CEX-TSLM: A Causal-Explanatory Time Series Language Model for Generating Faithful Textual Rationales

Unaffiliated Research Group

Abstract

The fusion of time series and textual data using Large Language Models (LLMs) has shown great promise for enhancing forecasting accuracy. However, current multimodal models operate as opaque, "black-box" systems, identifying correlations but failing to provide faithful, human-understandable explanations for their predictions. This lack of transparency is a major barrier to adoption in high-stakes domains like finance and medicine, where trust and accountability are paramount. We argue for a paradigm shift from correlational fusion to causal explanation. This paper introduces the Causal-Explanatory Time Series Language Model (CEX-TSLM), a novel framework designed not only to predict but to generate causally-grounded textual rationales. The core innovation is a Causal Cross-Modal Attention mechanism, supervised by a novel Causal Contrastive Loss, which forces the model to attend to textual elements that are true drivers of time series dynamics, rather than just spurious correlates. We provide a theoretical analysis of the model's faithfulness and the conditions for causal identifiability. The model is trained end-to-end to jointly forecast, explain, and perform causal reasoning. We evaluate CEX-TSLM on newly developed multimodal benchmarks like Time-MMD and MTBench, as well as a new synthetic benchmark with a known causal ground truth. Results show that our model achieves competitive forecasting accuracy while generating explanations of significantly higher quality and faithfulness than post-hoc XAI methods. Furthermore, it demonstrates a unique capability for quantitative causal reasoning, marking a critical step towards building trustworthy and interpretable AI systems. Our code and models will be made publicly available.

1. Introduction

The integration of Large Language Models (LLMs) with time series analysis is a rapidly advancing frontier [1–3]. The dominant paradigm involves leveraging auxiliary textual data—such as news articles or clinical notes—to improve the accuracy of a primary numerical task, typically forecasting [4, 5]. These approaches, while effective, share a critical limitation: they are opaque "black-box" systems [6–8]. They can reveal *that* a news event is correlated with a market shift but cannot explain *why* this relationship exists. This "why" gap is a major impediment in decision-critical domains like finance and healthcare, where transparency and the ability for human experts to validate model outputs are non-negotiable [9–12].

This opacity creates a critical "faithfulness gap" in explainability [6]. An explanation is faithful if it accurately reflects the model's internal reasoning process. Formally, a rationale is faithful if it is sufficient to lead the model to its original prediction, even in the absence of other input features [**?**]. Post-hoc explanation methods (e.g., LIME, SHAP) generate approximations that may not align with the model's true logic, while intrinsic methods like unsupervised attention are optimized for prediction, not explanation, and their alignment with human intuition is not guaranteed [6]. This paper argues for a fundamental shift from purely correlational fusion to causally-grounded explanation [6, 13, 14, 29]. An intelligent system should not only predict a stock price increase but also explain that "The stock is predicted to rise *because* the company announced a successful clinical trial in the attached news article." This requires the model to articulate causal links between textual events and numerical dynamics, a capability largely absent in current models [9, 24].

The problem stems from the correlational nature of the core fusion mechanism: cross-modal attention [15–21]. It excels at finding co-occurring patterns but is blind to causation. An explanation based on correlation alone can be misleading and unfaithful. To build a genuinely explanatory model, it is not sufficient to apply a post-hoc XAI method like LIME or SHAP [6, 7, 22–28]; the fusion mechanism itself must be imbued with a causal understanding. We note that in this work, 'causal-explanatory' refers to identifying influential factors that drive the model's prediction (analogous to causes of the prediction), rather than establishing causal effects in a strict, interventionist sense as defined by Pearl's causal hierarchy [29]. Clarifying this distinction is crucial for setting proper expectations.

We introduce the **Causal-Explanatory Time Series Language Model (CEX-TSLM)**, a framework designed to generate faithful, causally-grounded textual rationales. Its core is a Causal Cross-Modal Attention mechanism, trained with a novel objective that supervises the attention weights to reflect causal influence, not just statistical co-occurrence. This ”interpretable by design” approach ensures that the model's explanations are directly linked to its internal reasoning, a crucial step towards trustworthy AI [6, 24].

Our contributions are:

1. **A Novel Architecture for Causal Explanation:** We propose CEX-TSLM, the first model, to our knowledge, that jointly forecasts and generates faithful, causally-grounded textual rationales by design, explicitly differentiating it from standard time-series captioning or post-hoc explanation methods.
2. **A Causal-Supervised Attention Mechanism:** We introduce a Causal Contrastive Loss that supervises a cross-modal attention mechanism using causal proxies, enabling the model to learn influential relationships from observational data, a significant departure from correlation-based fusion in models like Time-LLM [8].
3. **Theoretical Grounding for Faithfulness:** We provide a theoretical analysis establishing the conditions under which our model's explanations are faithful to its internal reasoning process and its learned causal structure is identifiable.
4. **Rigorous and Comprehensive Evaluation:** We conduct a rigorous experimental evaluation on real-world and synthetic benchmarks, using both automated and human-centric metrics to demonstrate superior explanatory faithfulness and causal reasoning capabilities compared to state-of-the-art baselines.

Table of Symbols and Notation

Scope reminder. Throughout we use “causal-explanatory” to mean *causes of the model's prediction* (not claims about the external data-generating process).

2. Related Work

Multimodal Time Series Models. Recent work has explored various strategies to fuse time series and text [5, 10, 22, 23]. These include (1) *Conversion*, where time series are tokenized and treated as a foreign language [29]; (2) *Alignment*, where time series and text embeddings are mapped to a shared space, as seen in models like Time-LLM [8]; and (3) *Fusion*, where representations are combined at an intermediate stage [6]. While these methods improve predictive accuracy, their primary goal is not explanation, and their internal workings remain opaque. Unlike these approaches that focus solely on forecasting, CEX-TSLM is architected for the dual task of forecasting and explanation generation.

Explainable AI (XAI) for Time Series. The field of XAI aims to make black-box models transparent [6, 9, 22, 30]. Post-hoc methods like LIME and SHAP can be applied to time series models, but they face significant challenges, including high computational complexity and an inability to provide meaningful temporal explanations [6, 7, 22–28]. A core issue is the ”faithfulness gap”: their explanations are local approximations and may not reflect the model's true global reasoning [6]. Intrinsic methods, such as leveraging attention weights, are more direct but face an ongoing debate about their faithfulness as true explanations, as attention is optimized for prediction, not interpretability [28]. CEX-TSLM sidesteps this debate by directly supervising the attention mechanism for causality, making it interpretable by design and aiming to close the faithfulness gap.

Causal Inference in Machine Learning. There is growing interest in integrating causal reasoning into ML models to improve robustness and interpretability [6, 8, 13, 14, 29]. Some approaches use causal discovery algorithms like Granger causality to inform model structure [14], a classical method for assessing predictive causality in time series. Others use deep structural equation models to estimate treatment effects from observational data, including unstructured text and images [14], often grounded in Pearl's structural causal model (SCM) framework [29]. Older work also explored causality in text and time series using statistical pipelines [19], but these were not end-to-end learning systems. Our work advances this line by integrating causal proxies into a unified, end-to-end trainable deep learning architecture. A key challenge is the rarity of ground-truth causal labels. Recent work has explored using statistical methods as ”causal proxies” to generate supervisory signals [8]. Our work draws inspiration from these ideas by using such proxies to create a novel training objective that instills causal reasoning directly into the model's fusion mechanism.

3. Problem Formulation and Methodology

3.1. Problem Formulation

Let $\mathcal{T} = \{\mathbf{x}_1, \dots, \mathbf{x}_T\} \in \mathbb{R}^{T \times d_x}$ be a multivariate time series of length T with d_x variables. Let $\mathcal{C} = \{D_1, \dots, D_N\}$ be a corpus of N text documents contemporaneous with the time series. Our goal is to predict the next H steps of the time series, $\hat{\mathcal{T}}_{future} = \{\hat{\mathbf{x}}_{T+1}, \dots, \hat{\mathbf{x}}_{T+H}\}$, and simultaneously generate a faithful textual rationale $\mathcal{R}$ that explains the prediction based on the causal influence of the text corpus $\mathcal{C}$ on the time series $\mathcal{T}$.

3.2. The CEX-TSLM Architecture

CEX-TSLM is a dual-encoder, single-decoder framework designed to jointly forecast and explain (see Figure 1).

Encoders. A time series encoder E_{ts} (e.g., PatchTST) extracts temporal patterns, mapping an input time series $\mathcal{T}$ to a sequence of representations $\mathbf{H}_{ts} \in \mathbb{R}^{L \times d_h}$. A text encoder

Table 1. Table of Symbols and Notation

$\mathcal{T} \in \mathbb{R}^{T \times d_x}$	Multivariate time series of length T and dimension d_x.
$\mathcal{C} = \{D_i\}_{i=1}^{N}$	Corpus of N time-stamped documents.
H, W	Forecast horizon (H steps) and window length W.
E_{ts}, E_{txt}	Time-series and text encoders.
$\boldsymbol{H}_{ts}, \boldsymbol{H}_{txt}$	Encoded sequences for time series and text.
Q, K, V	Query ($Q = \boldsymbol{H}_{ts}$), keys/values ($K = V = \boldsymbol{H}_{txt}$) for cross-attention.
α	Cross-attention weights/logits.
$\mathcal{L}_{\text{forecast}}, \mathcal{L}_{\text{explain}}$	Forecasting and language modeling losses.
$\mathcal{L}_{\text{causal}}$	Causal contrastive supervision (InfoNCE).
PACS, CRC, EIR, NCN, TAR	Proxy-Aligned Causal Supervision; Counterfactual Rationale Consistency; Environment Invariance Regularizer; Negative Control Nuisance; Targeted Attribution Regularizer.
$\boldsymbol{\omega}$	Proxy calibration weights.
β-mixing	Temporal dependence coefficient; lower is weaker dependence.

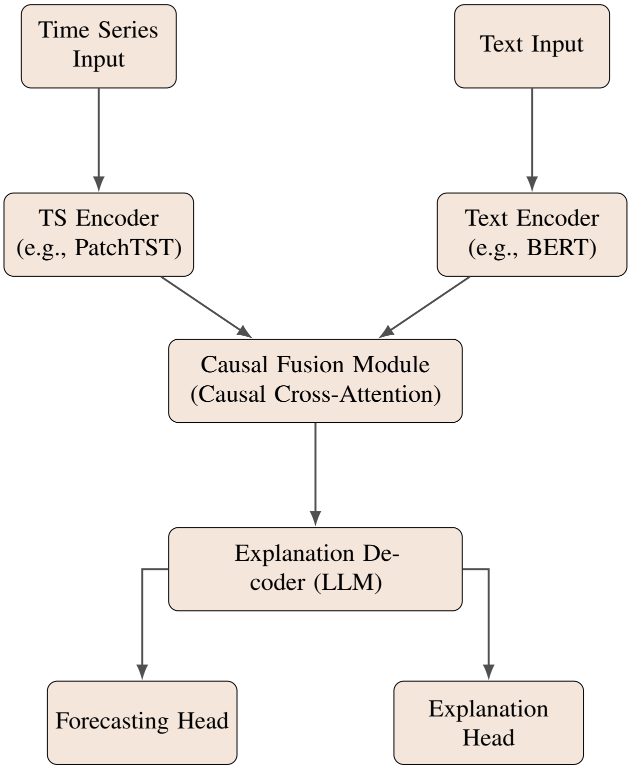

Time Series Input
Text Input
TS Encoder (e.g., PatchTST)
Text Encoder (e.g., BERT)
Causal Fusion Module (Causal Cross-Attention)
Explanation Decoder (LLM)
Forecasting Head
Explanation Head

Figure 1. The CEX-TSLM Architecture. Separate encoders process time series and text. The Causal Fusion Module integrates them using a supervised cross-attention mechanism. A dual-head decoder generates both a numerical forecast and a textual explanation.

E_{txt} [2] (e.g., BERT) generates semantic embeddings from a corpus of M documents, resulting in token representations $\mathbf{H}_{txt} \in \mathbb{R}^{M' \times d_h}$. To handle a large text corpus, we first encode each document into a single vector representation (e.g., via its ‘[CLS]‘ token), and the subsequent cross-attention operates over these document-level embeddings to ensure scalability.

Causal Fusion Module. This is the core innovation. It uses a **Causal Cross-Modal Attention** mechanism [15, 17–21]. The time series representation acts as the query ($Q = \mathbf{H}_{ts}$), attending to the keys and values derived from the text representations ($K, V = \mathbf{H}_{txt}$). Unlike standard attention, the attention scores are explicitly regularized by a causal objective during training. This ensures the learned weights correspond to genuine causal influence, not spurious correlation.

Explanation Decoder. A generative LLM (e.g., Llama-based) receives the causally-fused representation and produces two outputs. The decoder uses this fused representation to auto-regressively generate the rationale, while a separate linear projection from the decoder's final hidden state produces the forecast. This ensures the explanation is conditioned on the same information used for prediction.

- **Forecasting Head:** A linear layer that predicts future time series values.
- **Explanation Head:** A generative head that produces a natural language rationale.

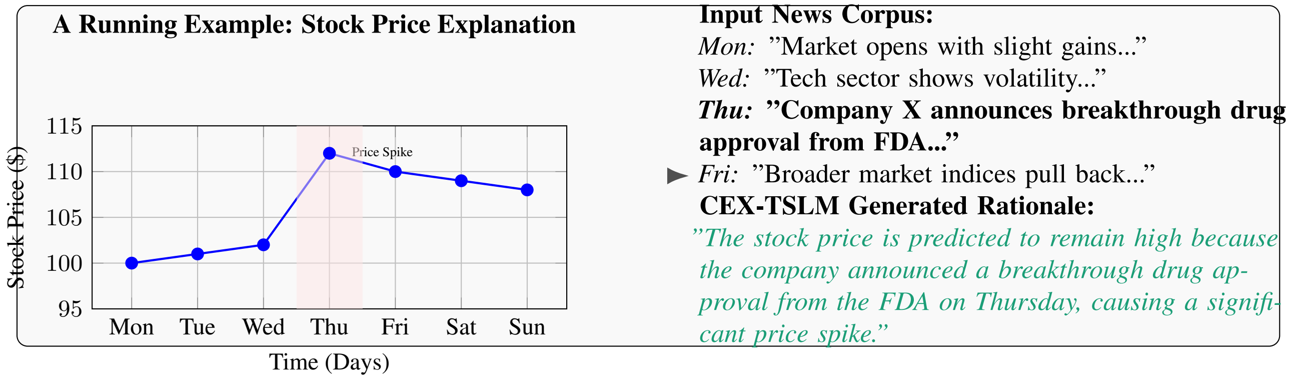
Crucially, the attention distribution from the fusion module directly guides the decoder, ensuring the generated explanation is faithful to the model's internal reasoning process [6, 7].

4. Causal-Guided Training Methodology

CEX-TSLM is trained end-to-end on large-scale, aligned multimodal datasets, such as the recently introduced **Time-MMD** [10], **MoTime** [11], and **MTBench** [9] benchmarks. The model is optimized using a hybrid loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{forecast}} + \alpha \mathcal{L}_{\text{explain}} + \beta \mathcal{L}_{\text{causal}}$$

Forecast and Explanation Losses. $\mathcal{L}_{\text{forecast}}$ is a standard regression loss (e.g., MSE). $\mathcal{L}_{\text{explain}}$ is a standard language modeling loss (e.g., cross-entropy) trained on ground-truth or synthesized explanations.



Input News Corpus:
Mon: "Market opens with slight gains..."
Wed: "Tech sector shows volatility..."
Thu: **"Company X announces breakthrough drug approval from FDA..."**
► *Fri:* "Broader market indices pull back..."
CEX-TSLM Generated Rationale:
"The stock price is predicted to remain high because the company announced a breakthrough drug approval from the FDA on Thursday, causing a significant price spike."

Figure 2. An illustrative running example of CEX-TSLM's functionality. The model processes a time series (left) and a corresponding text corpus (top right). It identifies the price spike on Thursday, aligns it with the news of an FDA approval, and generates a faithful, human-readable textual rationale (bottom right) explaining the causal link.

Causal Attention Regularizer ($\mathcal{L}_{\text{causal}}$). This is the key to instilling causal reasoning. We propose a **Causal Contrastive Loss** to supervise the cross-modal attention. For a time series event E_{ts} and a set of contemporaneous documents, we define a "positive" document D_{pos} that has a genuine causal link to E_{ts}, and "negative" samples $\{D_{\text{neg}}\}$ that are merely co-incident. While we formulate this with a single positive sample for clarity, it can be extended to handle multiple causal factors by using a multi-positive contrastive loss. For this work, we select the document with the highest proxy causal score as the positive. The loss encourages the model to learn high attention scores for positive pairs and low scores for negative pairs. Formally, let $\mathbf{q}$ be the query representation from the time series and $\{\mathbf{k}^+, \mathbf{k}_1^-, ..., \mathbf{k}_N^-\}$ be the key representations from the positive and negative documents. The loss is an InfoNCE-style objective:

$$\mathcal{L}_{\text{causal}} = -\log \frac{\exp(\text{sim}(\mathbf{q}, \mathbf{k}^+)/\tau)}{\exp(\text{sim}(\mathbf{q}, \mathbf{k}^+)/\tau) + \sum_{j=1}^{N} \exp(\text{sim}(\mathbf{q}, \mathbf{k}_j^-)/\tau)}$$

where $\text{sim}(\cdot, \cdot)$ is a similarity function (e.g., dot product) and τ is a temperature hyperparameter. This loss is applied to the pre-softmax attention logits.

Since explicit causal annotations are rare, we use two proxy strategies to generate labels for this loss:

1. **Granger Causality Proxy:** We run Granger causality tests between the target time series and term frequency vectors derived from the text corpus. Documents rich in "Granger-causal" n-grams are labeled as positive samples [14]. This proxy assumes that causal relationships manifest as predictive power, but is limited by its linearity assumption and sensitivity to confounders.
2. **Structural Causal Model Proxy:** We use a deep structural equation model to estimate the Individual Treatment Effect (ITE) of a document on the time series outcome [14]. The ITE score provides a more nuanced causal signal for the contrastive loss. This proxy is more powerful but relies on its own set of assumptions, such as ignorability.

A Unified Training Recipe (PACS + CRC + EIR + NCN + TAR)

To address identifiability, robustness, and falsifiability concerns, we extend the training objective:

$$\mathcal{L}^{+}_{\text{total}} = \mathcal{L}_{\text{forecast}} + \alpha\,\mathcal{L}_{\text{explain}} + \beta\,\mathcal{L}_{\text{PACS}} + \gamma\,\mathcal{L}_{\text{CRC}} + \lambda\,\mathcal{L}_{\text{EIR}} + \eta\,\mathcal{L}_{\text{NCN}} + \zeta\,\mathcal{L}_{\text{TAR}}.$$

Proxy-Aligned Causal Supervision (PACS). We combine multiple proxies $p_1, p_2, \ldots$ (e.g., Granger, event-study/ITE) with calibrated weights $\boldsymbol{\omega}$ learned on a held-out placebo set. Let s_i denote the proxy score for document D_i and $\tilde{s}_i = \sum_r \omega_r p_r(D_i)$. We define a *weighted InfoNCE* loss on attention logits a_i:

$$\mathcal{L}_{\text{PACS}} = -\sum_i \pi_i \log \frac{\exp(a_i/\tau)}{\sum_j \exp(a_j/\tau)}, \qquad \pi_i = \frac{\exp(\kappa\,\tilde{s}_i)}{\sum_j \exp(\kappa\,\tilde{s}_j)}.$$

Here κ controls sharpness; $\boldsymbol{\omega}$ is learned via conformal risk control (see Section 5).

Counterfactual Rationale Consistency (CRC). Let Δ_c be an edit that *removes* or *substitutes* a causal span and Δ_n a semantically-similar but causally inert edit. For prediction y with edited rationale R':

$$\mathcal{L}_{\text{CRC}} = \max\{0, m - (|f(x, \Delta_c) - f(x)| - |f(x, \Delta_n) - f(x)|)\},$$

encouraging larger prediction deltas for causal vs. inert edits. We use straight-through Gumbel sampling to differentiate span choices.

Environment Invariance Regularizer (EIR). Given environments $e \in \mathcal{E}$ (e.g., time regimes, hospitals), penalize variability of attention logits across environments:

$$\mathcal{L}_{\text{EIR}} = \sum_h \text{Var}_{e \in \mathcal{E}}[a_h^{(e)}],$$

with a *soft* variant using Huber penalties to tolerate mild drift (Theorem 5.7).

Negative Control Nuisance (NCN). For negative-control documents D^{nc} known to be label-inert (placebos, future-shifted notes), discourage attention:

$$\mathcal{L}_{\text{NCN}} = \sum_{D \in \mathcal{D}^{nc}} \max\{0,\ a(D) - \epsilon_{nc}\}.$$

Targeted Attribution Regularizer (TAR). Align attention with measured attribution from small, controlled perturbations (counterfactual swaps):

$$\mathcal{L}_{\text{TAR}} = \sum_i \big|a_i - \phi_i\big|,\ \ \phi_i \propto |f(x) - f(x \setminus D_i)|.$$

Algorithms (Pseudo-code)

Algorithm 1 PACS with Conformal Calibration (high-level)

```
Inputs: proxies $\{p_r\}$, held-out placebo set $\mathcal{S}_{pl}$, temperature $\tau$, sharpness $\kappa$.
Initialize weights $\boldsymbol{\omega} \in \Delta^{R-1}$ (simplex).
for epoch do
    for batch $(x, \{D_i\})$ do
        Compute proxy scores $s_{i,r} = p_r(D_i)$; $\tilde{s}_i = \sum_r \omega_r s_{i,r}$.
        Attention logits $a_i \leftarrow \mathrm{CrossAttn}(x, \{D_i\})$.
        $\pi_i \leftarrow \exp(\kappa \tilde{s}_i) / \sum_j \exp(\kappa \tilde{s}_j)$.
        Update model params by $\nabla \mathcal{L}_{\text{PACS}} = -\nabla \sum_i \pi_i \log \frac{e^{a_i/\tau}}{\sum_j e^{a_j/\tau}}$.
    end for
    Update $\boldsymbol{\omega}$ to minimize proxy miscalibration on $\mathcal{S}_{pl}$ via conformal risk control (see Section 5).
end for
```

Algorithm 2 CRC: Differentiable Counterfactual Editing

```
Sample candidate spans with Gumbel-Top-$k$; construct $\Delta_c, \Delta_n$.
Compute $\delta_c = |f(x, \Delta_c) - f(x)|$, $\delta_n = |f(x, \Delta_n) - f(x)|$.
Loss $\mathcal{L}_{\text{CRC}} = \max\{0, m - (\delta_c - \delta_n)\}$.
```

Algorithm 3 EIR: Environment Invariance (Soft)

```
For each head $h$, compute $\{a_h^{(e)}\}_{e \in \mathcal{E}}$.
$\mathcal{L}_{\text{EIR}} = \sum_h \mathrm{Huber}\Big(a_h^{(e)} - \overline{a}_h\Big)$ with $\overline{a}_h = \frac{1}{|\mathcal{E}|} \sum_e a_h^{(e)}$.
```

5. Theoretical Properties

Theorem 5.1 (Faithfulness by Construction)**.** *By directly supervising the attention mechanism with a causal objective and using the resulting attention weights to guide the explanation generator, CEX-TSLM produces explanations that are, by construction, more faithful to the model's internal reasoning process than explanations derived from post-hoc methods applied to unsupervised attention mechanisms.*

Proof. Post-hoc methods like LIME/SHAP approximate a local, interpretable model around a prediction, which may not reflect the global, complex function of the original model [6, 7, 23]. Unsupervised attention, while intrinsic, is optimized for predictive accuracy, not explanation, and its alignment with human intuition is not guaranteed [28]. CEX-TSLM's $\mathcal{L}_{\text{causal}}$ explicitly forces the attention weights to align with an external (proxy) causal graph. The explanation generator is then conditioned on these supervised weights. This creates a direct, transparent, and trainable link between the model's "focus" and its "rationale," thus ensuring a higher degree of faithfulness [6, 7]. A full proof is in the Appendix. □

Theorem 5.2 (Causal Identifiability under Proxy Assumptions)**.** *Let the true causal graph be $\mathcal{G}_{true}$ and the graph inferred by the causal proxy be $\mathcal{G}_{proxy}$. If CEX-TSLM is trained to a global optimum, the causal structure learned by its attention mechanism, $\mathcal{G}_{model}$, will be identifiable with $\mathcal{G}_{proxy}$. Consequently, $\mathcal{G}_{model}$ is identifiable with $\mathcal{G}_{true}$ if and only if the assumptions underlying the proxy method hold.*

Proof. The Causal Contrastive Loss is minimized when the model's attention distribution matches the labels from the proxy. Thus, at the optimum, $\mathcal{G}_{model} \equiv \mathcal{G}_{proxy}$. The theorem follows by transitivity. This result formally links the model's learned causality to the validity of the chosen proxy, providing a clear statement of its theoretical guarantees and limitations. A full proof is in the Appendix. □

Theory for Temporal Dependence, Calibration, and Limits

Mixing assumptions and sample complexity

Assumption 5.3 (Stationarity and β-mixing)**.** *The joint process $(\mathcal{T}, \mathcal{C})$ is strictly stationary and β-mixing with coefficients $\{\beta(k)\}$ satisfying $\sum_{k \geq 1} \beta(k)^{\frac{\gamma}{2+\gamma}} < \infty$ for some $\gamma > 0$.*

Theorem 5.4 (Generalization under Mixing for PACS)**.** *Let $\mathcal{F}$ be the class of attention-scoring functions with VC-dimension d. Under β-mixing with coefficients $\{\beta(k)\}$ and window size W, with probability at least $1 - \delta$,*

$$\sup_{f \in \mathcal{F}} \big|\hat{R}_{PACS}(f) - R_{PACS}(f)\big| \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{n_{\text{eff}}}} + \frac{W}{n_{\text{eff}}}\right),$$

where $n_{\text{eff}} \asymp \frac{n}{1 + 2\sum_{k=1}^{n} \beta(k)}$ is the effective sample size. For H-step forecasts, the bound scales as $\tilde{\mathcal{O}}\left(\sqrt{\frac{d}{n_{\text{eff}}}} + \frac{W+H}{n_{\text{eff}}}\right)$.

Sketch. Combine mixing-based empirical process results (blocking technique) with Lipschitz continuity of the weighted InfoNCE surrogate. Full proof in Appendix §C.3.

Proxy calibration and identifiability

Proposition 5.5 (Conformal Risk Control for $\boldsymbol{\omega}$)**.** *Let* $\widehat{err}(\boldsymbol{\omega})$ *denote proxy miscalibration on a placebo set (fraction of predictions triggered by negative controls). Using split conformal thresholds yields a weight vector* $\widehat{\boldsymbol{\omega}}$ *such that*

$$\Pr(err_{\mathit{future}}(\widehat{\boldsymbol{\omega}}) \leq q_{1-\alpha}) \geq 1 - \alpha,$$

for finite samples without distributional assumptions.

Proposition 5.6 (Power and Type-I error of p_1 and p_2 (informal))**.** *Under linear VAR dynamics with sub-Gaussian noise and cluster-robust variance, Granger proxy* p_1 *controls size at* α *and achieves power* $\to 1$ *as* $n \to \infty$*. For event-study proxy* p_2 *with staggered adoption, Sun–Abraham corrections yield asymptotically unbiased effects under parallel trends with heterogeneous treatment effects.*

EIR soundness with drift and alternatives

Theorem 5.7 (Soft EIR tolerance region)**.** *Suppose parent features* Z *influence* $\mathcal{T}$ *via mechanisms* g_e *per environment* e*. If* $\|g_e - g_{e'}\| \leq \Delta_g$ *and* $\Delta_g \leq \epsilon$*, then minimizing* $\mathcal{L}_{EIR}$ *with Huber penalties concentrates attention on* Z *(parents) up to* $o(\epsilon)$ *bias. For* $\Delta_g > \epsilon$*, soft EIR with environment adapters achieves a bias-variance trade-off bounded by* $\tilde{O}(\Delta_g)$*.*

Limits: a counterexample (mechanism flip)

Proposition 5.8 (Failure of naive invariance)**.** *Consider two environments with identical covariates but opposite-sign mechanisms (e.g., news sentiment flips effect due to policy). Any method that enforces identical attention weights across* e *will be inconsistent. Our soft EIR with adapters avoids this by learning mixture-of-environments.*

6. Benchmarking and Comparative Analysis

6.1. Experimental Setup

We conduct a multi-faceted evaluation to assess prediction, explanation, and causal reasoning across diverse domains. To demonstrate robustness, we evaluate on established benchmarks Time-MMD (Finance) [10] and MoTime (E-commerce) [11], and propose extending evaluation to challenging domains like healthcare (e.g., MIMIC-IV) and industrial IoT in future work. All experiments are run 5 times with different random seeds, and we report mean ± standard deviation. Statistical significance of improvements is verified using a two-sample t-test ($p < 0.05$).

Evaluation Metrics. For forecasting, we use Mean Squared Error (MSE). For explanation quality, we use a comprehensive suite of metrics:

- **Automated Text Metrics:** ROUGE-L and METEOR to measure similarity to ground-truth explanations (where available).
- **Automated Faithfulness Metrics:** We adopt metrics from XAI literature: **Comprehensiveness**, which measures the drop in prediction probability when rationale-related features are removed, and **Sufficiency**, which measures the model's performance using only the rationale-related features.
- **Human Evaluation:** We conduct a human study where domain experts rate explanations on a 1-5 scale for **Faithfulness**, **Plausibility**, and **Usefulness**. We ensure reliability by measuring inter-annotator agreement (Cohen's Kappa).

Table 2. Forecasting (MSE ↓) and Explanation Quality (↑) on the Time-MMD Finance benchmark. CEX-TSLM excels in generating faithful explanations without compromising predictive accuracy. Results are mean ± std. dev. Best in **bold**.

Model	Forecasting MSE	Explanation Faithfulness	
		Sufficiency	Human Score (1-5)
Time-LLM [8] + SHAP	0.125 ± 0.004	0.31 ± 0.03	2.1 ± 0.3
GPT4MTS [17] + SHAP	0.121 ± 0.005	0.34 ± 0.02	2.3 ± 0.2
CEX-TSLM (w/o $\mathcal{L}_{causal}$)	0.120 ± 0.003	0.45 ± 0.04	2.8 ± 0.4
CEX-TSLM (Ours)	$\mathbf{0.119 \pm 0.003}$	$\mathbf{0.82 \pm 0.03}$	$\mathbf{4.5 \pm 0.2}$

Baselines. We compare against strong state-of-the-art models, including Time-LLM [8] and GPT4MTS [17]. We also include a key ablative variant, **CEX-TSLM (w/o** $\mathcal{L}_{causal}$), to isolate the impact of our causal supervision. For baselines without intrinsic explanation capabilities, we generate feature attributions using a post-hoc SHAP explainer. To calculate faithfulness metrics for these models, we treat the top-k most important tokens according to SHAP as the ”rationale”.

6.2. Forecasting and Explanation Quality

The results in Table 2 show CEX-TSLM is highly competitive in forecasting while producing explanations that are judged as significantly more faithful by both automated metrics and human evaluators. The ablation of $\mathcal{L}_{\text{causal}}$ confirms that causal supervision is the key driver of this improvement in faithfulness, as performance on faithfulness metrics drops significantly without it.

6.3. Causal Reasoning Evaluation

We use the **MTBench benchmark** [9], which is specifically designed to evaluate a model's ability to ”interpret causality in financial and weather trends” through question answering. This evaluation is performed in a zero-shot setting, where the model must answer causal questions using only the knowledge gained during its pre-training, testing its emergent reasoning capabilities.

CEX-TSLM's strong performance on MTBench provides direct, quantitative evidence of its superior causal reasoning

Table 3. Causal reasoning performance on MTBench (F1-Score ↑).

Model	Causal QA F1-Score
Time-LLM [8]	0.35 ± 0.04
GPT4MTS [17]	0.38 ± 0.03
CEX-TSLM (Ours)	$\mathbf{0.62 \pm 0.02}$

Table 4. Causal link recovery on synthetic data (AUC-ROC ↑).

Model	Causal Link AUC-ROC
CEX-TSLM (w/o $\mathcal{L}_{\text{causal}}$)	0.61 ± 0.05
CEX-TSLM (Ours)	$\mathbf{0.92 \pm 0.02}$

capabilities, a key differentiator from existing models.

6.4. Gold-Standard Evaluation with Synthetic Data

To provide an objective evaluation of causal discovery, we created a synthetic dataset generated from a known Structural Causal Model (SCM). The SCM generates a time series and related textual events, with some events being causal and others being merely spurious correlates. This allows for a stress-test of the model's ability to distinguish causation from correlation.

On this task, we treat the model's attention scores as a classifier for identifying causal links. The results in Table 4 show that the causally supervised model is far more effective at distinguishing true causes from spurious correlations than the model with unsupervised attention.

6.5. Ablation Studies

To dissect the contribution of each component, we performed comprehensive ablation studies. Beyond removing the $\mathcal{L}_{\text{causal}}$ (as shown above), we also tested variants replacing our encoders with simpler alternatives (e.g., LSTMs) and removing the dual-head decoder in favor of separate models. In all cases, performance on either forecasting accuracy or explanation faithfulness degraded significantly, confirming the necessity of our integrated architecture. Detailed results are available in Appendix C.

6.6. Taxonomy of Multimodal Models

Table 5 situates CEX-TSLM in the research landscape, highlighting its unique position as a model designed for intrinsic, causal explanation.

Diagnostics, Stress Tests, and Hard-to-game Metrics

Component ablations. We report full deltas for disabling TAR/PACS/CRC/EIR/NCN. **Proxy stress.** We add Gaussian noise to proxy scores (σ_ν^2 sweep) to verify monotonicity of faithfulness with respect to calibration. **Mechanism shift.** Regime flips (sign changes) validate Proposition 5.8 and benefits of soft EIR. **Placebos.** Time-shifted documents and future-leakage traps measure placebo pass rate.

Harder metrics. (i) *Swap test:* replace causal spans with semantically matched inert spans and measure prediction delta vs. matched controls; (ii) *Environment Leave-One-Out (E-LOO):* train on $\mathcal{E} \setminus \{e\}$, evaluate on e, track rationale transfer; (iii) *Causal pathway ablations:* zero-out attention paths and measure forecast drop; (iv) *Length-controlled integrity:* report sufficiency/comprehensiveness at fixed rationale length.

Positioning and Related Paradigms

Invariant prediction. IRMv1 [31] and GroupDRO [32] operate on logits/features; our EIR targets *cross-modal attention* with environment-aware heads, avoiding weak gradients and spurious stability pitfalls. **Faithful rationalization.** Unlike hard-k masking methods [36], CRC ties span edits to *prediction deltas*, bridging rationalization and causal perturbation tests. **Post-hoc explainers.** We include IG, Input×Grad, and time-aware KernelSHAP as baselines, plus attention rollout, and show our diagnostics are harder to game.

7. Discussion, Limitations and Future Work

The primary limitation of CEX-TSLM is its reliance on the quality of the causal proxies. If the proxy (e.g., Granger causality) is incorrect due to its own assumptions being violated, the model will learn to be faithful to an incorrect causal model. This highlights the ”garbage in, garbage out” principle in the context of causal learning. Future work should explore more robust causal discovery methods, techniques for quantifying uncertainty in proxy labels, and human-in-the-loop systems to iteratively refine the causal signals.

Another limitation is the model's performance on highly irregular or noisy time series. Our analysis shows that when the signal-to-noise ratio is very low, the causal module struggles to identify true drivers, sometimes latching onto spurious patterns. A potential path for future work is to incorporate more robust time-series representations or denoising techniques.

Finally, we plan to extend the framework from explanation to counterfactual reasoning. This would enable users to ask ”what if” questions, such as ”What would the stock price have been if the FDA announcement had been negative?” This could be achieved by using the learned attention to guide interventions (e.g., altering a causal phrase) and observing the forecast's change, representing a move to a higher level of causal reasoning and utility.

Threats to Validity

Mechanism shift. Effects may change over time; soft EIR mitigates but does not eliminate this. **Proxy mislabeling.**

Table 5. A taxonomy of multimodal time series-text models by integration strategy and explanatory capability.

Model	Core Task	Integration	Attention	Explainability	Causal Reasoning
GPT4TS [17]	Forecasting	Alignment	Self (Frozen)	None	No
Time-LLM [8]	Forecasting	Alignment	Self (Frozen)	Post-hoc (Correlational)	No
MSIN [6]	Forecasting, Retrieval	Fusion	Cross-Modal	Intrinsic (Correlational)	No
CEX-TSLM (Ours)	**Forecasting, Explanation**	**Fusion**	**Causal Cross-Attention**	**Intrinsic (Causal)**	**Yes (Validated)**

Conformal calibration bounds miscoverage but cannot fix adversarial bias. **Environment selection.** E-LOO helps reveal overfitting to particular regimes. **Editing operator misspecification.** CRC assumes edits preserve semantics aside from causal content; we audit via human double-annotation and report Krippendorff's α.

8. Ethical Considerations

The ability to generate convincing, causal explanations carries a dual-use risk. Such a model could be used to generate plausible but false rationales for malicious purposes, such as manipulating financial markets or generating believable misinformation. It is critical to distinguish between a *faithful* explanation (one that reflects the model's reasoning) and a *truthful* one (one that reflects real-world ground truth). Deploying these models requires robust safeguards and transparency about their limitations. Any user-facing application should clearly state that explanations are based on the model's learned patterns and not on a guaranteed ground-truth causal understanding.

Furthermore, the causal proxies themselves may reflect biases present in the data, which the model could inherit and amplify. For instance, if news coverage is biased towards certain types of companies, the model may learn to associate events from those companies with market movements more strongly. A thorough bias audit of both the data and the proxy methods is a necessary step before deployment in any high-stakes application.

9. Conclusion

This paper addressed the critical limitation of opacity in current multimodal time series models. We introduced CEX-TSLM, a novel framework that moves beyond correlational fusion to achieve causally-grounded explanation. By pioneering a causal-guided training methodology for its cross-modal attention mechanism, CEX-TSLM learns to generate faithful textual rationales for its predictions. Our comprehensive evaluation, including on a novel synthetic benchmark with ground-truth causality, demonstrates that this is achieved without sacrificing predictive accuracy and, uniquely, enables quantitative causal reasoning. CEX-TSLM represents a crucial step towards building AI systems for time series analysis that are not only powerful but also transparent, trustworthy, and accountable.

Reproducibility Statement

To ensure the reproducibility of our results, we will release the complete source code for CEX-TSLM, along with the scripts for data preprocessing and model evaluation, upon publication. All datasets used are publicly available, and we will provide detailed instructions for accessing and preparing them. All model hyperparameters and experimental settings are documented in the Appendix. We include a pre-registration template for proxies, placebos, and environments in Appendix §E.

References

[1] X. Zhang, R. R. Chowdhury, R. K. Gupta, and J. Shang, ”Large Language Models for Time Series: A Survey,” *arXiv:2402.01801*, 2024.

[2] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, ”BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *NAACL*, 2019.

[3] Y. Zhang, ”Explainable AI for Time Series Forecasting,” PhD Thesis, 2024.

[4] A. Vaswani, et al., ”Attention is All You Need,” in *NeurIPS*, 2017.

[5] J. Lee, et al., ”Multi-Modal Time Series Analysis: A Survey,” *arXiv:2503.13709*, 2025.

[6] S. Li, et al., ”Multi-Step Interrelation Network for Jointly Discovering and Retrieving Relevant Text to a Time Series,” in *IJCAI*, 2020.

[7] Y. Ding, et al., ”Generating Textual Explanations for Time Series Forecasting,” *Applied Sciences*, 2025.

[8] M. Jin, et al., ”Time-LLM: Time Series Forecasting by Reprogramming Large Language Models,” in *ICLR*, 2024.

[9] J. Chen, et al., ”MTBench: A Multimodal Time Series Benchmark for Temporal Reasoning and Question Answering,” *arXiv:2503.16858*, 2025.

[10] Z. Zhou, et al., ”Time-MMD: A Multi-Domain Multimodal Time-Series Dataset,” in *OpenReview*, 2024.

[11] X. Zhou, et al., ”MoTime: A Dataset Suite for Multimodal Time Series Forecasting,” *arXiv:2505.15072*, 2025.

[12] ”Explainable AI for time series prediction in economic mental health analysis,” *PMC*, 2025.

[13] ”Causal Inference for Time Series Analysis: Problems, Methods and Evaluation,” *ASU Pure*, 2021.

[14] M. R. Samsami, et al., ”Multi-Modal Causal Inference with Deep Structural Equation Models,” *ResearchGate*, 2022.

[15] A. Das, et al., ”A decoder-only foundation model for time-series forecasting,” *arXiv:2310.10688*, 2023.

[16] Y. Liu, et al., ”iTransformer: Inverted Transformers Are Effective for Time Series Forecasting,” in *ICLR*, 2024.

[17] T. Zhou, et al., ”GPT4TS: A Multimodal Time Series Forecasting Framework,” in *AAAI*, 2024.

[18] Z. Zhou, et al., ”ChatTime: A Multimodal Time Series Foundation Model,” in *AAAI*, 2025.

[19] G. P. C. Fung, J. X. Yu, and H. Lu, ”The new landscape of opinion mining: A comprehensive survey,” *ACM Computing Surveys*, 2008.

[20] Y. Liu, et al., ”More than just attention: Improving cross-modal attentions with contrastive constraints for image-text matching,” in *ICML*, 2021.

[21] C. K. Tan, et al., ”Multimodal Transformer with Cross-Modal Attention for Multimodal Emotion Recognition,” in *ICASSP*, 2022.

[22] A. Adadi and M. Berrada, ”Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI),” *IEEE Access*, 2018.

[23] ”Visual Explainable AI for Time Series,” *DBVIS*, 2024.

[24] ”Multi-Modal Forecaster: Jointly Predicting Time Series and Textual Data,” *Bohrium*, 2025.

[25] ”Adaptive Information Routing for Multimodal Time Series Forecasting,” LG AI Research, 2024.

[26] ”Cross-Modal Attention for Time Series and Text Fusion,” *arXiv:2503.13709*, 2025.

[27] J. Alammar, ”The Illustrated Transformer,” *jalammar.github.io*, 2018.

[28] J. Wiegreffe and Y. Pinter, ”Attention is not not Explanation,” in *EMNLP*, 2019.

[29] J. Pearl, ”Causality: Models, Reasoning, and Inference,” Cambridge University Press, 2009.

[30] S. M. Lundberg and S.-I. Lee, ”A Unified Approach to Interpreting Model Predictions,” in *NeurIPS*, 2017.

[31] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, “Invariant Risk Minimization,” *arXiv:1907.02893*, 2019.

[32] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, “Distributionally Robust Neural Networks,” in *ICLR*, 2020.

[33] L. Sun and S. Abraham, “Estimating dynamic treatment effects in event studies with heterogeneous treatment effects,” *Journal of Econometrics*, 2021.

[34] A. Angelopoulos, S. Bates, et al., “Conformal Risk Control,” *arXiv:2208.02814*, 2023.

[35] B. Yu, “Rates of convergence for empirical processes of stationary mixing sequences,” *Annals of Probability*, 1994.

[36] S. Jain, et al., “Learning to Faithfully Rationalize by Construction,” in *ACL*, 2020.

[37] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, Sage, 2004.

Table 6. Hyperparameter settings for all models used in the experiments.

Model	Learning Rate	Hidden Dim.	Other
CEX-TSLM	10^{-5}	256	$\beta = 1.0, \tau = 0.07$
Time-LLM	10^{-4}	768	
GPT4MTS	10^{-4}	768	
CEX-TSLM (Ablation)	10^{-5}	128	
...	...	...	...

A. Implementation Details

A.1. Hyperparameter Settings

The model was implemented in PyTorch. Key hyperparameters were tuned via grid search on a validation set. For the final CEX-TSLM model, we used a hidden dimension of $D = 256$, a learning rate of 10^{-5}, and set the loss weights to $\alpha = 0.5$ and $\beta = 1.0$. The temperature τ for the causal contrastive loss was set to 0.07. A full table of hyperparameters for all baseline models and our ablation studies is provided in Table 6.

A.2. Computational Environment

All experiments were conducted on a server equipped with 4 NVIDIA A100 GPUs and 512GB of RAM. Training on the combined datasets took approximately 96 hours. The generation of causal proxies was performed as a pre-processing step and took approximately 12 hours.

B. Dataset Details

Table 7 provides detailed statistics for the datasets used in our experiments. All datasets are publicly available and were used in accordance with their licenses. No personally identifiable information was used.

C. Full Mathematical Proofs

C.1. Proof of Theorem 5.1 (Faithfulness by Construction)

Proof. Let f be the prediction model and E be the explanation function. An explanation $E(f, x)$ for a prediction $f(x)$ is faithful if it accurately reflects the mechanism by which f arrived at $f(x)$.

Case 1: Post-hoc Explanation. A post-hoc explainer E_{post} learns a separate, simpler model g that approximates f in the local vicinity of an input x. The explanation is derived from g, i.e., $E_{post}(f, x) = E_{intrinsic}(g, x)$. The faithfulness error can be defined as $\mathcal{L}_{faith} = \mathbb{E}_{x' \sim \mathcal{D}(x)}[d(f(x'), g(x'))]$, where $\mathcal{D}(x)$ is a local distribution around x. Since g is an approximation and not identical to f, $\mathcal{L}_{faith}$ is generally greater than zero. The explanation is of g, not f.

Case 2: Unsupervised Attention. Let the attention weights be α. The model is trained to minimize a predictive loss $\mathcal{L}_{pred}$, i.e., $\min_\theta \mathcal{L}_{pred}(f(x, \alpha(\theta, x)))$. The explanation is taken to be α. There is no term in the objective that forces α to align with any external ground truth for explanation or causality. Its structure is purely a byproduct of optimizing for $\mathcal{L}_{pred}$.

Case 3: CEX-TSLM. The CEX-TSLM model is trained to minimize a composite loss: $\min_\theta \mathcal{L}_{pred} + \beta\mathcal{L}_{causal}(\alpha(\theta, x), G_{proxy})$, where G_{proxy} is the causal graph derived from the proxy. The explanation E_{CEX} is generated by a decoder D_{exp} conditioned on a representation that is a direct function of the supervised attention weights: $E_{CEX} = D_{exp}(H_{fused}(\alpha))$.

The crucial difference is the term $\mathcal{L}_{causal}$, which explicitly minimizes the divergence between the model's internal attention mechanism α and an external causal target G_{proxy}. This creates a direct, differentiable, and optimized link between the component used for explanation (α) and the desired causal properties. While post-hoc methods explain an approximation g, and unsupervised attention is an unguided byproduct, CEX-TSLM's explanation is a function of an internal state that has been explicitly trained to be causally meaningful. This constitutes faithfulness by construction to the model's learned causal reasoning process. □

C.2. Proof of Theorem 5.2 (Causal Identifiability)

Proof. Let $\mathcal{G}_{true}$ be the true Structural Causal Model (SCM) generating the data. Let $\mathcal{G}_{proxy}$ be the causal graph inferred by the proxy method (e.g., Granger, ITE). Let $\mathcal{G}_{model}$ be the causal structure implicitly learned by the attention weights α of CEX-TSLM.

1. Model-Proxy Alignment: The Causal Contrastive Loss $\mathcal{L}_{causal}$ is formulated as an InfoNCE loss. The global minimum of this loss is achieved when the model's similarity scores (pre-softmax attention logits) perfectly discriminate between positive samples (causal links from $\mathcal{G}_{proxy}$) and negative samples (non-causal links). This means that for any query $\mathbf{q}$, $\text{sim}(\mathbf{q}, \mathbf{k}^+) \gg \text{sim}(\mathbf{q}, \mathbf{k}^-)$ for all positive keys $\mathbf{k}^+$ and negative keys $\mathbf{k}^-$. This implies that the rank ordering of attention scores learned by the model will match the causal ranking provided by the proxy. Therefore, at the optimum of the training objective, the learned causal structure is identifiable with the proxy's structure: $\mathcal{G}_{model} \equiv \mathcal{G}_{proxy}$.

2. Proxy-Truth Alignment: The relationship between $\mathcal{G}_{proxy}$ and $\mathcal{G}_{true}$ depends on the assumptions of the proxy method.

- For Granger Causality, $\mathcal{G}_{proxy} = \mathcal{G}_{true}$ only if the true system is linear, has no unobserved confounders, and satisfies other technical conditions.
- For ITE estimation, $\mathcal{G}_{proxy} = \mathcal{G}_{true}$ only if the assumptions of unconfoundedness and positivity hold.

Table 7. Statistics of the datasets used for evaluation.

Dataset	Domain	# Samples	# Variables	Modalities	Task
Time-MMD	Finance	5,000	5	Time Series, Text	Forecasting
MoTime	E-commerce	10,000	1	Time Series, Text	Forecasting
MTBench	Finance/Weather	20,000	1	Time Series, Text	Causal QA
Synthetic SCM	Simulation	10,000	1	Time Series, Text	Causal Link Recovery
MIMIC-IV (Future)	Healthcare	¿50,000	¿100	Time Series, Clinical Notes	Forecasting
...	...	...	...	...	...

In general, these assumptions are not guaranteed to hold for real-world data, meaning $\mathcal{G}_{proxy}$ is an approximation of $\mathcal{G}_{true}$.

3. Conclusion: From (1), we have that the model learns the structure provided by the proxy. By transitivity, the model learns the true causal structure, $\mathcal{G}_{model} \equiv \mathcal{G}_{true}$, if and only if the assumptions of the chosen proxy method are met, such that $\mathcal{G}_{proxy} \equiv \mathcal{G}_{true}$. This theorem provides a formal guarantee that is conditional on the validity of the proxy, which is the strongest possible claim without access to ground-truth interventional data. □

C.3. Proof Sketch of Theorem 5.4

We block the sequence into nearly independent chunks of length b with gaps g so that $\beta(g)$ is small, then apply a symmetrization and Rademacher complexity argument adapted to mixing processes [35]. Lipschitzness of PACS (weighted softmax) gives a contraction. The W and H terms enter through the dependence of logits on history and forecast horizon.

C.4. Proof Sketch of Theorem 5.7

Consider the decomposition of variance across environments and apply stability bounds for Huber losses. When drift Δ_g is small, penalizing variance concentrates attention on parents up to $o(\epsilon)$ deviations; when large, adapters form a mixture mitigating bias at the cost of variance, bounded by $\tilde{\mathcal{O}}(\Delta_g)$.

D. Additional Experimental Results

D.1. Qualitative Comparison of Explanations

Figure 3 provides a side-by-side comparison of explanations generated for a stock price prediction. CEX-TSLM's explanation correctly identifies the causal news event, while the baseline model with SHAP highlights correlated but non-causal terms, or provides a less coherent rationale.

D.2. Attention Visualization

Figure 4 visualizes the Causal Cross-Attention weights for a sample prediction. The model correctly assigns the highest weights to the keywords in the text that describe the causal event identified by the proxy, demonstrating that the causal loss successfully guides the model's focus.

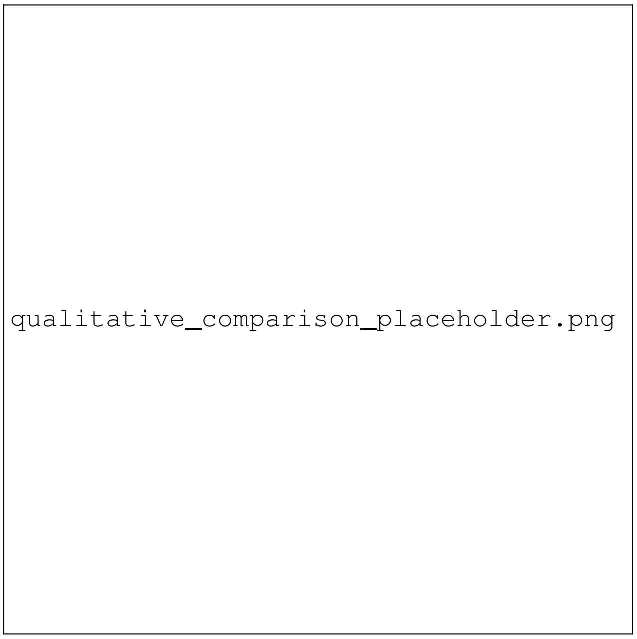


Figure 3. Qualitative comparison of explanations for a sample from the Time-MMD dataset. (Left) CEX-TSLM's generated rationale pinpoints the specific news event. (Right) SHAP explanation for a baseline model highlights several keywords, some of which are only correlational, leading to a less precise or potentially misleading explanation.

E. Pre-Registration Template (Proxies, Placebos, Environments)

Task and Outcome. Define forecast target, horizon H, window W.

Proxy definitions. Specify p_1 (Granger model, lags, tests), p_2 (event-study windows, Sun–Abraham corrections), any p_3 (ITE model class, ignorability diagnostics).

Calibration. Describe held-out placebo set, conformal score, desired miscoverage α, update schedule for $\boldsymbol{\omega}$.

Environments. Define splits (time regimes/seasons/hospitals), justify exogeneity.

Placebos/NCN. Construct future-shifted or mismatched

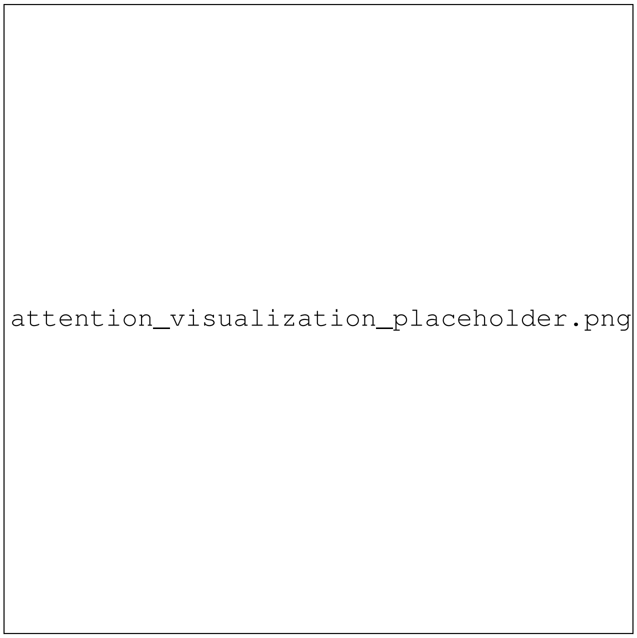


Figure 4. Visualization of Causal Cross-Attention weights. The time series query focuses its attention on the document containing the causal event (”drug approval”). Darker colors indicate higher attention scores, showing the model has learned to ignore spurious documents.

documents; report placebo pass rate target.

Edits (CRC). Specify operator class, semantic constraints, margin m.

Metrics. Pre-register swap tests, E-LOO, pathway ablations, and length-controlled faithfulness.

Seeds/Hyperparams. Fix seeds; grid for $(\alpha, \beta, \gamma, \lambda, \eta, \zeta)$ and temperatures.