbf{e}_j = \mathbf{W}^{(3)} \cdot \text{GELU}(\mathbf{W}^{(2)} \cdot \text{GELU}(\mathbf{W}^{(1)}\mathbf{h}_j + \mathbf{b}^{(1)}) + \mathbf{b}^{(2)}) + \mathbf{b}^{(3)}, \tag{3}$$

where $\mathbf{W}^{(1)} \in \mathbb{R}^{768\times1024}$, $\mathbf{W}^{(2)} \in \mathbb{R}^{1024\times1024}$, and $\mathbf{W}^{(3)} \in \mathbb{R}^{1024\times768}$. The final encoding vector $\mathbf{E}_i$ for the script is computed as the average of the transformed embeddings, $\mathbf{E}_i = \frac{1}{m}\sum_{j=1}^{m} \mathbf{e}_j$.

5.2 Time-Series Modality Encoder (CAMEF$_{\text{Time-Series}}$)

To encode the time series data, we employ a pretrained time series encoder, MOMENT [21], which generates a fixed-dimensional vector for an input time series segment. Subsequently, we design a multi-residual layer to further refine the encoding vectors, as shown below:

$$\mathcal{X}_i = \textbf{MOMENT}(\{X_1, X_2, \ldots, X_n\}), \tag{4}$$

where $\mathcal{X}_i \in \mathbb{R}^d$ is the encoded vector for the input time series segment $\{X_1, X_2, \ldots, X_n\}$, and d represents the dimensionality of the encoded vector. To enhance this representation, we introduce a multi-residual projection layer:

$$\mathbf{Z}_i = \mathcal{X}_i + f_{\text{residual}}(\mathcal{X}_i), \tag{5}$$

where $f_{\text{residual}}(\mathcal{X}_i)$ represents the transformation applied through the residual projection layer, which consists of multiple linear layers interleaved with GELU activations:

$$f_{\text{residual}}(\mathcal{X}_i) = \mathbf{W}_3 \cdot \text{GELU}(\mathbf{W}_2 \cdot \text{GELU}(\mathbf{W}_1 \cdot \mathcal{X}_i + \mathbf{b}_1) + \mathbf{b}_2) + \mathbf{b}_3, \tag{6}$$

where $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{1024\times1024}$, $\mathbf{W}_3 \in \mathbb{R}^{1024\times768}$ are the weight matrices, and $\mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}^{1024}$, $\mathbf{b}_3 \in \mathbb{R}^{768}$ are the respective biases. Finally, $\mathbf{Z}_i \in \mathbb{R}^{1024}$ serves as the refined vector for the time series segment.

5.3 Feature Fusion and Time Series Decoder

After obtaining the encoded vectors from the textual and time series data, denoted as $\mathbf{E}_i$ and $\mathbf{Z}_i$, respectively, we concatenate them to form a unified representation:

$$\mathbf{E}_{\text{combined}} = \text{Concat}(\mathbf{E}_i, \mathbf{Z}_i), \tag{7}$$

which captures both semantic content from macroeconomic texts and temporal patterns from time series inputs. To fuse these modalities, $\mathbf{E}_{\text{combined}}$ is passed through a two-layer feedforward network with GELU activation:

$$\mathbf{E}_{\text{fused}} = \mathbf{W}^{(f2)} \cdot \text{GELU}(\mathbf{W}^{(f1)} \cdot \mathbf{E} * \text{combined} + \mathbf{b}^{(f1)}) + \mathbf{b}^{(f2)}, \quad (8)$$

where $\mathbf{W}^{(f1)} \in \mathbb{R}^{(2\times768)\times1024}$, $\mathbf{W}^{(f2)} \in \mathbb{R}^{1024\times1024}$, and $\mathbf{b}^{(f1)}, \mathbf{b}^{(f2)} \in \mathbb{R}^{1024}$ are the corresponding weight matrices and bias terms. This fusion block enables interaction across modalities and produces a refined joint embedding for downstream decoding.

We then employ GPT-2 [41] as the decoder to decode the fused vector by leveraging its effective auto-regressive ability:

$$\mathbf{H}^{(l)} = f^{(l)}_{\text{GPT2\_layer}}(\mathbf{H}^{(l-1)}), \quad (9)$$

where $\mathbf{H}^{(0)} = \mathbf{E}_{\text{fused}}$, and $f^{(l)}_{\text{GPT2\_layer}}$ represents the transformation function of the l-th GPT-2 layer, where $l = 12$. After the final layer, the output is normalized using a layer normalization function:

$$\mathbf{H}_{\text{final}} = \text{LayerNorm}(\mathbf{H}^{(l)}), \quad (10)$$

The final output $\mathbf{H}_{\text{final}}$ is then used to generate predictions based on the combined multi-modal information.

5.4 Time-Series Forecasting Post-Regressor and Learning Objectives

We designed a **Post-Regressor** that applies a linear transformation to the concatenated vector $\mathbf{H}_{\text{final}}$, followed by GELU activation and a dropout layer with a rate of 0.1:

$$\mathbf{R}^{(k)} = \text{GELU}(\mathbf{W}^{(k)} \cdot \mathbf{R}^{(k-1)} + \mathbf{b}^{(k)}), \quad (11)$$

where $\mathbf{R}^{(0)} = \mathbf{H}_{\text{final}}$, $\mathbf{W}^{(k)}$ and $\mathbf{b}^{(k)}$ are the weight matrix and bias of the k-th linear layer, respectively. $k = 4$ is the total number of layers in the regressor. The final linear layer maps the representation to a vector of shape ($d \times$ pred_len), where d is the forecast dimensionality and pred_len is the number of predicted time steps:

$$\hat{\mathbf{Y}} = \mathbf{W}_{\text{out}} \cdot \mathbf{R}^{(K)} + \mathbf{b}_{\text{out}}, \quad (12)$$

where $\hat{\mathbf{Y}}$, represents the predicted time series values.

Learning Objectives for Time Series: We employ a combination of Mean Squared Error (MSE) loss and Mean Absolute Error (MAE) loss to optimize the model. The MSE loss minimizes the squared differences between the predicted time series values, $\hat{\mathbf{Y}}$, and the ground truth values, $\mathbf{Y}$, while the MAE loss minimizes the absolute differences. These are defined as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{Y}}_i - \mathbf{Y}_i)^2, \quad \mathcal{L}_{\text{MAE}} = \frac{1}{n}\sum_{i=1}^{n}|\hat{\mathbf{Y}}_i - \mathbf{Y}_i|, \quad (13)$$

where n is the number of predicted values (e.g., 35, 70, or 140, as defined in Section 6.1). The total loss function combines both objectives to balance optimization for large and small errors:

$$\mathcal{L}_{\text{Time}} = \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{MAE}}. \quad (14)$$

5.5 Counterfactual Events Sampling and Causal Learning Objective

Causal learning enhances the robustness of the CAMEF model by enabling it to identify the correct event script among sampled counterfactual events (CEs). To achieve this, we first design a **Diverse Counterfactual Event Sampling Mechanism**, which generates two types of CEs. These counterfactuals, along with their corresponding time-series data, are then encoded using the textual and time-series modalities of CAMEF. This process helps the model learn causal relationships between events and their corresponding time-series movements.

5.5.1 Diverse Counterfactual Event Sampling Mechanism. We propose a **Diverse Counterfactual Event Sampling Mechanism** to enhance the model's ability to both identify the ground-truth event and distinguish between different event types. This mechanism is designed with two objectives: (1) to help the model recognize the ground-truth event among similar counterfactuals of the same type, and (2) to enable the model to differentiate between events of different types.

To achieve these objectives, we generate two categories of counterfactual events for each factual event:

(1) **Identical Type Sampling:** Counterfactual events of the same type as the ground-truth event, created by modifying sentiment-relevant components and key numerical variables, as detailed in Sec. 4.2.
(2) **Diverse Type Sampling:** Counterfactual events of a different type, sampled by substituting the ground-truth event with 5 other event type occurring on the closest date.

This mechanism provides a diverse set of counterfactual events, collectively denoted as $\mathbb{E}^{CF} := \{\mathcal{E}^{CF}_1, \mathcal{E}^{CF}_2, \ldots, \mathcal{E}^{CF}_{|\mathbb{E}^{CF}|}\}$, we set the total number of diverse-type samples to be 5, and the default number of identical-type samples to be 10 as introduced Sec. 4.2.

5.5.2 Causal Learning Objective. The causal learning process utilizes both the textual (see Sec. 5.1) and time-series (see Sec. 5.2) encoders of CAMEF to capture the relationships between events and market movements. The textual encoder is used to encode both the ground-truth event $\mathcal{E}^{GT}$ and the sampled CEs $\mathcal{E}^{CF}$:

$$\{\mathbf{P}^{GT}, \mathbf{P}^{CF}_1, \ldots, \mathbf{P}^{CF}_{|\mathbb{E}^{CF}|}\} = \mathbf{CAMEF}_{\text{Textual}}(\{\mathcal{E}^{GT}\} \cup \{\mathcal{E}^{CF}_i\}_{i=1}^{|\mathbb{E}^{CF}|}), \quad (15)$$

where $\mathbf{P}^{GT}$ represents the embedding of the ground-truth event, and $\{\mathbf{P}^{CF}_i\}_{i=1}^{|\mathbb{E}^{CF}|}$ represents the embeddings of the sampled counterfactual events.

The time-series encoder is used to encode the historical time-series segment $\mathcal{X}$ aligned with the ground-truth event, resulting in the time-series embedding:

$$\mathbf{T} = \mathbf{CAMEF}_{\text{Time-Series}}(\mathcal{X}). \quad (16)$$

Triplet Loss: The triplet loss is applied to enforce that the ground-truth event embedding $\mathbf{P}^{GT}$ is closer to the time-series embedding $\mathbf{T}$ than any counterfactual event embedding $\mathbf{P}^{CF}_i$, by a margin α (set to 1.0):

Table 2: Financial Forecasting Results (MSE and MAE Scores) for CAMEF and Baselines Across Various Financial Assets: S&P500 (SPX), Dow Industrials (INDU), Nasdaq100 (NDX) Index, US 1-Month Treasury Bond (USGG1M), and US 5-Year Treasury Bond (USGG5YR).

Model / Datasets	Forecasting Length	SP500 (SPX)		Dow Industrial (INDU)		NASDAQ (NDX)		USGG1M		USGG5YR	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ARIMA	35	0.0032628	0.0308016	0.0121513	0.0523810	0.0065253	0.0435694	0.0072223	0.0262512	0.0038973	0.0340227
	70	0.0035361	0.0352004	0.0139245	0.0656002	0.0082324	0.0520933	0.0028710	0.0304132	0.0039441	0.0366630
	140	0.0051080	0.0439935	0.0219147	0.0793471	0.0118692	0.0665155	0.0089949	0.0338275	0.0050746	0.0455617
DLinear	35	0.0144331	0.0896910	0.0395136	0.1380539	0.0183989	0.0999178	0.0108576	0.0706170	0.0147211	0.0876379
	70	0.0120578	0.0817373	0.0406573	0.1282699	0.0189431	0.0972439	0.0093591	0.0678380	0.0146430	0.0881574
	140	0.0178138	0.0931039	0.0747210	0.1724027	0.0344153	0.1287803	0.0101465	0.0645396	0.0179185	0.0977673
Autoformer	35	0.0068136	0.0540556	0.0277636	0.0975948	0.0135249	0.0796486	0.0047505	0.0388076	0.0092717	0.0622862
	70	0.0088997	0.0628341	0.0375279	0.1185710	0.0185264	0.0933398	0.0065554	0.0471531	0.0113750	0.0727373
	140	0.0158248	0.0829188	0.0772580	0.1640365	0.0334504	0.1262163	0.0086212	0.0533337	0.0152298	0.0875821
FEDformer	35	0.0072377	0.0576221	0.0304808	0.1044447	0.0128668	0.0758742	0.0063596	0.0519141	0.0094995	0.0494306
	70	0.0088056	0.0621386	0.0399824	0.1229772	0.0171008	0.0889995	0.0062841	0.0448822	0.0099615	0.0664197
	140	0.0157429	0.0819469	0.0786426	0.1677705	0.0313433	0.1214097	0.0083784	0.0518365	0.0131231	0.0782861
iTransformer	35	0.0064209	0.0516341	0.0270860	0.0927605	0.0125008	0.0751656	0.0011000	0.0155975	0.0056660	**0.0183524**
	70	0.0069612	0.0540920	0.0304566	0.1038424	0.0151214	0.0816262	0.0021811	0.0221315	**0.0011721**	**0.0226975**
	140	0.0128021	0.0718771	0.0680991	0.1486782	0.0254479	0.1047119	0.0052255	0.0327429	**0.0017441**	**0.0282537**
PatchTST	35	0.0063304	0.0507462	0.0293131	0.0989455	0.0122679	0.0764552	0.0012060	0.0163610	0.0063078	0.0520734
	70	0.0072471	0.0547738	0.0339444	0.1116023	0.0153753	0.0824677	0.0021643	0.0223036	0.0079617	0.0606007
	140	0.0130219	0.0712171	0.0688582	0.1452592	0.0256001	0.1047749	0.0054544	0.0341517	0.0118553	0.0747537
GPT4MTS	35	0.00795088	0.0674255	0.0026558	0.0417553	0.0011035	0.0240469	0.0017016	0.0309320	0.0028713	0.0389277
	70	0.00171038	0.0305950	0.0027033	0.0393112	0.0016205	0.0336116	0.0019098	0.0279222	0.0023371	0.0354777
	140	0.00212612	0.0330497	0.0045029	0.0458267	0.0025175	0.0341671	0.0013648	0.0260887	0.0037004	0.0449272
TEST	35	0.00073333	0.0199733	0.0026572	**0.0296887**	0.0006593	0.0194091	0.0003252	0.0137511	0.0013633	0.0265036
	70	0.00078762	0.0205412	0.0088903	0.0599179	0.0010678	0.0248594	0.0010515	0.0182706	0.0034630	0.0415002
	140	0.00278572	0.0467150	0.0070749	0.0557744	0.0020130	0.0362325	0.0006995	0.0207850	0.0024356	0.0375164
CAMEF	35	**0.00048860**	**0.0154050**	**0.0025349**	0.0366245	**0.0005468**	**0.0178845**	**0.0002883**	**0.0118010**	**0.0013234**	0.0260618
	70	**0.00064780**	**0.0178691**	**0.0025042**	**0.0365500**	**0.0005814**	**0.0162882**	**0.0004402**	**0.0139699**	0.0020701	0.0326371
	140	**0.0010756**	**0.0210284**	**0.0039313**	**0.0383459**	**0.0010159**	**0.0207716**	**0.0004938**	**0.0148485**	0.0022458	0.0336680

$$\mathcal{L}_{\text{Causal-TL}} = \max\left(0, d(\mathbf{P}^{GT}, \mathbf{T}) - d(\mathbf{P}_i^{CF}, \mathbf{T}) + \alpha\right), \tag{17}$$

where $d(\cdot,\cdot)$ denotes the distance between two embeddings (e.g., cosine similarity or Euclidean distance). This loss function encourages the model to capture the causal relationships between events and time-series movements by penalizing counterfactual events that deviate from the causal signal of the ground-truth event.

The combination of diverse counterfactual sampling and causal learning ensures that CAMEF effectively learns the true causal drivers of financial market movements, improving its robustness and predictive power.

Total Loss: The overall training loss for CAMEF is defined as:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Time}} + \mathcal{L}_{\text{Causal-TL}}, \tag{18}$$

where $\mathcal{L}_{\text{Time}}$ is the objective for time series forecasting, as defined in Equation 14.

6 Experiments

In this section, we evaluate CAMEF by addressing the following key questions: **RQ1. Accuracy:** How accurately does CAMEF forecast fina ncial market based on events? **RQ2. Model Effectiveness:** How do different components enhance CAMEF's predictive performance? **RQ3. Event Analysis:** Which types of events exhibit stronger influences to financial market?

6.1 Experimental Settings

6.1.1 Datasets. We utilized the collected event scripts and time-series data as outlined in Sec. 4 and detailed in Appendix B. The dataset was divided into training, validation, and testing sets in a 6:2:2 ratio. The training set was used to train the models, while the validation set was used for convergence checking and early stopping to prevent overfitting. The final results, based on the test set, are reported in Table 2.

6.1.2 Baselines. To evaluate our proposed method, we compare it against both **uni-modal** and **multi-modal** time series forecasting approaches. For the **uni-modal baselines**, we considered the traditional yet robust ARIMA model [6], the linear neural model DLinear [59], and several state-of-the-art transformer-based time-series models, including AutoFormer [8], FEDformer [61], and iTransformer [30]. For the **multi-modal baselines**, we included TEST [51] and GPT4MTS [24].

6.1.3 Test Settings. We evaluate the baselines and CAMEF across three time horizons—short, medium, and long run, to simulate real investment behavior. For each aligned pair of event and time-series data ($[\mathcal{X}_{i-\tau:i+\tau} \mapsto \mathcal{E}_i]$) as defined in Sec. 3, we use the event script ($\mathcal{X}$) and the time-series segment preceding the event time point i, i.e., $\mathcal{X}_{i-\tau:i}$, to forecast the future time-series segment $\mathcal{X}_{i+1:i+\tau}$. The value of τ is adjusted based on the time horizon: we set τ to 35, 70, and 140 for short, medium, and long-term forecasts, respectively. These correspond to 175 minutes (about half a trading day), 350 minutes (about one trading day), and 700 minutes (about two trading days).

Table 3: Ablation Study Results (MSE) Evaluating the CAMEF Model Components at Forecasting Lengths of 35, 70, and 140.

Textual	Causal	Feature Fusion	GPT2 Decoder	Post-Regressor	SPX			INDU			NDX			USGG1M			USGG5YR			AVERAGE		
					35	70	140	35	70	140	35	70	140	35	70	140	35	70	140	35	70	140
✗	✓	✓	✓	✓	0.00074	0.00340	0.00120	0.00340	0.00330	0.01033	0.00131	0.00097	0.00197	0.00080	0.00179	0.00092	0.00222	0.00276	0.00380	0.00169	0.00199	0.00364
✓	✗	✓	✓	✓	0.00068	0.00095	0.00106	0.02939	0.00281	0.00533	0.00064	0.00073	0.00121	0.00065	0.00083	0.00099	0.00160	0.00220	0.00308	0.00659	0.00170	0.00233
✓	✓	✗	✓	✓	0.00080	0.00079	0.00110	0.01173	0.00391	0.00581	0.00069	0.00073	0.00168	0.00047	0.00046	0.00216	0.00251	0.00306	0.00426	0.00324	0.00212	0.00300
✓	✓	✓	✗	✓	0.00073	0.00067	0.00114	0.26768	0.72387	0.18091	0.00062	0.00069	0.00158	0.00043	0.02110	0.75057	0.00220	0.00309	0.00566	0.05433	0.14593	0.03801
✓	✓	✓	✓	✗	0.00662	0.03911	0.00979	0.50841	0.57007	0.22267	0.00774	0.00852	0.00950	0.28932	0.02110	0.75057	0.56705	0.06188	0.08251	0.27583	0.14014	0.21501
Full CAMEF Model					**0.00048**	**0.00064**	**0.00107**	**0.00253**	**0.00250**	**0.00393**	**0.00054**	**0.00058**	**0.00101**	**0.00028**	**0.00044**	**0.00049**	**0.00132**	**0.00207**	**0.00224**	**0.00104**	**0.00124**	**0.00174**

Table 4: Ablation Study Results on Different Type of Events on S&P500 Index

Event Type	Forecasting Length = 35		Forecasting Length = 70		Forecasting Length = 140	
	MSE	MAE	MSE	MAE	MSE	MAE
Unemployment Insurance	0.0004870 ↓	0.0159497 ↑	0.0005199 ↓	0.0157778 ↓	0.0006948 ↓	0.0190141 ↓
Employment Situation	0.0003923 ↓	0.0142684 ↓	0.0004612 ↓	**0.0154431** ↓	0.0013767 ↑	0.0218677 ↑
GDP Adcance	0.0006203 ↑	0.0175397 ↑	0.0005897 ↓	0.0175464 ↓	0.0011554 ↑	0.0217548 ↑
FOMC Minutes	**0.0003401** ↓	**0.0127694** ↓	**0.0004448** ↓	0.0170094 ↓	**0.0006433** ↓	**0.0185657** ↓
CPI Report	0.0005645 ↑	0.0160723 ↑	0.0010660 ↑	0.0212536 ↑	0.0008187 ↓	0.0197824 ↓
PPI Report	0.0005275 ↑	0.0148434 ↓	0.0008054 ↑	0.0201844 ↑	0.0017646 ↑	0.0251856 ↑
Full Selection	0.0004886	0.0152405	0.0006478	0.0178691	0.0010756	0.0210284

6.1.4 Implementation Overview. For the single-modality approaches (ARIMA, DLinear, AutoFormer, FEDformer, iTransformer, and PatchTST), we tested two methods: (1) training the models on continuous historical time-series data and testing on aligned event-based time-series segments from the test set, and (2) training the models directly on event-based time-series segments, and also test on the aligned event-based time-series segments from the test set. The second approach produced more accurate results, and these are the results presented in this paper. Specifically, for each aligned event data pair ($[\mathcal{X}_{i-\tau:i+\tau} \mapsto \mathcal{E}_i]$), single-modality approaches use only the time-series segment preceding the event, $\mathcal{X}_{i-\tau:i}$, to train the models and forecast the subsequent segment, $\mathcal{X}_{i+1:i+\tau}$; and testing follows the same approach. Detailed settings for each model are provided in Appendix C.

For the multi-modality approaches (GPT4MTS, TEST, and CAMEF), both the event script $\mathcal{E}_i$ and the time-series segment preceding the event, $\mathcal{X}_{i-\tau:i}$, are used as input to train the models to forecast the post-event time-series segment, $\mathcal{X}_{i+1:i+\tau}$. Implementation details, including model configurations and training settings, are explained in Appendix C.

6.2 Experimental Results for Event-Driven Time-Series Forecasting (RQ1)

Table 2 presents the forecasting results for the five datasets across short, medium, and long forecasting horizons. CAMEF outperformed other models in 24 out of 30 settings, achieving first-place rankings, and ranked second in the remaining 6 settings. Specifically, CAMEF demonstrated the best performance across all forecasting lengths for the Stock Market Indices (SPX, INDU, and NDX), except for short-horizon forecasting on INDU, highlighting its effectiveness in event-driven stock market forecasting. For treasury bonds, CAMEF achieved the best results across all forecasting lengths for the 1-month treasury bond (USGG1M) and ranked second for the 5-year treasury bond (USGG5YR). The slightly lower performance on USGG5YR suggests that long-run treasury bonds may be less sensitive to event-driven factors and more influenced by historical trends.

Compared to single-modality models (e.g., DLinear, Autoformer, FEDformer, PatchTST, and iTransformer), CAMEF achieved an average MSE reduction of 62.55% relative to the best-performing single-modality model, iTransformer. Among multi-modality models, CAMEF surpassed TEST, the second-best performer, with an average MSE reduction of 33.55%.

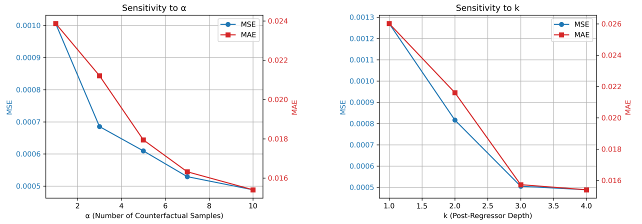
These results highlight three key insights: (1) effectively leveraging multi-modality information provides significant performance gains, particularly for SPX, INDU, NDX, and USGG1M; (2) transformer-based methods consistently outperform classical models such as ARIMA; and (3) CAMEF's superior training and feature fusion strategies establish it as the most effective method for event-driven financial forecasting tasks.

6.3 Ablation Studies on Model Components (RQ2)

To evaluate the effectiveness of each component in CAMEF, we conduct comprehensive ablation studies on Textual Modality, Causal Learning, Feature Fusion, GPT2 Decoder, and the Post-Regressor. Based on the full CAMEF model, we remove the corresponding neural layers, such as the RoBERTa encoder for textual modality or the causal learning component, to assess their individual contributions. The ablation results are presented in Table 3, where the left part of the table uses ✓ or ✗ to indicate whether the specific component is included or excluded in the test. From the results, three key findings are: (1) The full CAMEF model achieves the best performance across all datasets, demonstrating the critical importance of utilizing both textual and time-series modalities; (2) Causal learning provides incremental improvements, confirming its value in capturing cause-effect relationships within the data; (3) The proposed feature fusion layers and GPT2 decoder effectively integrate and leverage multi-modality features, significantly enhancing the model's ability to decode time-series data. These findings underscore the necessity of each component in achieving optimal performance for event-driven financial forecasting tasks.

6.4 Ablation Studies on Different Type of Events (RQ3)

Table 4 shows the predictive performance of different events on the S&P500 Index. **FOMC Minutes** achieve the lowest MSE and MAE, confirming their critical importance for market prediction. **Unemployment Insurance Claims** and **Unemployment Situation Reports** also come with lower errors than the full selection, however the latter becomes less effective at long forecasting length. In contrast, **CPI** and **PPI Reports** show weaker predictive power, with PPI yielding the highest errors and CPI improving slightly at long forecasting length. These results emphasize the importance of FOMC and unemployment-related events for financial forecasting.



(a) Effect of counterfactual sample size (α)

(b) Effect of post-regressor depth (k)

Figure 4: Sensitivity analysis of CAMEF on the SP500 dataset with respect to key hyperparameters.

6.5 Parameter Sensitivity Analysis

We examine the impact of two key hyperparameters in CAMEF on SPX dataset with 35 predictive length: the number of counterfactual samples (α) and the depth of the post-regressor network (k). The number of GPT-2 layers (l) remains fixed, as we adopt the pre-trained model without modification. As shown in Figure 4, increasing α enhances forecasting accuracy through stronger contrastive supervision, with performance gains saturating around $\alpha = 10$. Similarly, increasing k improves decoding capacity, though with diminishing returns beyond $k = 4$. These trends suggest that CAMEF performs robustly across a range of configurations, benefiting from moderate complexity increases without overfitting.

7 Conclusion and Future Work

This paper proposed **CAMEF**, a multi-modality model for event-driven financial forecasting, which integrates effective causal learning and an LLM-based counterfactual event augmentation strategy. Alongside the model, we introduced a novel synthetic dataset comprising 6 types of salient macroeconomic event scripts, their counterfactual samples, and high-frequency time-series data for 5 key financial assets, aligned with real-world investment practices. Extensive experiments demonstrated CAMEF's superior predictive performance compared to prior deep time-series and multi-modality methods. Ablation studies testified the importance of causal learning and other designed components of CAMEF. Additionally, it is found that FOMC and unemployment-related events provided the most predictive value among the tested event types.

For future work, we plan to leverage advanced LLMs for enhanced textual encoding to extract deeper semantic information, refine the cross-modality causal inference mechanisms, and expand the dataset to include additional event types, such as political events and corporate market-sensitive news.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant Nos. 62402396 and 72471197. We would like to thank the anonymous reviewers for their valuable comments and suggestions that helped improve the quality of this paper.

References

[1] Adebiyi A. Ariyo, Adewumi O. Adewumi, and Charles K. Ayo. [n. d.]. Stock Price Prediction Using the ARIMA Model. In *UKSim-AMSS 2014*.

[2] Wuzhida Bao, Yuting Cao, Yin Yang, Hangjun Che, Junjian Huang, and Shiping Wen. 2025. Data-driven stock forecasting models based on neural networks: A review. *Information Fusion* 113 (2025), 102616.

[3] Leonardo Bartolini, Linda Goldberg, and Adam Sacarny. 2008. How economic news moves markets. *Curr. Issues Econ. Finance* 14, Aug (2008), 6.

[4] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv* (2020).

[5] Dinesh Bhuriya, Girish Kaushal, Ashish Sharma, and Upendra Singh. 2017. Stock market predication using a linear regression. In *ICECA 2017*.

[6] George Edward Pelham Box and Gwilym M. Jenkins. 1994. *Time Series Analysis: Forecasting and Control*. Prentice Hall PTR.

[7] C. Q. Cao and R. S. Tsay. 1992. Nonlinear time-series analysis of stock volatilities. *J. Appl. Econom.* 7, S1 (1992), 165–185.

[8] Minghao Chen, Houwen Peng, Jianlong Fu, and Haibin Ling. 2021. AutoFormer: Searching Transformers for Visual Recognition. In *ICCV 2021*.

[9] Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2023. DISCO: Distilling Counterfactuals with Large Language Models. In *ACL 2023*.

[10] ANNA CIESLAK, ADAIR MORSE, and ANNETTE VISSING-JORGENSEN. 2019. Stock Returns over the FOMC Cycle. *J. Finance* 74, 5 (2019), 2201–2248.

[11] Kiyoshi Izumi Daigo Tashiro, Hiroyasu Matsushima and Hiroki Sakaji. 2019. Encoding of high-frequency order information and prediction of short-term stock price by deep learning. *Quant. Finance* 19, 9 (2019), 1499–1506.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL 2019*.

[13] Patti Domm. 2021. One Year Ago, Stocks Dropped 12% in a Single Day—What Investors Have Learned Since Then. https://www.cnbc.com/2021/03/16/one-year-ago-stocks-dropped-12percent-in-a-single-day-what-investors-have-learned-since-then.html Accessed: 2024-10-06.

[14] Dr. M. Durairaj and B. H. Krishna Mohan. 2022. A convolutional neural network based approach to financial time series prediction. *Neural Comput. Appl.* 34, 16 (2022), 13319–13337.

[15] Eugene F. Fama. 1965. The Behavior of Stock-Market Prices. *J. Bus.* 38, 1 (1965), 34–105.

[16] Eugene F. Fama, Lawrence Fisher, Michael C. Jensen, and Richard Roll. 1969. The Adjustment of Stock Prices to New Information. *Int. Econ. Rev.* 10, 1 (1969), 1–21.

[17] Pavel Gertler and Roman Horvath. 2018. Central bank communication and financial markets: New high-frequency evidence. *J. Financ. Stab.* 36 (2018), 336–345.

[18] T Gilbert, C Scotti, G Strasser, and C Vega. 2010. Why do certain macroeconomic news announcements have a big impact on asset prices? In *Appl. Econom. Forecast. Macro. Finance Workshop*.

[19] Thomas Gilbert, Chiara Scotti, Georg Strasser, and Clara Vega. 2017. Is the intrinsic value of a macroeconomic news announcement related to its asset price impact? *J. Monet. Econ.* 92 (2017), 78–95.

[20] Linda Goldberg and Christian Grisse. 2013. *Time variation in asset price responses to macro announcements*. Working Papers 2013-11. Swiss National Bank.

[21] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. 2024. MOMENT: A Family of Open Time-series Foundation Models. In *ICML 2024*.

[22] Huy D. Huynh, L. Minh Dang, and Duc Duong. 2017. A New Model for Stock Price Movements Prediction Using Deep Neural Network. In *SoICT 2017*.

[23] Tae Hyup Roh. 2007. Forecasting the volatility of stock price index. *Expert Syst. Appl.* 33, 4 (2007), 916–922.

[24] Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. 2024. GPT4MTS: Prompt-based Large Language Model for Multimodal Time-series Forecasting. *AAAI 2024* (2024).

[25] Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning The Difference That Makes A Difference With Counterfactually-Augmented Data. In *ICLR 2020*.

[26] Suk Joong Kim, Michael D. McKenzie, and Robert W. Faff. 2004. Macroeconomic news announcements and the role of expectations: Evidence for US bond, stock and foreign exchange markets. *J. Multinatl. Financ. Manag.* 14, 3 (2004), 217–232.

[27] Geon Lee, Wenchao Yu, Wei Cheng, and Haifeng Chen. 2024. MoAT: Multi-Modal Augmented Time Series Forecasting. OpenReview

[28] Qing Li, TieJun Wang, Ping Li, Ling Liu, Qixu Gong, and Yuanzhu Chen. 2014. The effect of news and public mood on stock movements. *Information Sciences* 278 (2014), 826–840.

[29] Huicheng Liu. 2018. Leveraging Financial News for Stock Trend Prediction with Attention-Based Recurrent Neural Network. *arXiv* (2018).

[30] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. 2024. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In *ICLR 2024*.

[31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* (2019).

[32] Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. 2024. Timer: Generative Pre-trained Transformers Are Large Time Series Models. In *ICML 2024*.

[33] DAVID O. LUCCA and EMANUEL MOENCH. 2015. The Pre-FOMC Announcement Drift. *J. Finance* 70, 1 (2015), 329–371.

[34] Donato Masciandaro, Oana Peia, and Davide Romelli. 2024. Central bank communication and social media: From silence to Twitter. *J. Econ. Surv.* 38, 2 (2024), 365–388.

[35] Puneet Mathur, Atula Neerkaje, Malika Chhibber, Ramit Sawhney, Fuming Guo, Franck Dernoncourt, Sanghamitra Dutta, and Dinesh Manocha. 2022. MONOPOLY: Financial Prediction from MONetary POLicY Conference Videos Using Multimodal Cues. In *MM 2022*.

[36] Thien Hai Nguyen, Kiyoaki Shirai, and Julien Velcin. 2015. Sentiment analysis on social media for stock movement prediction. *Expert Syst. Appl.* 42, 24 (2015), 9603–9611.

[37] Kun Ouyang, Yi Liu, Shicheng Li, Ruihan Bao, Keiko Harimoto, and Xu Sun. 2024. Modal-adaptive Knowledge-enhanced Graph-based Financial Prediction from Monetary Policy Conference Calls with LLM. In *FinNLP-KDF-ECONLP 2024*.

[38] Douglas K Pearce and V. Vance Roley. 1984. *Stock Prices and Economic News.* Working Paper. National Bureau of Economic Research.

[39] G. Pui Cheong Fung, J. Xu Yu, and Wai Lam. 2003. Stock prediction: Integrating text mining approach using real-time news. In *CIFEr*. 395–402.

[40] Yu Qin and Yi Yang. 2019. What You Say and How You Say It Matters: Predicting Stock Volatility Using Verbal and Vocal Cues. In *ACL 2019*.

[41] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[42] Carlo Rosa. 2011. Words that shake traders: The stock market's reaction to central bank communication in real time. *J. Empir. Finance* 18, 5 (2011), 915–934.

[43] Carlo Rosa. 2013. The financial market effect of FOMC minutes. *Econ. Policy Rev.* (2013), 67–81.

[44] Carlo Rosa. 2016. Fedspeak: Who Moves U.S. Asset Prices? *Int. J. Cent. Bank.* 12, 4 (2016), 223–261.

[45] Alexis Ross, Tongshuang Wu, Hao Peng, Matthew Peters, and Matt Gardner. 2022. Tailor: Generating and Perturbing Text with Semantic Controls. In *ACL 2022*.

[46] Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Ratn Shah. 2020. Deep Attentive Learning for Stock Movement Prediction From Social Media Text and Company Correlations. In *EMNLP 2020*.

[47] Ramit Sawhney, Puneet Mathur, Ayush Mangal, Piyush Khanna, Rajiv Ratn Shah, and Roger Zimmermann. 2020. Multimodal Multi-Task Financial Risk Forecasting. In *MM 2020*.

[48] Sreelekshmy Selvin, R Vinayakumar, E. A Gopalakrishnan, Vijay Krishna Menon, and K. P. Soman. 2017. Stock price prediction using LSTM, RNN and CNN-sliding window model. In *ICACCI 2017*.

[49] Agam Shah, Suvan Paturi, and Sudheer Chava. 2023. Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis. In *ACL 2023*.

[50] Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. 2013. Exploiting Topic based Twitter Sentiment for Stock Prediction. In *ACL 2013*.

[51] Chenxi Sun, Hongyan Li, Yaliang Li, and Shenda Hong. 2024. TEST: Text Prototype Aligned Embedding to Activate LLM's Ability for Time Series. In *ICLR 2024*.

[52] Raul Cruz Tadle. 2022. FOMC minutes sentiments and their impact on financial markets. *J. Econ. Bus.* 118 (2022), 106021.

[53] Sabera J Talukder, Yisong Yue, and Georgia Gkioxari. 2024. TOTEM: TOkenized Time Series EMbeddings for General Time Series Analysis. *TMLR* (2024).

[54] Stephen J Taylor. 2007. *Modelling Financial Time Series.* World Scientific Publishing.

[55] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *ACL 2023*.

[56] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models. In *ACL 2021*.

[57] Yumo Xu and Shay B. Cohen. 2018. Stock Movement Prediction from Tweets and Historical Prices. In *ACL 2018*.

[58] Linyi Yang, Tin Lok James Ng, Barry Smyth, and Riuhai Dong. 2020. HTML: Hierarchical Transformer-based Multi-task Learning for Volatility Prediction. In *WWW 2020*.

[59] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2023. Are Transformers Effective for Time Series Forecasting?. In *AAAI 2023*.

[60] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. In *AAAI 2021*.

[61] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. 2022. FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting. In *ICML 2022*.

[62] Zhihan Zhou, Liqian Ma, and Han Liu. [n. d.]. Trade the Event: Corporate Events Detection for News-Based Event-Driven Trading. In *ACL Findings 2021*.

A Prompt Templates for Counterfactual Events Generations

A.1 Summarization Prompt Template

Due to the script's length, we firstly prompt LLaMA-3 summarizes each chunk of an event script, followed by a second prompt to generate the full summary. Below is the chunk-level prompt, which instructs the model to summarize within a word limit. Text in " " indicates input variables.

Prompt Template for Event Script Chunk Summarization

> You are given chunk {*chunk_idx*} of a {*text_type*} report. Your task is to generate a summary within {*number_of_words*} words.
> The content of chunk {*chunk_idx*} is as follows:
> {*original_text*}
> Please provide a concise summary, while keep the key variables:

The second prompt combines the chunk summaries as input to generate a final summary for the entire event script:

Prompt Template for Final Summarization

> You are given {*chunk_num*} summaries of different chunks from a {*text_type*} report. Your task is to generate an overall summary within {*number_of_words*} words.
> The chunk summaries are as follows: {*chunk_summaries*}
> Please provide a comprehensive summary of the entire report, while keep the key variables:

A.2 Sentiment Analytical Prompt Template

For sentiment analysis, the prompt asks the model to rate an event scirpt's sentiment from 0 (negative) to 10 (positive) and explain the rating. Below is the prompt template, with " " denoting input variables.

Sentiment Analysis Prompt

> Please analyze the sentiment of the following {*text_type*} summary and rate it on a scale from 0 to 10, where:
> 0 = Extremely Negative; 1 = Strongly Negative; 2 = Very; Negative; 3 = Moderate Negative; 4 = Slightly Negative; 5 = Neutral; 6 = Slightly Positive; 7 = Moderate Positive; 8 = Very Positive; 9 = Strongly Positive; 10 = Extremely Positive {*text_type*} summary: {*text*}
> Output the sentiment analysis as:
> Sentiment rating: (0 to 10), Explanation:

A.3 Counterfactual Event Generation Prompt Template

To generate counterfactual versions of a text with different sentiment levels, we use a prompt that instructs the model to modify key facts and information to align with a target sentiment rating. The model is provided with the original sentiment rating and sentiment description and is asked to adjust the text to reflect a specified target sentiment rating. Below is the prompt, where the text within "{ }" indicates the input variables:

Counterfactual Text Generation Prompt

> The original text has been identified with a sentiment rating of {*current_sentiment_rating*} ({*current_sentiment*}). Your task is to generate a counterfactual version of the text that aligns with a sentiment rating of {*target_sentiment_rating*} ({*target_sentiment*}) by modifying the key facts and information to reflect the specified target sentiment score about the economy, while keep the overall format and the sentiment-neural content unchanged.
> Original text: {*original_text*}
> Counterfactual text with a sentiment rating of {*target_sentiment_rating*} ({*target_sentiment*}):

B Implementation Details of Dataset Collection and Preprocessing

B.1 Dataset Collection

To construct the CAMEF dataset, raw macroeconomic event data are crawled from official government sources (Tab. 1) in various formats (PDF, HTML, TXT). We utilize the following Python libraries:

- **Requests** – to send HTTP requests for archive access;
- **Selenium** – to locate and download files via HTML headers;
- **BeautifulSoup** – to parse HTML and extract file links.

The dataset covers six event types: FOMC minutes, CPI, PPI, unemployment insurance, unemployment rate, and GDP advance releases.

B.2 Preprocessing

Raw files are converted into a unified text format. `PyPDF2` and `pdfplumber` extract content from PDFs, while `BeautifulSoup` parses HTML using depth-first traversal. TXT files are used as-is. We convert all the tables into structured text using the delimiter "|", as shown:

```
<Table>
Header 1 | Header 2 | Header 3
----
Cell 1  | Cell 2    | Cell 3
</Table>
```

This preprocessing ensures clean, consistent, machine-readable data, supporting robust analysis and model training.

C Implementation Details

Single-modality baselines (ARIMA, DLinear, AutoFormer, FEDFormer, iTransformer, PatchTST) were adapted from Time-Series-Library to the event-driven setting, using pre-event segments to predict 35-, 70-, and 140-step trends. All models were trained for 10 epochs with batch size 32.

For multi-modality baselines, we used TEST and a re-implemented GPT4MTS (with LongFormer replacing BERT). Both were aligned to the same forecasting horizons and training settings.

CAMEF followed a two-stage pipeline: MOMENT was pre-trained on time series, then the full model fine-tuned on aligned text–series pairs. MOMENT and RoBERTa (excluding final layer) were frozen, while GPT2 and fusion layers were trainable, enabling robust multimodal forecasting.

C.1 Implementation of Baseline Models

Baseline models were developed with objectives distinct from our event-driven forecasting approach, as single-modality methods are primarily designed for continuous time-series forecasting. To enable a fair comparison, we adapted these models to align with an event-driven context, where past time-series data preceding an event is used to forecast trends within a defined forecasting horizon.

Single-Modality Models: For time-series-based single-modality models, including **DLinear, AutoFormer, FEDFormer, iTransformer, and PatchTST**, we utilized the open-source library **Time-Series-Library**[2]. Time-series segments were extracted with lengths covering both the input and forecasting periods surrounding each event, ensuring alignment with the respective event's time step. The input segment represents the historical trend information, while the forecasting period is treated as "unseen" data to be predicted. Input and forecasting lengths were consistently set to 35, 70, and 140 time steps across all experiments, as described in Sec. 6.1. All models were trained for 10 epochs, with a batch size of 32. These settings ensured convergence of the loss function and maintained consistency with CAMEF's training configuration.

Multi-Modality Models: For multi-modality methods, including **TEST** and **GPT4MTS**, we followed implementation strategies specific to each model.

- **TEST:** We followed the instructions provided in the open-source TEST repository[3], which involved two training stages. In the first stage, the encoder was trained by selecting 10 prototype words based on GPT vocabulary clustering (as instructed in the repository) to align textual representations with time-series data. The second stage involved training the time-series decoder, where input and forecasting steps were set to 35, 70, and 140, similar to the single-modality models. The batch size was set to 7, and training epochs were kept at 10.
- **GPT4MTS:** As the official implementation of GPT4MTS was unavailable, we re-implemented the model based on its original paper. Instead of using BERT [12] as the textual encoder, we employed LongFormer [4], which is better suited for encoding longer contexts, as our texts tend to be relatively lengthy. Input time-series segments were used to forecast

[2] https://github.com/thuml/Time-Series-Library
[3] https://github.com/SCXsunchenxi/TEST

Table 5: Learning Rates for CAMEF Components

Model Component	Learning Rate
MOMENT Model	1×10^{-6}
RoBERTa Model	5×10^{-7}
GPT2 Model	1×10^{-5}
Embedding Layer	1×10^{-5}
Residual Layer	1×10^{-5}
Fusion Layer	5×10^{-7}
Output Layer	1×10^{-5}

the time-series data at the forecasting lengths of 35, 70, and 140. We kept training epochs at 10 and the batch size at 7 to ensure consistency with other baselines.

C.2 Implementation of CAMEF

The batch size is set to 10, and the training epochs are set to 10, consistent with the baseline models.

The implementation of **CAMEF** involves two distinct training stages to ensure effective learning of both time-series and textual features, while leveraging pre-trained components:

(1) **Pre-training MOMENT:** In the first stage, we pre-train the MOMENT model using the time-series segments from the training data. This step focuses on learning representations of the past time-series patterns.

(2) **Training Entire CAMEF:** In the second stage, we train the entire CAMEF model using both event scripts and their corresponding aligned time-series segments. During this stage, certain parameters are frozen to retain the pre-trained knowledge, while others remain open for fine-tuning:
   - **Frozen Parameters:** All parameters of the MOMENT model and RoBERTa (except for the last hidden layer) are frozen. Additionally, the token embedding layer of GPT2 in the decoder is frozen to preserve its pre-trained representations.
   - **Trainable Parameters:** The remaining components of GPT2, the last hidden layer of RoBERTa, and other components such as the embedding, residual, fusion, and output layers are open for training.

The hidden sizes of the different components of the CAMEF model are detailed in Sec. 5. To optimize the model effectively, we adopt component-specific learning rates as listed in Table 5.