Causal-LLM: Towards Predictive and Interpretable Spatiotemporal Foundation Models

Zhiqing Cui[1, 2]

[1]Nanjing University of Information Science and Technology
[2]University of Reading
dh803755@student.reading.ac.uk

Abstract

Spatiotemporal Graph Neural Networks (STGNNs) have revolutionized forecasting accuracy but suffer from inherent opacity, failing to explain the underlying drivers of their predictions. This lack of interpretability creates a critical trust bottleneck in high-stakes domains such as meteorology and urban planning. To bridge this gap, I propose Causal-LLM, a novel neuro-symbolic foundation model designed to be both predictively powerful and causally interpretable. Furthermore, I introduce a causal data synthesis training paradigm that explicitly teaches the model to align numerical forecasts with human-understandable causal narratives. This work provides a blueprint for trustworthy AI in science, enabling models that not only predict future states but also articulate the physical *why* behind them.

1 Introduction

Spatiotemporal forecasting has seen remarkable progress with the advent of deep learning, particularly with Spatiotemporal Graph Neural Networks (STGNNs). These models excel at answering the *what* question: predicting future numerical values with high accuracy. However, they fail to answer the crucial *why* question. In high-stakes domains such as meteorology, urban planning, and public health, this opacity creates a critical bottleneck for adoption. A model that predicts a severe pollution event without explaining its atmospheric drivers is a black box, limiting its trustworthiness and utility for decision-makers who need actionable, causal insights.

To address this critical gap, I propose a long-term research project to develop **Causal-LLM**, a new class of foundation models for spatiotemporal data that are both predictively powerful and causally interpretable. My central thesis is that genuine interpretability cannot be an afterthought; it must be designed into the model's core learning process. Through a process I term causal data synthesis, Causal-LLM will learn to forecast future states while articulating the human-understandable causal narratives behind them.

This research will make two primary contributions: (1) a novel hybrid architecture that synergizes the perceptual power of GNNs with the reasoning capabilities of LLMs for complex physical systems, and (2) a new training paradigm that explicitly teaches this mapping. A successful project would provide a blueprint for a new class of trustworthy foundation models for science, enabling applications such as a climate model that not only predicts a flood but also explains the atmospheric river causing it, empowering authorities to make more informed and trusted decisions.

2 Background

2.1 Related Work

The integration of Large Language Models (LLMs) with numerical time series data represents a vibrant and rapidly evolving research frontier(Qiu et al. 2024, 2025c,a,b). Motivated by the intuition that the vast, pre-trained knowledge and reasoning capabilities of LLMs can enhance predictive signals beyond what is captured in the time series alone, recent work has explored various methods for this fusion(Nie et al. 2022). These approaches can be broadly categorized into two dominant paradigms: aligning-based methods and prompting-based methods(Jin et al. 2024).

Aligning-based methods seek to fuse multimodal information by creating a shared latent space, for instance by projecting time series embeddings into the LLM's representation space (Pan et al. 2024; Liu et al. 2025; Jia et al. 2024). However, this paradigm has two fundamental weaknesses for my research focus. First, the fused latent space remains a "black box" optimized for numerical accuracy, not for generating human-understandable causal explanations (Jin et al. 2023). Second, these methods are not natively equipped to handle the rich, graph-structured topology inherent in spatiotemporal data (Tian et al. 2024).

In contrast, **prompting-based methods** leverage the in-context learning capabilities of LLMs like Deepseek-R1 or Gemini 2.5-pro by formatting numerical data directly into natural language prompts for zero-shot forecasting (Xue and Salim 2023; Gruver et al. 2023). While simple and powerful, this approach suffers from a critical lack of physical grounding, as LLMs have no intrinsic understanding of the physical laws that govern scientific spatiotemporal data.

2.2 Prior Work by the Applicant

My proposed research is a direct extension of my prior work (Cui et al. 2025a,c; Wang et al. 2025), which has equipped me with a strong foundation in two key areas: modeling

complex spatiotemporal systems and bridging perception with symbolic reasoning(Yuan et al. 2025).

My expertise in spatiotemporal forecasting is demonstrated by two first-author or co-first-author publications. In my work on CauAir (Ma et al. 2025), I developed a novel causal dataset for nationwide air quality prediction. Similarly, my first-author work on Prithvi-TC (Cui, Meng, and Luo 2025) involved building a foundation model for tropical cyclone prediction. While these models achieved state-of-the-art predictive accuracy, they highlighted a critical limitation: their inability to provide human-understandable explanations for their forecasts, which directly motivates my current proposal.

Concurrently, my experience in multimodal reasoning comes from leading the Draw with Thought (DWT) project (Cui et al. 2025b). I designed a framework to translate unstructured diagram images into structured, symbolic XML code. Together, these experiences have given me both the domain expertise and the technical skills required to successfully undertake this project.

3 Approach

My proposed **Causal-LLM** will adapt the powerful Time-LLM architecture to create a model capable of both high-fidelity spatiotemporal forecasting and causal explanation.

3.1 Architecture: A Three-Stage Pipeline

The model's architecture is a three-stage pipeline designed for a clear separation of tasks:

1. **Spatiotemporal Encoder:** A Graph Neural Network (GNN) is used to perceive the physical state, as it is natively designed to handle the graph-structured topology of sensor networks. It processes the raw, multivariate data to learn a condensed latent vector representing the entire system's state.
2. **Reprogramming Module:** This acts as a neuro-symbolic bridge, projecting the GNN's continuous latent vectors into the LLM's word embedding space. This creates a sequence of special "physical state tokens," where each token represents a learned, recurring physical phenomenon. This allows the frozen, pre-trained LLM to process these complex dynamics using its powerful reasoning capabilities without altering its weights.
3. **Multi-Task Decoders:** The LLM's final output is directed to two heads to generate both a numerical forecast via a simple linear head and a natural language causal explanation via its own language head.

3.2 Training via Causal Data Synthesis

The central innovation lies in how this architecture is trained. Instead of relying on the model to implicitly discover correlations from raw data—a process that can be unreliable—I will guide it explicitly through **causal data synthesis**. This involves curating a novel training dataset where each instance is a carefully constructed triplet. For each significant historical event, this triplet will consist of the raw data leading up to the event as the input, the ground-truth numerical values of the event as the predictive target, and a ground-truth, human-written causal explanation for that event as the explanatory target. These explanations will be sourced from scientific literature, such as, "A high-pressure system created a temperature inversion, trapping pollutants."

By training the model end-to-end with a combined multi-task loss on these triplets, the Causal-LLM learns a direct and powerful mapping. It is explicitly taught that a specific physical state perceived by the GNN corresponds not only to a numerical outcome but also to a specific causal narrative. This methodology forms the foundation for a system that is interpretable by design, not by chance.

4 Evaluation

The success of this project will be measured across two dimensions: predictive power and explanatory quality.

Predictive accuracy will be rigorously validated using standard metrics (MAE, RMSE) on a diverse range of benchmarks. This includes domain-specific datasets (Wang et al. 2020), as well as general forecasting datasets like **ETT**, to test for generalizability. Performance will be compared against a comprehensive suite of state-of-the-art baselines, and extensive **ablation studies** will be conducted to quantify the contribution of each architectural component.

Explanatory quality will be assessed using a mixed-methods approach. I will use quantitative NLP metrics, such as ROUGE and BERTScore (Huang et al. 2025), to compare the model's generated text against the ground-truth explanations from my curated dataset. Critically, this will be supplemented by **qualitative human evaluation**, where domain experts (e.g., meteorologists) will rate the factual correctness, coherence, and scientific utility of the model's explanations for significant historical events. The overall goal is to demonstrate that the Causal-LLM is not only predictively competitive but also provides trustworthy and useful scientific insights.

5 Discussion

The primary expected outcome is an AI system that generates factually correct and scientifically useful causal explanations for its predictions, potentially with a slight trade-off in raw numerical accuracy for a significant gain in trustworthiness. This work would represent a major step forward in neuro-symbolic AI, providing a concrete methodology for creating foundation models for complex, dynamic domains. For society, it paves the way for more reliable and transparent AI decision-support systems in critical areas like emergency management and environmental policy.

6 Conclusion

To address the opacity of current spatiotemporal models, I propose Causal-LLM, a novel architecture that synergizes the perceptual power of GNNs with the reasoning capabilities of LLMs. Through a principled adaptation of the Time-LLM framework and a novel causal data synthesis training method, this work aims to pioneer a new class of models that are both predictively powerful and causally interpretable.

References

Cui, Z.; Meng, F.; and Luo, J. 2025. Breaking through tropical cyclone intensity prediction: a foundation model Prithvi-TC. *Frontiers of Computer Science*, 19(12): 1–3.

Cui, Z.; Wang, B.; Liu, Q.; Wang, Y.; Zhou, Z.; Liang, Y.; and Wang, Y. 2025a. Augur: Modeling Covariate Causal Associations in Time Series via Large Language Models. *arXiv preprint arXiv:2510.07858*.

Cui, Z.; Yuan, J.; Wang, H.; Li, Y.; Du, C.; and Ding, Z. 2025b. Draw with Thought: Unleashing Multimodal Reasoning for Scientific Diagram Generation. *arXiv preprint arXiv:2504.09479*.

Cui, Z.; Yuan, J.; Xu, H.; Wei, Y.; and Ding, Z. 2025c. RTL-Net: real-time lightweight Urban traffic object detection algorithm. *Complex & Intelligent Systems*, 11(7): 304.

Gruver, N.; Finzi, M.; Qiu, S.; and Wilson, A. G. 2023. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36.

Huang, Y.; Fang, Z.; Zeng, Z.; Chen, L.; and Gao, Y. 2025. Causal Spatio-Temporal Prediction: An Effective and Efficient Multi-Modal Approach. *arXiv preprint arXiv:2505.17637*.

Jia, F.; Wang, K.; Zheng, Y.; Cao, D.; and Liu, Y. 2024. Gpt4mts: Prompt-based large language model for multimodal time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 23343–23351.

Jin, M.; Wang, S.; Ma, L.; Chu, Z.; Zhang, J. Y.; Shi, X.; Chen, P.-Y.; Liang, Y.; Li, Y.-F.; Pan, S.; et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*.

Jin, M.; Zhang, Y.; Chen, W.; Zhang, K.; Liang, Y.; Yang, B.; Wang, J.; Pan, S.; and Wen, Q. 2024. Position paper: What can large language models tell us about time series analysis. *CoRR*.

Liu, C.; Xu, Q.; Miao, H.; Yang, S.; Zhang, L.; Long, C.; Li, Z.; and Zhao, R. 2025. Timecma: Towards llm-empowered multivariate time series forecasting via cross-modality alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 18780–18788.

Ma, J.; Cui, Z.; Wang, B.; Wang, P.; Zhou, Z.; Zhao, Z.; and Wang, Y. 2025. Causal learning meet covariates: Empowering lightweight and effective nationwide air quality forecasting. In *International Joint Conference on Artificial Intelligence*.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2022. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*.

Pan, Z.; Jiang, Y.; Garg, S.; Schneider, A.; Nevmyvaka, Y.; and Song, D. 2024. S^2IP-LLM: Semantic space informed prompt learning with LLM for time series forecasting. In *Forty-first International Conference on Machine Learning*.

Qiu, X.; Hu, J.; Zhou, L.; Wu, X.; Du, J.; Zhang, B.; Guo, C.; Zhou, A.; Jensen, C. S.; Sheng, Z.; and Yang, B. 2024. TFB: Towards Comprehensive and Fair Benchmarking of Time Series Forecasting Methods. In *Proc. VLDB Endow.*, 2363–2377.

Qiu, X.; Li, Z.; Qiu, W.; Hu, S.; Zhou, L.; Wu, X.; Li, Z.; Guo, C.; Zhou, A.; Sheng, Z.; Hu, J.; Jensen, C. S.; and Yang, B. 2025a. TAB: Unified Benchmarking of Time Series Anomaly Detection Methods. In *Proc. VLDB Endow.*, 2775–2789.

Qiu, X.; Wu, X.; Cheng, H.; Liu, X.; Guo, C.; Hu, J.; and Yang, B. 2025b. DBLoss: Decomposition-based Loss Function for Time Series Forecasting. In *NeurIPS*.

Qiu, X.; Wu, X.; Lin, Y.; Guo, C.; Hu, J.; and Yang, B. 2025c. DUET: Dual Clustering Enhanced Multivariate Time Series Forecasting. In *SIGKDD*, 1185–1196.

Tian, J.; Liang, Y.; Xu, R.; Chen, P.; Guo, C.; Zhou, A.; Pan, L.; Rao, Z.; and Yang, B. 2024. Air quality prediction with physics-guided dual neural odes in open systems. *arXiv preprint arXiv:2410.19892*.

Wang, H.; Wang, S.; Zhong, Y.; Yang, Z.; Wang, J.; Cui, Z.; Yuan, J.; Han, Y.; Liu, M.; and Ma, Y. 2025. Affordance-r1: Reinforcement learning for generalizable affordance reasoning in multimodal large language model. *arXiv preprint arXiv:2508.06206*.

Wang, S.; Li, Y.; Zhang, J.; Meng, Q.; Meng, L.; and Gao, F. 2020. Pm2. 5-gnn: A domain knowledge enhanced graph neural network for pm2. 5 forecasting. In *Proceedings of the 28th international conference on advances in geographic information systems*, 163–166.

Xue, H.; and Salim, F. D. 2023. PromptCast: A New Prompt-based Learning Paradigm for Time Series Forecasting. In *The 32nd ACM International Conference on Information and Knowledge Management*.

Yuan, J.; Di, Z.; Cui, Z.; Yang, G.; and Naseem, U. 2025. ReflectDiffu: Reflect between Emotion-intent Contagion and Mimicry for Empathetic Response Generation via a RL-Diffusion Framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 25435–25449.