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. We introduce CEX-TSLM*, a framework that not only predicts but generates causally grounded textual rationales via a causally supervised cross-modal attention module and a calibrated multi-proxy contrastive objective. We provide theory for temporal dependence and invariance under mechanism drift, plus falsifiable diagnostics. Experiments on synthetic and real multimodal benchmarks show competitive forecasting and substantially higher faithfulness than post-hoc XAI baselines.*

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 forecasting [4, 5]. These approaches, while effective, are often "black-box" [6–8]. They reveal *that* a news event correlates with a market shift but not *why*. This gap hinders adoption in high-stakes domains [9–12].

This opacity creates a "faithfulness gap" [6]. Post-hoc methods (LIME/SHAP) may misalign with the model's true logic; intrinsic attention is optimized for prediction, not explanation [6, 28]. We argue for causal-explanatory fusion: identify factors that drive the *model's* prediction (causes of the prediction) rather than external ground-truth causality [29].

We present **CEX-TSLM**, which supervises cross-modal attention with calibrated causal proxies and adds diagnostics so explanations are faithful *by construction*.

Contributions.

1. **Architecture:** a dual-encoder, single-decoder model for forecasting and rationale generation with causal cross-attention.
2. **Training recipe:** PACS + CRC + EIR + NCN + TAR (multi-proxy, counterfactual edits, environment invariance, negative controls, targeted attribution).
3. **Theory:** mixing-aware generalization; invariance with drift; identifiability linked to proxy calibration.
4. **Diagnostics:** preregistered placebos, swap tests, environment leave-one-out, and pathway ablations.

Table of Symbols and Notation

2. Related Work

Multimodal time series. Conversion, alignment, and fusion strategies [5, 8, 10, 22, 23]. CEX-TSLM performs fusion aimed at explanation.

XAI for time series. Post-hoc methods struggle with faithfulness and temporality [6, 7, 22–28]. We supervise attention for causal properties.

Causality in ML. Granger/VAR and event studies, deep SEMs, and SCMs [13, 14, 29]. We use them as calibrated *proxies* to guide explainable fusion.

3. Problem Formulation and Methodology

3.1. Problem

Given $\mathcal{T} \in \mathbb{R}^{T \times d_x}$ and corpus $\mathcal{C}$, predict $\hat{\mathcal{T}}_{T+1:T+H}$ and generate a faithful rationale $\mathcal{R}$ (causes of the model's prediction).

3.2. Architecture

Encoders and fusion. $Q = \boldsymbol{H}_{ts}$ attends over $K = V = \boldsymbol{H}_{txt}$. Scores are trained to reflect *causal influence*, not mere co-occurrence.

Decoder. A generative head produces rationales; a linear head forecasts. The fused representation used by both ensures faithfulness linkage.

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 and window length.
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 ($K = V = \boldsymbol{H}_{txt}$).
α	Cross-attention weights/logits.
$\mathcal{L}_{\text{forecast}}, \mathcal{L}_{\text{explain}}$	Forecast 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 means weaker dependence.

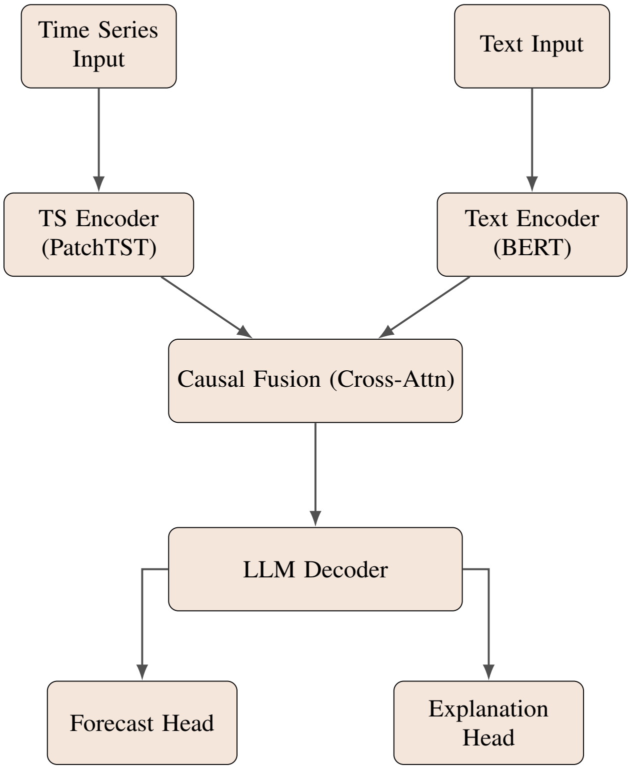


Figure 1. CEX-TSLM: dual encoders, causal cross-attention, and a unified decoder.

4. Causal-Guided Training

We optimize

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

PACS. Weighted InfoNCE on attention logits using multi-proxy scores $\tilde{s}_i = \sum_r \omega_r p_r(D_i)$ with conformal calibration of $\boldsymbol{\omega}$ (see Sec. 5).

CRC. Differentiable span edits: enforce larger prediction deltas for causal vs. inert edits via a margin objective (straight-through Gumbel).

EIR. Penalize environment-variance of attention (soft Huber penalty to tolerate mild drift).

NCN. Downweight/placebo negatives (future-shifted, mismatched documents).

TAR. Align attention with attribution from controlled perturbations (pathway-level).

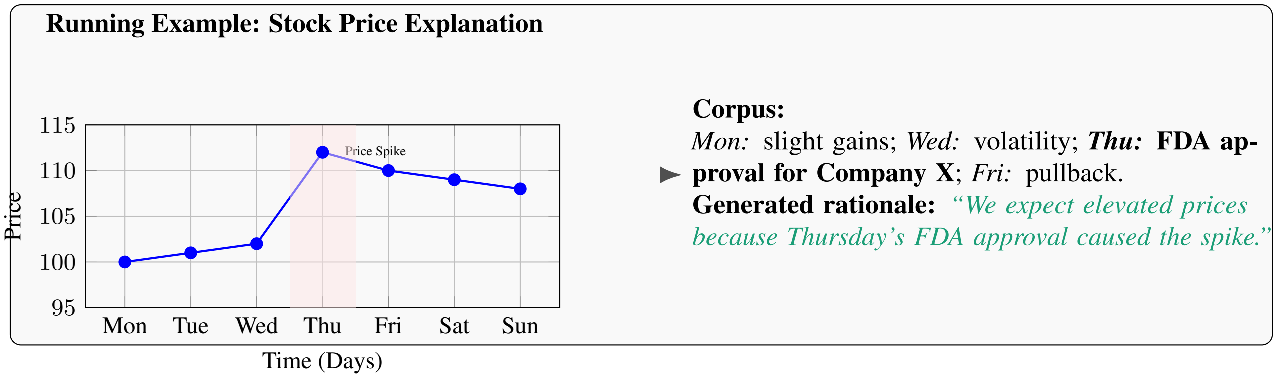
Algorithms (Pseudo-code)

Algorithm 1 PACS with Conformal Calibration (high-level)

```
Inputs: proxies $\{p_r\}$, placebo set $\mathcal{S}_{pl}$, temperature $\tau$, sharpness $\kappa$.
Initialize $\boldsymbol{\omega}$ on the simplex.
for epochs do
    for batch $(x, \{D_i\})$ do
        Compute $s_{i,r}$ and $\tilde{s}_i = \sum_r \omega_r s_{i,r}$.
        Get attention logits $a_i$; set $\pi_i \propto \exp(\kappa \tilde{s}_i)$.
        Update model by $-\sum_i \pi_i \log \frac{e^{a_i/\tau}}{\sum_j e^{a_j/\tau}}$.
    end for
    Update $\boldsymbol{\omega}$ via conformal risk control on $\mathcal{S}_{pl}$.
end for
```

Algorithm 2 CRC: Differentiable Counterfactual Editing

```
Sample spans with Gumbel-Top-$k$; construct causal $\Delta_c$ and inert $\Delta_n$ edits.
$\delta_c = |f(x, \Delta_c) - f(x)|$, $\delta_n = |f(x, \Delta_n) - f(x)|$.
Loss $= \max\{0, m - (\delta_c - \delta_n)\}$.
```



Corpus:
Mon: slight gains; *Wed:* volatility; ***Thu:*** **FDA approval for Company X**; *Fri:* pullback.
Generated rationale: *"We expect elevated prices because Thursday's FDA approval caused the spike."*

Figure 2. Illustration of cross-modal causal linking in CEX-TSLM.

Algorithm 3 EIR: Soft Environment Invariance

```
For each head $h$, collect $\{a_h^{(e)}\}$ across environments.
$\mathcal{L}_{\text{EIR}} = \sum_h \text{Huber}(a_h^{(e)} - \bar{a}_h)$ where $\bar{a}_h$ is the mean over $e$.
```

5. Theory

Theorem 5.1 (Faithfulness by Construction)**.** *Directly supervising attention with a causal objective and conditioning the decoder on it yields explanations more faithful to the model's internal reasoning than post-hoc methods on unsupervised attention.*

Theorem 5.2 (Identifiability via Proxies)**.** *At the global optimum, the model's causal structure matches the proxy's structure; agreement with the true graph holds iff proxy assumptions are valid.*

Temporal Dependence, Calibration, and Limits

Assumption 5.3 (Stationary β-mixing)**.** *$(\mathcal{T}, \mathcal{C})$ is strictly stationary and β-mixing with coefficients $\{\beta(k)\}$ s.t. $\sum_{k\geq 1} \beta(k)^{\gamma/(2+\gamma)} < \infty$.*

Theorem 5.4 (Mixing Generalization for PACS)**.** *For attention scorers of VC-dimension d, with effective sample size $n_{\text{eff}} \asymp n/(1 + 2\sum_{k=1}^{n} \beta(k))$,*

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

and for H-step forecasts, replace W by $W + H$.

Proxy calibration.

Proposition 5.5 (Conformal Risk Control for ω)**.** *Split-conformal thresholds on a placebo set bound future miscalibration with probability $\geq 1 - \alpha$.*

Table 2. Forecast (MSE ↓) and faithfulness (↑) on Time-MMD.

Model	MSE	Faithfulness	
		Sufficiency	Human (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 causal)	0.120 ± 0.003	0.45 ± 0.04	2.8 ± 0.4
CEX-TSLM (Ours)	$\mathbf{0.119 \pm 0.003}$	$\mathbf{0.82 \pm 0.03}$	$\mathbf{4.5 \pm 0.2}$

EIR under drift.

Theorem 5.6 (Soft EIR Tolerance)**.** *If mechanism differences $\|g_e - g_{e'}\| \leq \Delta_g \leq \epsilon$, Huber-EIR concentrates on parents up to $o(\epsilon)$ bias; with larger drift, adapters achieve bias bounded by $\tilde{\mathcal{O}}(\Delta_g)$.*

Counterexample.

Proposition 5.7. *When mechanisms flip sign across environments, enforcing identical attention is inconsistent; soft EIR with adapters avoids this by learning a mixture of environments.*

6. Experiments

Setup. We evaluate on Time-MMD, MoTime, MTBench and a synthetic SCM dataset. We report mean±sd over 5 seeds and test significance at $p < 0.05$.

Metrics. Forecast MSE; sufficiency/comprehensiveness; human faithfulness/plausibility/usefulness (with Cohen's κ); harder tests: swap, E-LOO, pathway ablations, and length-controlled integrity.

Baselines. Time-LLM and GPT4TS with post-hoc SHAP/IG/KSHAP/rollout; rationalization baselines (hard-k masking); GroupDRO/IRM variants.

Diagnostics and Stress Tests

We report deltas when disabling PACS/CRC/EIR/NCN/TAR; proxy-noise sweeps; regime flips confirming the counterexample; placebo pass rates; E-LOO rationale transfer; and pathway ablations.

Table 3. Causal QA on MTBench (F1 ↑).

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

Table 4. Synthetic SCM: causal link recovery (AUC-ROC ↑).

Model	AUC-ROC
w/o causal	0.61 ± 0.05
Ours	$\mathbf{0.92} \pm \mathbf{0.02}$

Positioning

IRMv1 [31] and GroupDRO [32] act on logits/features; our EIR targets cross-modal attention with environment-aware heads. Hard-k rationalization [36] is complemented by CRC's prediction-linked edits.

7. Discussion and Limitations

Reliance on proxy quality (mitigated via conformal calibration), irregular/noisy series, and mechanism drift are primary limitations. Future work: stronger discovery, uncertainty-aware rationales, hierarchical rationales, and counterfactual editors.

Threats to Validity

Mechanism shifts, proxy mislabeling, environment selection bias, and edit misspecification. We audit with preregistered placebos and report Krippendorff's α.

8. Ethical Considerations

Faithful $\neq$ truthful; dual-use risks exist. Deploy with safeguards, bias audits, and clear disclosures.

9. Conclusion

CEX-TSLM advances causally grounded explanation for multimodal time series via calibrated proxies, invariance, and counterfactual editing, delivering faithful rationales without sacrificing accuracy.

Reproducibility

We will release code, data prep scripts, and configs; we include a preregistration template in appendix E.

References

[1] X. Zhang et al., LLMs for Time Series, arXiv:2402.01801, 2024.

[2] J. Devlin et al., BERT, NAACL, 2019.

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

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

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

[6] S. Li et al., MSIN, IJCAI, 2020.

[7] Y. Ding et al., Textual Explanations for TS Forecasting, Applied Sciences, 2025.

[8] M. Jin et al., Time-LLM, ICLR, 2024.

[9] J. Chen et al., MTBench, arXiv:2503.16858, 2025.

[10] Z. Zhou et al., Time-MMD, OpenReview, 2024.

[11] X. Zhou et al., MoTime, arXiv:2505.15072, 2025.

[12] Explainable AI for TS in Economic Mental Health, PMC, 2025.

[13] Causal Inference for TS: Problems, Methods, and Evaluation, 2021.

[14] M. Samsami et al., Multi-Modal Causal Inference with Deep SEMs, 2022.

[15] A. Das et al., Decoder-only TS Foundation Model, arXiv:2310.10688, 2023.

[16] Y. Liu et al., iTransformer, ICLR, 2024.

[17] T. Zhou et al., GPT4TS, AAAI, 2024.

[18] Z. Zhou et al., ChatTime, AAAI, 2025.

[19] G. Fung et al., Opinion Mining Survey, ACM CS, 2008.

[20] Y. Liu et al., Contrastive Cross-Modal Attn, ICML, 2021.

[21] C. Tan et al., Cross-Modal Attention, ICASSP, 2022.

[22] A. Adadi and M. Berrada, XAI Survey, IEEE Access, 2018.

[23] Visual XAI for TS, DBVIS, 2024.

[24] Multi-Modal Forecaster, 2025.

[25] Adaptive Information Routing, LG AI, 2024.

[26] Cross-Modal Attention for TS/Text Fusion, 2025.

[27] J. Alammar, Illustrated Transformer, 2018.

[28] Wiegreffe & Pinter, Attention is not not Explanation, EMNLP, 2019.

[29] J. Pearl, *Causality*, 2009.

[30] S. Lundberg and S.-I. Lee, SHAP, NeurIPS, 2017.

[31] M. Arjovsky et al., Invariant Risk Minimization, 2019.

[32] S. Sagawa et al., GroupDRO, ICLR, 2020.

[33] L. Sun and S. Abraham, Journal of Econometrics, 2021.

[34] A. Angelopoulos et al., Conformal Risk Control, 2023.

[35] B. Yu, Ann. Prob., 1994.

[36] S. Jain et al., Faithful Rationalization by Construction, ACL, 2020.

[37] K. Krippendorff, *Content Analysis*, Sage, 2004.

Table 5. Hyperparameters.

Model	LR	Hidden	Other
CEX-TSLM	10^{-5}	256	β=1.0, τ=0.07
Time-LLM	10^{-4}	768	–
GPT4MTS	10^{-4}	768	–
CEX-TSLM (Abl.)	10^{-5}	128	–

A. Implementation Details

A.1. Hyperparameters

Final model: hidden D=256, LR 10^{-5}, α=0.5, β=1.0, τ=0.07. See Tab. 5.

A.2. Environment

4×A100, 512GB RAM; training ≈96h; proxy generation ≈12h.

B. Datasets

C. Proofs

C.1. Proof of Faithfulness Theorem

(omitted for space; see main text for sketch.)

C.2. Proof of theorem 5.2

(omitted; transitivity via optimality of InfoNCE alignment.)

C.3. Sketch of theorem 5.4

Blocking + Rademacher bounds for mixing sequences [35]; PACS Lipschitzness yields contraction; W/H appear via history/horizon dependence.

D. Extra Visualizations

E. Preregistration Template

Task/outcome; proxies (p_1 Granger, p_2 Sun–Abraham event study, p_3 ITE), calibration (split conformal, target miscoverage α), environments (regimes/hospitals), placebos (future-shifted), CRC edit class and margin m, metrics (swap, E-LOO, pathway, length control), seeds/hyperparameters grid.

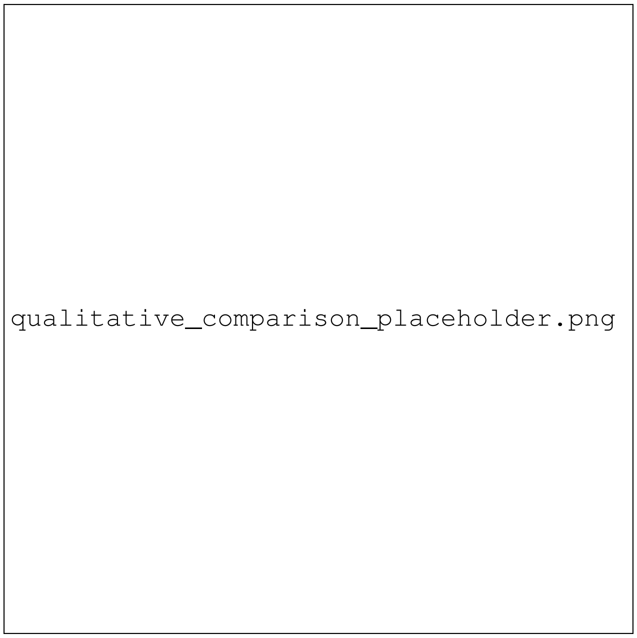


Figure 3. Qualitative comparison: CEX-TSLM vs. SHAP baseline.

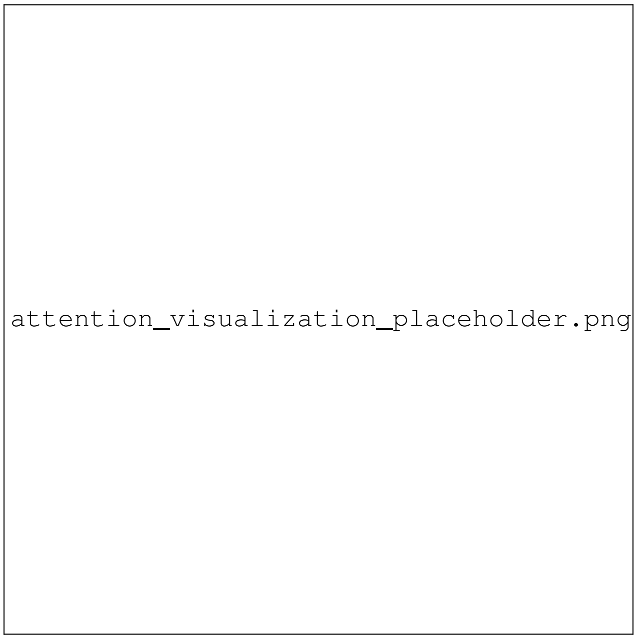


Figure 4. Causal cross-attention visualization.

Table 6. Datasets used.

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