Transformer Is Inherently a Causal Learner

Xinyue Wang
Halıcıoğlu Data Science Institute
University of California San Diego
La Jolla, CA 92093
xiw159@ucsd.edu

Stephen Wang
ABEL Intelligence, Inc.
Mountain View, CA 94040
stephen@abel.ai

Biwei Huang
Halıcıoğlu Data Science Institute
University of California San Diego
La Jolla, CA 92093
bih007@ucsd.edu

Abstract

We reveal that transformers trained in an autoregressive manner naturally encode time-delayed causal structures in their learned representations. When predicting future values in multivariate time series, the gradient sensitivities of transformer outputs with respect to past inputs directly recover the underlying causal graph, without any explicit causal objectives or structural constraints. We prove this connection theoretically under standard identifiability conditions and develop a practical extraction method using aggregated gradient attributions. On challenging cases such as nonlinear dynamics, long-term dependencies, and non-stationary systems, we see this approach greatly surpass the performance of state-of-the-art discovery algorithms, especially as data heterogeneity increases, exhibiting scaling potential where structure discovery accuracy improves with data volume, a property traditional methods lack. This unifying view opens a new paradigm where causal discovery operates through the lens of foundation models, and foundation models gain interpretability and enhancement through the lens of causality. [1]

1 Introduction

Causality drives scientific progress across domains, e.g., medicine [Doll and Hill, 1950, Popa-Fotea, 2021], economics [Chetty et al., 2015], and neuroscience [Roth, 2016]. As an evolving field, causal discovery aims to formalize theoretical frameworks for identification criteria and propose search algorithms to find the true causal structure from observational data [Pearl, 2009, Spirtes et al., 2000]. In this area, causal discovery from time series focuses on identifying temporal causal dynamics by exploiting the temporal ordering that naturally constrains the direction of causation. Granger causality [Granger, 1969, Tank et al., 2021, Nauta et al., 2019] formalizes this intuition: a variable X Granger-causes Y if past values of X contain information that helps predict Y beyond what is available from past values of Y alone. Additional methods extend this foundation, including constraint-based approaches like PCMCI and its variants that iteratively test conditional independence to examine the existence of causal edges [Runge et al., 2017], score-based methods like DYNOTEARS [Pamfil et al., 2020] that optimize graph likelihood with structural prior regularizations, and functional approaches like TiMINo and VAR-LiNGAM that leverage structural equation models and non-Gaussianity for identifiability [Peters et al., 2014, Hyvärinen et al., 2010].

[1] Project website: https://www.charonwangg.com/project/transformers-scale-discovery

Accepted to the NeurIPS 2025 Workshop on CauScien: Uncovering Causality in Science.

Real-world systems exhibit complex interactions among many variables. For example, financial markets are highly non-stationary and involve very large variable sets [Engle, 1982]; neural recordings exhibit strongly nonlinear population dynamics [Breakspear, 2017]; climate sensor networks display long and short-term teleconnections [Wallace and Gutzler, 1981, Newman et al., 2016]; and unstructured modalities such as video require modeling long-range spatiotemporal dependencies [Bertasius et al., 2021, Arnab et al., 2021]. Despite rigorous theoretical foundations, prevailing algorithms are often constrained in practice by complex heuristics. Specifically, constraint-based and score-based approaches scale poorly: the number of statistical tests grows rapidly with dimension and lag, and non-parametric tests are computationally expensive [Runge et al., 2017, Chickering, 2002]. Optimization approaches require careful tuning to achieve the right balance between likelihood and structural regularization [Zheng et al., 2018, Ng et al., 2020a, Pamfil et al., 2020, Zheng et al., 2019]. More fundamentally, these estimators are not scalable representation learners: their learning is not transferable and thus offers little generalizability for zero- or few-shot adaptation; their effective capacity and expressiveness are not well-suited for pretraining on diverse systems.

Motivated by the striking performance and scaling behavior of autoregressive foundation models [Brown et al., 2020, Kaplan et al., 2020, Hoffmann et al., 2022], we ask whether the properties that make transformers strong forecasters can help causal discovery. Building this connection is valuable in two directions: for discovery, it promises data efficiency by leveraging pretrained representations and a scalable learning paradigm suited to complex dependencies; for foundation models, causal principles offer diagnose limitations in memory and hallucinations, and guide architecture and objective choices. In this paper, we take a first step toward these goals: we revisit common identifiability assumptions in lagged data generation processes and show how decoder-only transformers trained for forecasting, together with input–output gradient attributions via Layer-wise Relevance Propagation (LRP) [Achtibat et al., 2024, Bach et al., 2015], reveal lagged causal structure. This view turns modern sequence models into practical, scalable estimators for temporal graphs while opening a path to analyze and strengthen foundation models through causal perspectives.

2 A Unifying View: Identification inside Robust Next Variables Prediction

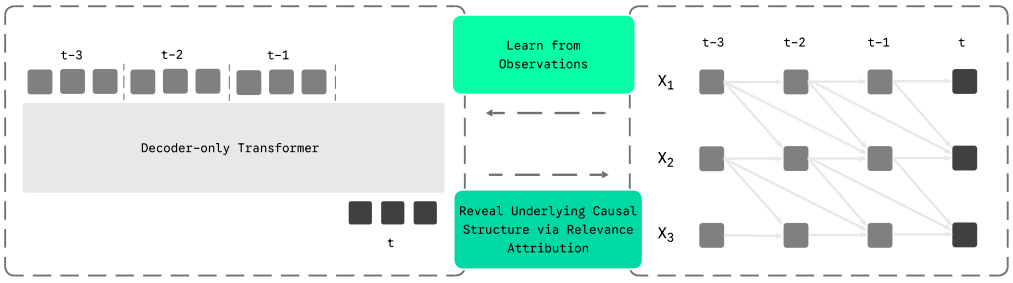


Figure 1: **Data generation and transformer-based causal discovery. Left:** A decoder-only transformer trained for next-step prediction. Tokens are lagged observations from $t-L$ to $t-1$; the model predicts X_t from $X_{t-1:t-L}$. **Right:** A lagged data-generating process with $p=3$ and window $L=3$. Each $X_{i,t}$ depends on selected past values $X_{j,t-\ell}$ per the true graph $\mathcal{G}^*$. The trained transformer learns the process, and relevance attribution help recover the causal structure.

2.1 From Prediction to Causation

Data-generating process. Consider a p-variate time series $X_t = (X_{1,t}, \ldots, X_{p,t})^\top$ and a lag window $L \geq 1$. Each variable follows

$$X_{i,t} = f_i(\mathrm{Pa}(i,t), N_{i,t}),$$

where $\mathrm{Pa}(i,t) \subseteq \{X_{j,t-\ell} : j \in [p],\ \ell \in [L]\}$ are the lagged parents and $N_{i,t}$ are independent noises. We write $j \xrightarrow{\ell} i$ if $X_{j,t-\ell}$ is a direct cause of $X_{i,t}$. The lagged graph $\mathcal{G}^*$ contains $j \xrightarrow{\ell} i$ iff $X_{j,t-\ell} \in \mathrm{Pa}(i,t)$.

Assumptions for lagged identifiability

A1 Causal sufficiency (no latent confounders).

A2 No instantaneous effects (all parents occur at lags $\ell \geq 1$).

A3 Lag-window coverage (the chosen L includes all true parents).

A4 Causal Markov and Faithfulness [Spirtes et al., 2000, Pearl, 2009].

This theorem reduces causal discovery to finding which lagged variables are predictively relevant for each target. The identifiability criterion most closely related to ours is Granger causality, where it is termed as *predictive causation*. Analytically, this can be captured by the population gradient energy $G^{\ell}_{j,i} := \mathbb{E}[(\partial_{x_{j,t-\ell}} f^*(X_i))^2]$, which is zero exactly for non-parents and positive for parents. In practice, we approximate $G^{\ell}_{j,i}$ by aggregated Layer-wise Relevance (LRP) $\tilde{G}^{(\ell)}_{j,i} := \mathbb{E}[\,|R^{(\ell)}_{ij}(X)|\,]$, then calibrate to recover $\mathcal{G}^*$. As we show next, decoder-only transformers are well aligned with these properties and suitably serve as scalable causal learners. When assumptions are violated (e.g., latent confounding, instantaneous effects), we can handle them by adjusting masking rules and combining traditional causal discovery methods as post-processing procedures. We illustrate straightforward ways to handle latent confounders and instantaneous relationships in Section 3 and Subsection A.7.8. Identifiability proof and LRP–gradient connection are provided in Appendix §A.1 and Appendix §A.2.

Causal Identifiability via Prediction

Theorem 1. *Under A1–A4, the lagged causal graph $\mathcal{G}^*$ is uniquely identifiable from conditional prediction dependencies: edge $j \xrightarrow{\ell} i$ exists iff $X_{j,t-\ell}$ is informative for helping to predict $X_{i,t}$ given all other lagged variables.*

2.2 Transformers inherit causal identifiability

We connect Theorem 1 to decoder-only transformers and make explicit why this architecture aligns with the identifiability program in Section 2.1, and how we extract a graph in practice. The connection has four parts: (i) alignment with assumptions A1–A4 and the forecasting objective, (ii) scalable sparsity and conditional-dependence selection, (iii) contextualized parameters for heterogeneity, and (iv) a structure extraction and binarization procedure.

Alignment with identifiability and objective. We use a decoder-only transformer on a length-L window. For each $t > L$, the input $\mathbf{s}_t = [X_{t-L}, \ldots, X_{t-1}] \in \mathbb{R}^{L \times p}$ is flattened to $L \cdot p$ tokens. We use separate learnable node embedding and time embedding to distinguish temporal dimension and node entities. Causal masking and autoregressive decoding enforce temporal precedence (A2); the window L bounds the maximum lag (A3). We assume there are no hidden confounders (A1). Unlike traditional structure learning approaches, which use a fixed input length to predict only the last token, our autoregressive training uses every token as a training signal via teacher forcing. We optimize:

$$\min_{\theta} \; -\frac{1}{(T-L)\,L} \sum_{i=1}^{T-L} \sum_{k=1}^{L} \log p_\theta(X_{i+k} \,|\, X_{i:i+k-1}) \;+\; \lambda\, \Omega(\theta). \tag{1}$$

where $p_\theta(\cdot \mid \cdot)$ denotes the conditional likelihood parameterized by transformer outputs $\widehat{f}_\theta : \mathbb{R}^{L \times p} \to \mathbb{R}^p$. For simplicity, we use a Gaussian likelihood (Mean Square Error objective), and $\Omega(\theta)$ is optional (e.g., sparsity or entropy regularization; by default, we do not use structural penalties).

Selectivity and scalable dependence selection. While explicit sparsity is not required for identifiability in the population, finite-sample recovery benefits from sparsity for both accuracy and efficiency. Constraint-based and score-based approaches control complexity via combinatorial conditioning and structural penalties, which limits scalability in high dimensions and long lags. Transformers exhibit implicit sparsification: finite capacity, weight decay, and the implicit bias of gradient descent favor low-complexity solutions; softmax attention induces competitive selection among candidates [Martins and Astudillo, 2016, Sutton et al., 1998]; and multi-head context supports selecting complementary parents. These priors make transformers well suited for scalable causal learning and can be complemented with explicit sparsity if desired.

Attention as contextual parameters. Attention matrices are input-conditioned and therefore act as contextualized parameters of pairwise dependencies rather than fixed population-level graph weights commonly used in optimization-based estimators [Zheng et al., 2018, Pamfil et al., 2020]. Unlike methods that learn a single static binary mask, input-conditioned attention adapts to heterogeneity and non-stationarity: different contexts (time, regime) induce distinct effective dependency patterns. This flexibility is desirable and scalable in practice, enabling a data-driven mixture-of-graphs view without committing to a single mask.

Structure extraction. After training, we recover structure via population gradient energy rather than raw attention. We use Layer-wise Relevance Propagation (LRP) [Achtibat et al., 2024] to compute relevance scores $R_{ij}^{(\ell)}$ that quantify the influence of variable j at lag ℓ on predicting variable i at time t:

$$R_{ij}^{(\ell)} = \sum_{m=1}^{M} \sum_{h=1}^{H} \mathrm{LRP}^{(m,h)}\big(\widehat{f}_\theta, X_t^{(i)}, X_{t-\ell}^{(j)}\big). \tag{2}$$

We aggregate these attributions across samples to estimate gradient energy $\tilde{G}_{j,i}^{(\ell)} = \mathbb{E}[\,|R_{ij}^{(\ell)}(X)|\,]$ and then calibrate to a sparse graph. We do not use raw attention weights as causal explanations since deep token mixing often misaligns attention scores with input and output dependence [Jain and Wallace, 2019]. See Appendix §A.2 for implementation and aggregation details.

Graph binarization. We propose two rules to binarize it: (i) *Top-k per target:* for each target variable (row), select the k largest entries as parents; this directly controls graph density and stabilizes precision. (ii) *Uniform-threshold rule:* assume a uniform baseline over $L \times p$ candidates and select entries whose normalized relevance exceeds $\frac{1}{L \times p}$. The two rules behave similarly at small scale; as context length grows, the uniform-threshold rule tends to degrade in precision compared to Top-k.

3 Experiments

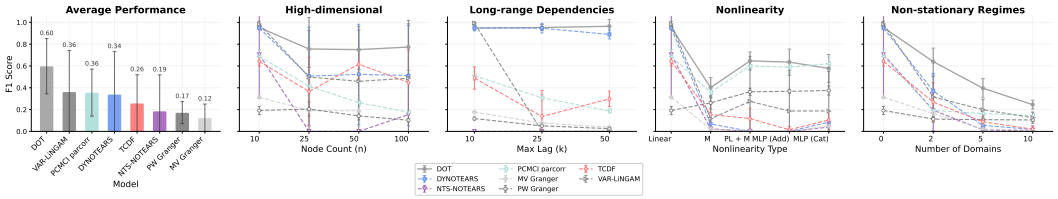


Figure 2: **F1 score analysis across regimes. (A)** Mean F1 across all experiments (averages exclude timeout cases). **(B)** High-dimensional input: F1 averaged across scales and seeds vs. the number of nodes. **(C)** Long-range dependencies: F1 averaged across scales and seeds vs. maximum lag. **(D)** Nonlinearity: F1 averaged across scales and seeds vs. different types of functional forms. **(E)** Non-stationarity: F1 averaged across scales and seeds vs. the number of domains. We run each method with three seeds. Missing results indicate method timeouts due to computational limits. DOT stands for Decoder-only Transformer. PL and M stand for piecewise linear and monotonic functions.

Setup. We evaluate decoder-only transformers for causal discovery using the simulator detailed in Appendix §A.3. We compare against PCMCI [Runge et al., 2017], DYNOTEARS [Pamfil et al., 2020], VAR-LiNGAM [Hyvärinen et al., 2010, Peters et al., 2014], NTS-NOTEARS [Sun et al., 2021], TCDF [Nauta et al., 2019] and pairwise/multivariate Granger tests [Granger, 1969] across variations in nonlinearity, maximum lag, dimensionality, noise type, and non-stationarity (see more discussions in Appendix §A.6.

General capability and complex dependencies. The transformer recovers lagged parents accurately and consistently across settings, achieving comparable or better performance to baseline methods (Figure 2). It maintains strong performance under nonlinearity, long-term dependencies, large variable sizes, and non-stationarity. Traditional methods degrade as dynamics and dimension grow, whereas the transformer remains robust without sensitive hyperparameter tuning. Its advantages stem from the model's expressivity and attention-based dependency selection. Performance improves steadily with sample size, making the approach suitable for complex real-world scenarios. More detailed results including additional settings, case studies, and analysis of transformer variants, are provided in Appendix §A.7.

Capacity and scaling potential. The transformer effectively leverages additional data to improve causal structure modeling accuracy (see Figure 6 and 7). Unlike traditional methods that are intractable with more data, the transformer shows consistent improvement across sample sizes. In non-stationary settings, the model learns to handle multiple local mechanisms within a single framework. As sample size increases, the transformer better separates and routes different causal structures corresponding to distinct regimes. This scaling behavior mirrors that of large pretrained models and distinguishes our approach from traditional causal discovery methods. The results also suggest that hallucinations in foundation models may arise when insufficient data prevents accurate regime separation and structure routing.

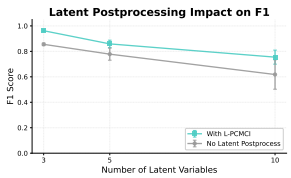


Figure 3: **Handling latent confounders with post-processing.** Comparison of F1 scores between using L-PCMCI as post-processing and using transformer alone.

The potential of handling latent confounders. Transformer performance degrades under latent confounding, and its architecture cannot generally model latent variables (see Figure 9). We show that it is possible to handle this by post-processing with a latent-aware causal discovery method: run L-PCMCI [Gerhardus and Runge, 2020] constrained by the transformer's predicted edges to refine the graph. Starting from the transformer's graph sharply reduces the expensive search space of latent-aware causal discovery methods. The combined pipeline is robust to latent confounders and yields substantially higher accuracy than the transformer alone (see Figure 3).

Integration with known domain indicators in non-stationary settings. Exploiting variation across environments and distributions helps identify causal structure and representations [Huang et al., 2020, Khemakhem et al., 2020]. Providing domain indicators lets the model separate cross-domain changes from invariants. We encode a domain index, proxying distribution shifts, as an additional input to a decoder-only transformer, improving data efficiency in both standard and highly complex settings (Figure 4) and helping disentangle structure within representations.

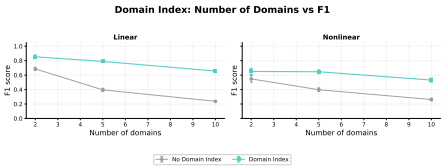
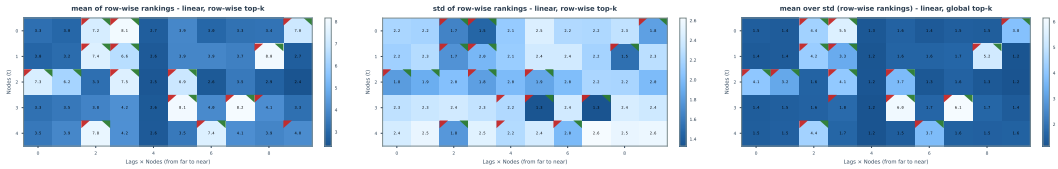


Figure 4: **Integrating with domain indicators helps structure learning in non-stationary settings.** Integrating domain index as a surrogate consistently improves data efficiency in both linear and nonlinear settings.

Uncertainty analysis. Statistical causal discovery outputs a population-level graph and estimates uncertainty via resampling (e.g., bootstrap). With transformers, we can aggregate per-sample point estimates and use their standard deviation to gauge consistency. Because larger mean relevance scores often have larger raw score variance, we rank each target's candidate parents within every sample and summarize these ranks by their mean and standard deviation. True edges show higher mean rank and lower rank standard deviation, indicating greater confidence (Figure 5). This offers a pragmatic way to surface the most reliable edges when precision is prioritized. In graphs with varied degrees, combining the mean and variance of ranks with a global top-k yields more accurate structures than using the mean of raw scores with row-wise top-k in both linear and nonlinear settings. More results are provided in Appendix §A.7.9.



(a) (b) (c)

Figure 5: **Uncertainty analysis of causal structure estimation.** Mean and variance of relevance score rankings across samples for potential parents of target variables. Larger mean rankings tend to have a lower variance in rankings, indicating the model's confidence in identifying true causal relationships. The top-left red triangle means that model predicts there is a causal edge and top-right green triangle means that there is a true edge between the two variables.

References

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Attnlrp: attention-aware layer-wise relevance propagation for transformers. *arXiv preprint arXiv:2402.05602*, 2024.

Robert A Adams and John JF Fournier. *Sobolev spaces*, volume 140. Elsevier, 2003.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Icml*, volume 2, page 4, 2021.

Michael Breakspear. Dynamic models of large-scale brain activity. *Nature neuroscience*, 20(3): 340–352, 2017.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Yu Chen, Nathalia Céspedes, and Payam Barnaghi. A closer look at transformers for time series forecasting: Understanding why they work and where they struggle. In *Forty-second International Conference on Machine Learning*.

Raj Chetty, Nathaniel Hendren, and Lawrence Katz. Nber working paper series the effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. 2015. URL `https://api.semanticscholar.org/CorpusID:3816986`.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024.

Richard Doll and Austin Bradford Hill. Smoking and carcinoma of the lung; preliminary report. *British medical journal*, 2 4682:739–48, 1950. URL `https://api.semanticscholar.org/CorpusID:41795917`.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019.

Robert F Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the econometric society*, pages 987–1007, 1982.

Lawrence C Evans. *Partial differential equations*, volume 19. American mathematical society, 2022.

Andreas Gerhardus and Jakob Runge. High-recall causal discovery for autocorrelated time series with latent confounders. *Advances in neural information processing systems*, 33:12615–12625, 2020.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.

Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International conference on machine learning*, pages 1724–1732. PMLR, 2017.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kenji Kawaguchi. Deep learning without poor local minima. *Advances in neural information processing systems*, 29, 2016.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pages 2207–2217. PMLR, 2020.

Andrew R. Lawrence, Marcus Kaiser, Rui Sampaio, and Maksim Sipos. Data generating process to evaluate causal discovery techniques for time series data. *Causal Discovery & Causality-Inspired Machine Learning Workshop at Neural Information Processing Systems*, 2020.

Zida Liang, Jiayi Zhu, and Weiqiang Sun. Why attention fails: The degeneration of transformers into mlps in time series forecasting. *arXiv preprint arXiv:2509.20942*, 2025.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. *arXiv preprint arXiv:2502.00816*, 2025.

Enzhe Lu, Zhejun Jiang, Jingyuan Liu, Yulun Du, Tao Jiang, Chao Hong, Shaowei Liu, Weiran He, Enming Yuan, Yuzhi Wang, et al. Moba: Mixture of block attention for long-context llms. *arXiv preprint arXiv:2502.13189*, 2025.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pages 1614–1623. PMLR, 2016.

Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):19, 2019.

Matthew Newman, Michael A Alexander, Toby R Ault, Kim M Cobb, Clara Deser, Emanuele Di Lorenzo, Nathan J Mantua, Arthur J Miller, Shoshiro Minobe, Hisashi Nakamura, et al. The pacific decadal oscillation, revisited. *Journal of Climate*, 29(12):4399–4427, 2016.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *ArXiv*, abs/2006.10201, 2020a. URL `https://api.semanticscholar.org/CorpusID:219792014`.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020b.

Ignavier Ng, Biwei Huang, and Kun Zhang. Structure learning with continuous optimization: A sober look and beyond. In *Causal Learning and Reasoning*, pages 71–105. PMLR, 2024.

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

Judea Pearl. *Causality*. Cambridge university press, 2009.

Ronan Perry, Julius Von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. *Advances in Neural Information Processing Systems*, 35:10904–10917, 2022.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. *Advances in neural information processing systems*, 26, 2013.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT press, 2017.

Nicoleta-Monica Popa-Fotea. Dexamethasone in hospitalized patients with covid-19. *Romanian Archives of Microbiology and Immunology*, 80, 2021. URL `https://api.semanticscholar.org/CorpusID:235443953`.

Bryan L. Roth. Dreadds for neuroscientists. *Neuron*, 89:683–694, 2016. URL `https://api.semanticscholar.org/CorpusID:11550590`.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on uncertainty in artificial intelligence*, pages 1388–1397. Pmlr, 2020.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting causal associations in large nonlinear time series datasets. *arXiv preprint arXiv:1702.07007*, 2017.

Sofia Serrano and Noah A Smith. Is attention interpretable? *arXiv preprint arXiv:1906.03731*, 2019.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, Kenneth Bollen, and Patrik Hoyer. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research-JMLR*, 12(Apr):1225–1248, 2011.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Xiangyu Sun, Oliver Schulte, Guiliang Liu, and Pascal Poupart. Nts-notears: Learning nonparametric dbns with prior knowledge. *arXiv preprint arXiv:2109.04286*, 2021.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2021.

John M Wallace and David S Gutzler. Teleconnections in the geopotential height field during the northern hemisphere winter. *Monthly weather review*, 109(4):784–812, 1981.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, YX Wei, Lean Wang, Zhiping Xiao, et al. Native sparse attention: Hardware-aligned and natively trainable sparse attention. *arXiv preprint arXiv:2502.11089*, 2025.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2023.

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

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Learning sparse nonparametric dags. *ArXiv*, abs/1909.13189, 2019. URL `https://api.semanticscholar.org/CorpusID:203593218`.

A Appendix

A.1 Identifiability of the causal structure

We formalize when gradients of the population regression recover the lagged causal parents. Let $X = (X_1, \dots, X_d)$ collect all covariates formed by stacking all variables over lags $1{:}L$ at time t, and let $Y := Y_t$. Write $S \subseteq \{1, \dots, d\}$ for the index set of the direct time-lagged parents $\mathrm{Pa}(Y)$ inside X.

Assumptions and definitions. We work under the following standard conditions (definitions inlined; references in parentheses):

- **Causal sufficiency**: all common causes of the modeled variables are observed; no latent confounders [Pearl, 2009, Spirtes et al., 2000].
- **No instantaneous effects**: edges from time t to t are absent; all parents of Y_t live at lags $\ell \geq 1$ (time-lagged SCM; see, e.g., Peters et al., 2013, Runge et al., 2017).
- **Lag-window coverage**: the constructed design vector X contains all true lagged parents of Y_t (the chosen maximum lag L is at least the causal horizon).
- **Causal Markov, and Faithfulness**: $Y_t \perp (\text{Past} \setminus \mathrm{Pa}(Y_t)) \mid \mathrm{Pa}(Y_t)$ (Causal Markov property), and the distribution is faithful to the underlying time-lagged graph so that no independences arise from measure-zero cancellations [Pearl, 2009, Spirtes et al., 2000, Peters et al., 2013].
- **Support and regularity**: the law of X admits a density supported on a rectangle $\Omega \subset \mathbb{R}^d$ (no deterministic constraints/collinearity), and the population regression

  $$f^*(x) \;:=\; \mathbb{E}[\, Y \mid X = x \,]$$

  lies in $W^{1,2}_{\mathrm{loc}}(\Omega)$, i.e., is weakly differentiable with square-integrable partial derivatives [Evans, 2022, Adams and Fournier, 2003].

Define the *gradient energy* of coordinate j by

$$G_j \;:=\; \mathbb{E}\Big[\big(\partial_{x_j} f^*(X)\big)^2\Big], \quad j = 1, \dots, d.$$

Lemma 1 (Markov projection)**.** *Under the Causal Markov property and no instantaneous effects, there exists a measurable g such that for all $x \in \Omega$,*

$$f^*(x) \;=\; g(x_S), \qquad S = \text{indices of } \mathrm{Pa}(Y_t).$$

In particular, $\mathbb{E}[Y \mid X = x] = \mathbb{E}[Y \mid X_S = x_S]$.

Proof. By the Causal Markov property and the absence of instantaneous effects, $Y \perp X_{S^c} \mid X_S$. Therefore $\mathbb{E}[Y \mid X = x] = \mathbb{E}[Y \mid X_S = x_S]$ for all $x \in \Omega$. Let $g(u) := \mathbb{E}[Y \mid X_S = u]$. Then $f^*(x) = g(x_S)$. The function g is measurable by standard properties of regular conditional expectations. □

Lemma 2 (Zero weak partial implies no dependence)**.** *Let $f \in W^{1,1}_{\mathrm{loc}}(\Omega)$ on a rectangle $\Omega \subset \mathbb{R}^d$. If $\partial_{x_j} f = 0$ almost everywhere on Ω, then there exists a measurable h with $f(x) = h(x_{-j})$ almost everywhere. Conversely, if f does not depend on x_j, then $\partial_{x_j} f = 0$ almost everywhere.*

Proof. Assume $\partial_{x_j} f = 0$ almost everywhere. Fix x_{-j}. For almost every line $t \mapsto (t, x_{-j})$, the one-dimensional fundamental theorem of calculus yields $f(t_2, x_{-j}) - f(t_1, x_{-j}) = \int_{t_1}^{t_2} \partial_{x_j} f(s, x_{-j})\, ds = 0$, so $f(t, x_{-j})$ is (a.e.) constant in t. Thus there is a measurable h with $f(x) = h(x_{-j})$ a.e. Conversely, if f does not depend on x_j, then its weak partial $\partial_{x_j} f$ is 0 almost everywhere. □

Connecting dependence and gradients. By Lemma 1, f^* depends only on the parent coordinates X_S. For any coordinate j, "f^* does not depend on x_j" is equivalent to "$\partial_{x_j} f^*(x) = 0$ almost everywhere," by Lemma 2. Hence $G_j = \mathbb{E}[(\partial_{x_j} f^*(X))^2]$ equals 0 exactly when f^* ignores x_j. Under Faithfulness, this happens precisely for non-parents and not for true parents.

Theorem 1 (Gradient characterization of lagged parents)**.** *Under the assumptions in this subsection, for each coordinate $j \in \{1, \dots, d\}$,*

$$G_j = 0 \iff j \notin S.$$

In particular, if $k \in S$ then $G_k > 0$.

Proof. ($\Leftarrow$) If $j \notin S$, then by Lemma 1 $f^*(x) = g(x_S)$ and thus it does not depend on x_j. Lemma 2 gives $\partial_{x_j} f^* = 0$ a.e., so $G_j = 0$.

($\Rightarrow$) If $G_j = 0$, then $\partial_{x_j} f^* = 0$ a.e., so by Lemma 2 f^* does not depend on x_j. Hence $Y \perp X_j \mid X_{-j}$. By Faithfulness, this is impossible for a true parent, so $j \notin S$. For any $k \in S$, the contrapositive implies $\partial_{x_k} f^*$ is nonzero on a set of positive measure, and therefore $G_k > 0$. □

A.2 Attention LRP as a surrogate for gradient energy

Layer-wise Relevance Propagation (LRP) decomposes a model's output $f(x)$ into relevance scores assigned to input coordinates. For efficiency and simplicity, we adopt the Input×Gradient formulation of ε-LRP, which expresses LRP as a single chain of Jacobian–vector products (one backward pass) with small, local modifications to the backward rule at nonlinearities and at attention/normalization layers. This implementation is equivalent to ε-LRP up to a layer-wise rescaling and closely follows the efficient Attention-LRP formulation used for transformers [Achtibat et al., 2024].

Concretely, for a trained forecaster $\hat{f}$ and a scalar prediction $z := \hat{f}(x)$ (e.g., the mean for regression or a logit/probability for classification), we define per-sample relevance by

$$R(x) \;:=\; x \;\odot\; \widetilde{\nabla}_x z,$$

where $\widetilde{\nabla}_x$ denotes a gradient computed with the modified local Jacobians described below. Aggregating coordinates gives a global score

$$\tilde{G}_j \;:=\; \mathbb{E}\big[\,|R_j(X)|\,\big],$$

used as a monotone proxy for $G_j = \mathbb{E}[(\partial_{x_j} f^*(X))^2]$.

Core (Input×Gradient) LRP equations. For computational efficiency, we use the gradient-input reformulation in attention-aware LRP [Achtibat et al., 2024]:

$$R(x) \;=\; x \;\odot\; \Big(J_1\, J_2 \cdots J_L\, e_i\Big) \qquad \text{(Input×Gradient with modified local Jacobians).} \quad \text{(IG-1)}$$

The same chain-of-Jacobian idea applies to attention and normalization layers in transformers. In practice, this yields LRP attributions in a single backward pass, after which token-level relevances are aggregated to $\tilde{G}_j$ as above.

A.3 Experiment setups

Data Generation and Simulation

Simulator. We use the CDML-NeurIPS2020 structural time series simulator to sample datasets [Lawrence et al., 2020]. We use a linear baseline and multiple variants in different dimensions such as the number of variables, maximum lag, noise type, non-stationarity, and latent variables. For the variants, we vary only the property of interest of the data generation process compared to the linear baseline and use multiple sample sizes to see how the performance changes with the sample size (5e4, 1e5, 1e6).

Variables and lags. For a system with N observed variables and maximum lag K. We disable instantaneous effects and set the transition probability to 0.3. Latent and noise autoregression are set to 0 unless noted. We use the number of variables $10, 25, 50$ to study the high-dimensional cases.

Control graph density via expected in-degree. To obtain comparable sparsity across N and K, we specify an expected in-degree $E_{\text{in}} = 3$ per node (aggregated across all parent candidates).

Structural functions and nonlinearity. We control the nonlinearity complexity by employing functional forms as follows (first 3 are additive noise models): (1) piecewise linear (PL): mixture of linear, piecewise linear, (2) and monotonic (sum-of-sigmoids) functions (3) MLP (add): multi-layer perceptron (MLP) with additive noise injection (4) MLP (concat): MLP aggregation with noise concatenation.

Noise types. We consider three noise types: Gaussian (in linear baseline), Uniform, and Mixed. The mixed noise is a fixed mixture over distributions [Gaussian, Uniform, Laplace, Student's t]. We also study non-equal variance noise, with a small range from 0.5 to 5 and a large range from 0 to 10.

Non-stationarity. To study how different approaches behave under time-varying causal structure, we partition the sequence into S contiguous segments ($S \in \{2, 5, 10\}$) and independently generate each segment with a randomly sampled graph. We construct two settings here; the first is the regular setting where each domain has the same maximum lag (5) and the data generation process is all linear. In the extreme setting, each domain might not have the same maximum lag (up to 5) and the data generation process is composed of random functions (from linear and nonlinear function set).

Latent variables. We examine the robustness of discovery methods in the presence of latent variables. We set the number of latent variables to $L \in \{3, 5, 10\}$.

A.4 Training details and model architecture

We train autoregressive Transformers on lag-K windows, after per-variable z-score normalization. We use an embedding dimension of 64, 4 attention heads, and either 1 ("shallow") or 4 ("deep") layers with pre-LayerNorm, residual connections, and a 2-layer ReLU feed-forward; causal masking, node/time embeddings. Models are optimized with Adam (learning rate 1e-3, batch size 256) under an MSE objective, with gradient clipping at 1.0. We use the official implementations for PCMCI [Runge, 2020], VAR-LiNGAM [Hyvärinen et al., 2010], TCDF [Nauta et al., 2019], NTS-NOTEARS [Sun et al., 2021] and Granger causality implementations from a collection repository of time series causal discovery algorithms (https://github.com/ckassaad/causal_discovery_for_time_series). We use the default hyperparameters from the official implementation.

A.5 Compute resources

All transformer experiments are implemented in PyTorch and executed in FP32 precision on a single NVIDIA A100 GPU with actual memory usage below 24GB. Experiments that exceed 6 hours of runtime, including both our transformer approach and baseline methods, are terminated and classified as timeouts.

A.6 More Discussions

The difficulty of non-convex optimization. Traditional continuous-optimization approaches to causal discovery struggle with non-convex loss landscapes. Even under identifiability and with the correct objective, nonconvexity from unequal noise variances and nonlinearities can make structure recovery nearly intractable; outcomes hinge on fragile initialization, especially with limited, homogeneous data [Ng et al., 2024]. By contrast, large-scale transformer pretraining operates in a different regime: overparameterized networks have benign landscapes with many global minima [Du et al., 2019]; in high dimensions, bad local minima are rare while saddle points dominate [Kawaguchi, 2016]; and the stochasticity of SGD helps escape saddles and favors flatter, more generalizable regions [Jin et al., 2017]. This geometry enables transformers to function as scalable causal learners, effectively sidestepping the non-convex barriers constraining classical methods.

The role of prediction objective. A Gaussian likelihood (MSE) is convenient and coherent with a Gaussian noise prior, but richer likelihoods can better capture heteroskedastic or multimodal dynamics and sharpen attribution. Objectives should match the data distribution and exogenous noise. For instance, adding degrees of freedom can accommodate unequal noise variances in highly heteroskedastic data. Promising alternatives include flow-matching and diffusion objectives, as well as quantile and energy-based losses; these better model stochasticity and complex distributions. As predictive fidelity improves, the implicitly learned structure should become more accurate.

The need of better structure priors. Our experiments indicate that vanilla decoder-only transformers are sample-hungry for recovering correct structures from a single nonlinear generator, and heterogeneous mixtures are harder—evidence of weak inductive bias for causal structure. Causal theory offers mature priors to close this gap. Independent Causal Mechanisms and minimality enforce mechanism independence and modular factorization, yielding cross-environment invariances and improved sample efficiency [Peters et al., 2017, Huang et al., 2020]. Sparsity further reduces the search space, making learning more tractable in noisy settings [Ng et al., 2020b, Zheng et al., 2018, Perry et al., 2022]. Recent large-language-model architectures echo these ideas: gated and block-sparse attention instantiate sparsity and modularity, mitigating spurious context coupling and improving long-context retrieval and robustness to distribution shift [Yuan et al., 2025, Lu et al., 2025]. Finally, scalable, native modeling of latent variables and instantaneous effects broadens the class of structures beyond lagged processes. Integrating such priors should improve efficiency, generalizability, and robustness.

Context with prior findings on Transformers for time series. A growing body of work has argued that transformers are not uniformly superior for time series forecasting, and that simple linear baselines can be competitive or even stronger under common settings [Zeng et al., 2023]. We view these results less as evidence against the architecture and more as a representation bottleneck: unlike language (byte-pair encoding) and vision (patches), time series still lacks a widely adopted, structure-aware tokenization that has a high compression rate and exposes salient inductive biases. In the absence of such tokens, architectural capacity is under-utilized and empirical gains hinge on data preprocessing and normalization choices, precisely the phenomenon documented by prior analyses [Chen et al., Liang et al., 2025]. Moreover, recent time series foundation models pretrained on large, heterogeneous corpora have demonstrated strong zero-shot forecasting performance [Ansari et al., 2025, Das et al., 2024, Liu et al., 2025], indicating the scaling potential of Transformer-based backbones for temporal data. The scalability demonstrated in language and vision, where model size and data scale reliably translate into better representation learning, suggests that the limiting factor for time series is the input representation, not the backbone. Our findings echo this perspective from a causal angle. Decoder-only Transformers trained autoregressively possess the capacity and expressivity to encode complex and diverse temporal causal graphs in a scalable way. We should expect that with better time series tokenization, the same machinery that powers foundation models can also serve as a practical, extensible vehicle for forecasting and causal discovery.

A.7 Complete experiment results

A.7.1 Sample scaling behavior in nonlinear settings

We examine the capability of the transformer in learning nonlinear interactions, considering settings from simple to complex: additive noise models with linear, monotonic, mixture of piecewise linear and monotonic, and multi-layer perceptron (MLP) as nonlinear functions, and non-additive noise models with MLP as mixing functions of variables and noise. We observe a trade-off between data efficiency and expressivity. While traditional methods employing simple estimators and search heuristics from human prior (e.g., DYNOTEARS, VAR-LiNGAM, PCMCI) can achieve good performance efficiently in simple cases like linear settings, a decoder-only transformer generally works better when the data scales and shows a consistent accuracy improvement as data increases.

A.7.2 Sample scaling behavior in non-stationary settings

Here we construct two kinds of non-stationarity: the regular setting randomly samples linear structures with a fixed maximum lag for each domain, and the extreme setting randomly samples structure and random functions (from linear and nonlinear function set) for each domain, within a range

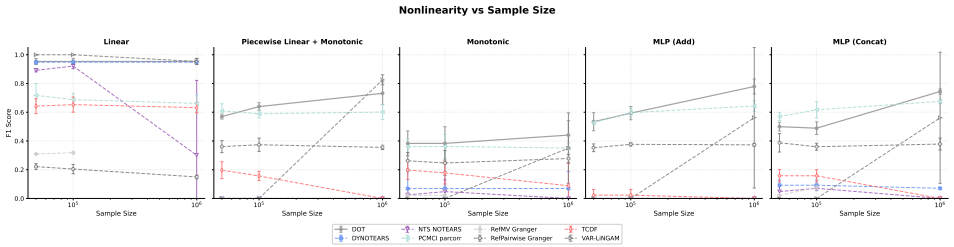


Figure 6: **Nonlinear dependencies.** F1 scores averaged across seeds vs. sample size in different nonlinear settings.

of maximum lags. Unlike traditional methods that are intractable with more data, the transformer shows consistent improvement across sample sizes. In non-stationary settings, the model learns to handle multiple local mechanisms within a single framework. As the sample size increases, the transformer better separates and routes different causal structures corresponding to distinct regimes (Figure 2E). This scaling behavior mirrors the remarkable zero-shot generalization of large language models and the rich world knowledge and rules they learned [Kaplan et al., 2020, Brown et al., 2020]. The pretraining provides the model chances to learn such diverse causal structures and find deeper cross-domain patterns by connecting them. On the other hand, these learning curves along with nonlinearity experiment results show the limitation of weak model prior and data hungriness. When the number of domains increases, learning to accurately model and switch between them becomes much harder and requires much more data. It implies the structure learning of foundation models might be limited because of data insufficiency; thus, they often rely on imperfect structures (spurious correlations) and induce hallucinations.

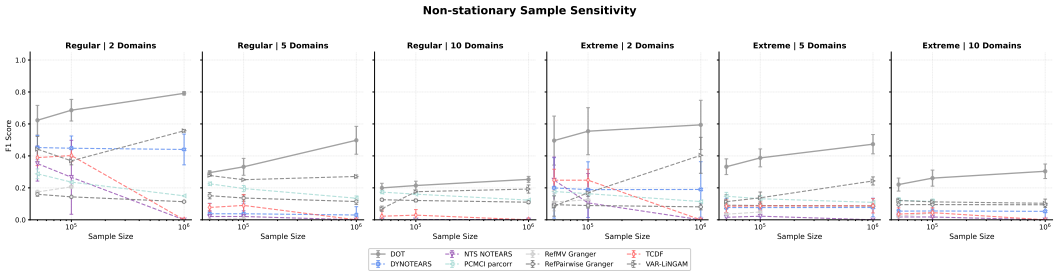


Figure 7: **Non-stationary dependencies.** F1 scores averaged across seeds vs. sample size in different non-stationary settings.

A.7.3 Performance under non-stationary settings with realistic minimal changes.

In real world settings, the changes across domains are often minimal and gradual. We construct a setting where a small part of the structure is rewired randomly compared to the previous regime. We see that the data efficiency is much higher compared to the randomly sampled setting in both regular and extreme settings (see paragraph A.7.2). Note that here we do not inject any prior and constraint about minimal changes to the architecture, and we should expect it to be much more data-efficient when we incorporate such prior knowledge natively into the transformer.

A.7.4 Noise and latent variable robustness.

The transformer demonstrates robust performance across different noise distributions, maintaining consistent accuracy regardless of noise type or the variance properties of noise (see Figure 9). While we observe a performance drop of continuous optimization methods like DYNOTEARS and TCDF in non-equal noise variance settings aligned with Ng et al. [2024], the decoder-only transformer remains stable and accurate. However, due to the lack of a latent variable modeling mechanism, transformers are prone to learn spurious links and degrade as the number of latent variables increases, while traditional methods considering sparsity alleviate this influence.

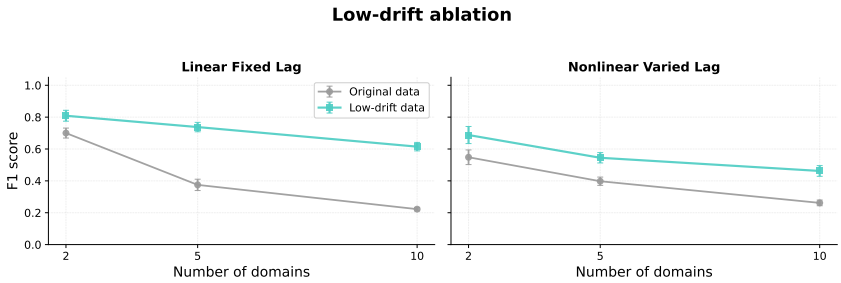


Figure 8: **Performance under non-stationary settings with minimal changes across regimes.** Comparison of F1 scores between randomly sampled regime changes (original data) and minimal changes (low-drift data) where only a small part of the structure is rewired. The minimal changes setting is more realistic and our approach shows much higher data efficiency in both linear and nonlinear non-stationary scenarios, demonstrating the applicability of our approach in real-world settings.

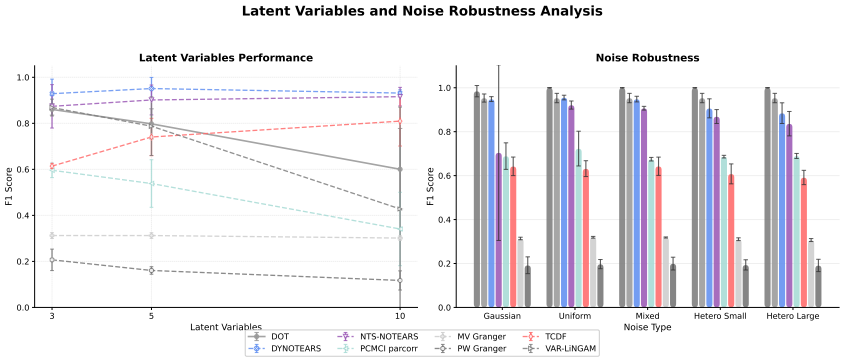


Figure 9: **Robustness to latent variables and noise. Left:** F1 sores on scenarios including different amount of latent variables. **Right:** F1 scores on different kinds of noise (equal variance and non-equal variance)

A.7.5 Attention and gradient attribution.

We also evaluate non-gradient proxies, such as raw attention scores, for recovering causal structure (see Figure 11). We see the relevance between attention scores and gradient attributions is different for deep and shallow transformers. In deep transformers, attention scores barely reveal any information about the structure model learned, while in one-layer transformers, structures extracted from attention scores are much more accurate and aligned with the LRP's outputs. This aligns with findings that as depth increases, repeated attention routing and residual MLP updates mix token representations, making the attention not faithfully represent token dependency [Serrano and Smith, 2019, Jain and Wallace, 2019].

A.7.6 The effect of model depth.

The depth of the transformer primarily affects capacity and the expressivity to capture complex dependencies. With more layers, the transformer can model more complex structures and longer-range effects. In our nonlinear and long-range settings, deeper transformers achieve higher accuracy in recovering causal structure and show clear advantages with LRP readout. This highlights the potential of deep transformers for highly heterogeneous, long-range dynamics, echoing the success of pretrained large language and vision models.

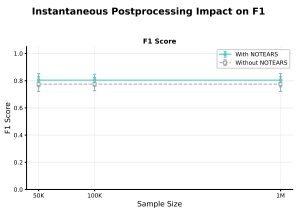


Figure 10: **Handling instantaneous relationships with post-processing.** Comparison of F1 scores between using NOTEARS as post-processing and using transformer alone.

A.7.7 The effect of graph binarization.

Different binarization rules can lead to distinct causal graphs. We compare thresholding and top-k. Thresholding performs similarly

to top-k when the number of variables and the lag window are moderate, but its precision degrades as the context length grows. The importance of variables varies non-uniformly across lags with longer contexts. Top-k provides a simple, effective way to control the precision–recall trade-off. Similar to max-depth limits in classical methods (PC, GES, etc.), choosing k with domain knowledge lets us control edge density (e.g., use a small k when the goal is to recover only the most important interactions).

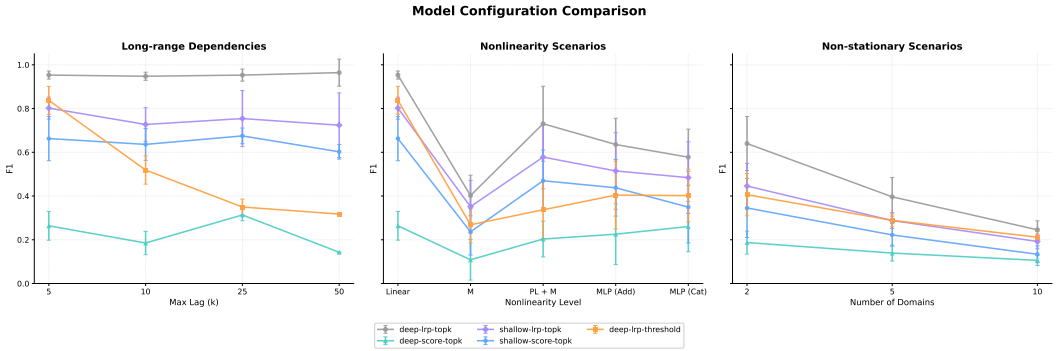


Figure 11: **Transformer variants performance comparison on challenging regimes. Left:** F1 scores on long-range dependencies. **Middle:** F1 scores on linear and nonlinear dynamics. **Right:** F1 scores on non-stationary dependencies.

A.7.8 The potential of handling instantaneous relationships.

Decoder-only transformer lacks the ability to model instantaneous relationships due to its autoregressive nature. We can take a similar approach as the one for latent confounders to handle instantaneous relationships, by separating the instantaneous parts to a statistical causal discovery method [Shimizu et al., 2011, Spirtes et al., 2000]. We apply NOTEARS [Zheng et al., 2018] to the output residuals of the transformer to estimate the instantaneous causal structure and combine it with the lagged part from the transformer. This procedure improves the accuracy when encountering instantaneous relationships. However, it is not a native capability of the transformer and the instantaneous part is largely limited by the prediction accuracy of the transformer. We leave the native instantaneous relationship and latent variable modeling in transformer for future work.

A.7.9 More results on uncertainty analysis.

Here we show additional results on uncertainty analysis. We use a linear gaussian dataset and a nonlinear dataset (sigmoid) with 5 variables and 2 lags for the ease of analysis and simplicity of visualization. We take two measures here, original relevance scores and the ranking of relevance scores. The second quantized measure is more stable and comparable across different pairs of variables. The mean over standard deviation of the ranking is introduced as a more general metric, considering the uncertainty and strength of the edges at the same time. We also show two different kinds of top-k for binarization, the row-wise top-k (choose the top-k variables with largest relevance scores for each child variable as its parents, mainly used in the main experiments) and the global top-k (consider the edges with top-k largest relevance scores across all rows/child variables as causal edges). For the row-wise top-k, we take top-3 for each child variable as its parents, and for the global top-k, we take top-15 across all rows as the true edges.

The standard deviation of the original scores is not straightforward to interpret, and we cannot directly compare them across different pairs of variables. It presents an overall trend that larger mean scores are more likely to have larger variance. We use the row-wise ranking of the original relevance scores to compute standard deviation, which is clear and aligns with our intuition that the true causal connections should be more stable than the false ones. We see that the model tends to have a higher consistency for the edges with stronger strengths. These high confidence (low standard deviation and high mean ranking) edges are often the true edges. Based on the ranking, we propose another metric, mean over standard deviation of the ranking, as a more general measure of edge existence. Furthermore, we find for the graph with varied degrees of different nodes, row-wise top-k is a hard

truncation and more focused on local structures that might miss and add some edges without any reason. In such circumstances, global top-k is more robust as it considers the most predominant edges in the whole system as true edges. By quantizing the original scores with row-wise rankings, it can also recognize the local structures even when their original strengths are weak. We surprisingly find that the ranking measure is less noisy and gives better results than the original continuous scores, in both linear and nonlinear settings. However, we do not observe any advantages of using the combined metric with global top-k and row-wise top-k compared to only using the mean of rankings, with respect to identification accuracy. We leave the precise calibration and more structure extraction strategies based on uncertainty for future work.

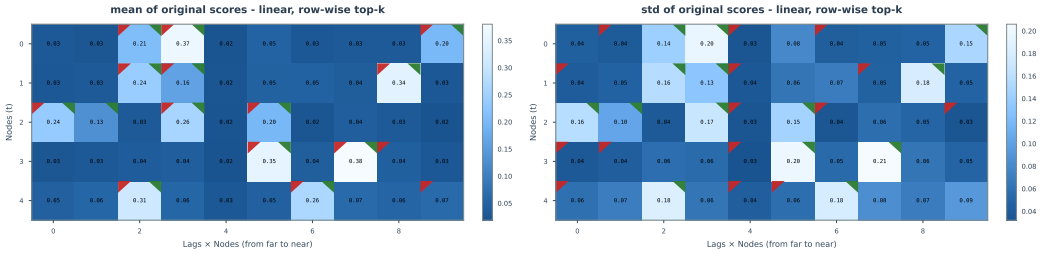


Figure 12: **Mean and standard deviation of relevance scores in the linear setting:** **(A)** Heatmap showing the mean of edge attributions across all samples. **(B)** Heatmap showing the standard deviation of edge attributions across all samples. The top-left red triangle means that model predicts there is a causal edge and top-right green triangle means that there is a true edge between the two variables.

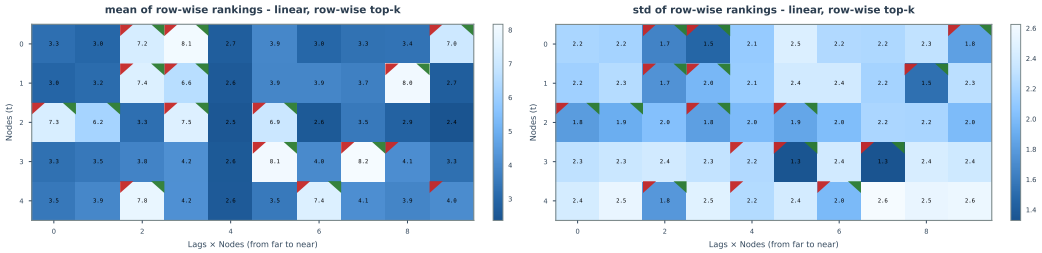


Figure 13: **Mean and standard deviation of row-wise rankings in the linear setting:** **(A)** Heatmap showing the mean of the ranking of the edge attributions across all samples. **(B)** Heatmap showing the standard deviation of the ranking of the edge attributions across all samples. The top-left red triangle means that model predicts there is a causal edge and top-right green triangle means that there is a true edge between the two variables.

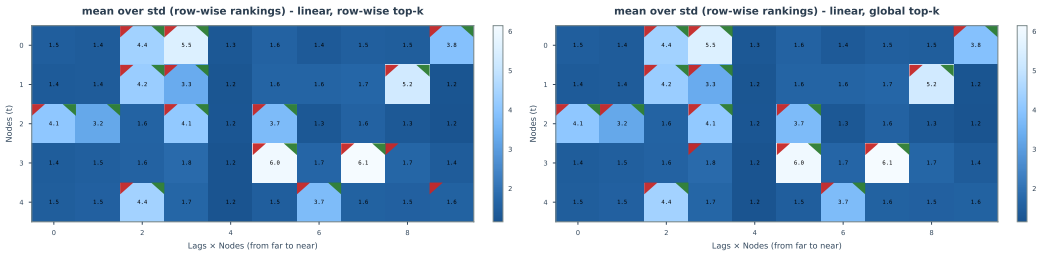


Figure 14: **Row-wise top-k and global top-k in the linear setting:** **(A)** Heatmap showing the mean over standard deviation of the ranking of the edge attributions across all samples. **(B)** Heatmap showing the standard deviation of the ranking of the edge attributions across all samples. The global top-k select more accurate causal edges than the row-wise top-k. The top-left red triangle means that model predicts there is a causal edge and top-right green triangle means that there is a true edge between the two variables.

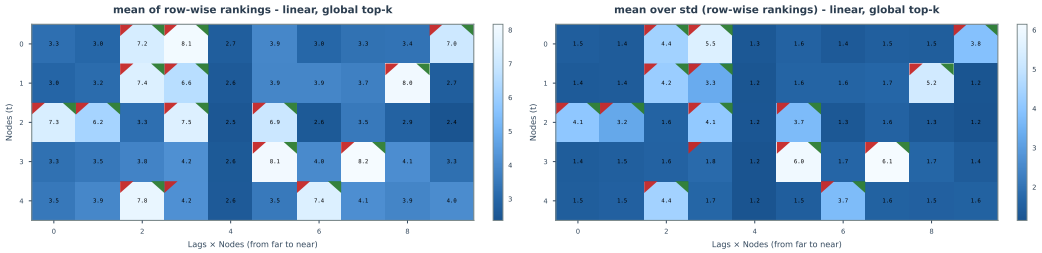


Figure 15: **Global top-k based on mean of rankings and mean over standard deviation of rankings in the linear setting: (A)** Heatmap showing the mean of the ranking of the edge attributions across all samples and predictions selected by the global top-k. **(B)** Heatmap showing the mean over standard deviation of the ranking of the edge attributions across all samples and predictions selected by the global top-k. The top-left red triangle means that model predicts there is a causal edge and top-right green triangle means that there is a true edge between the two variables.

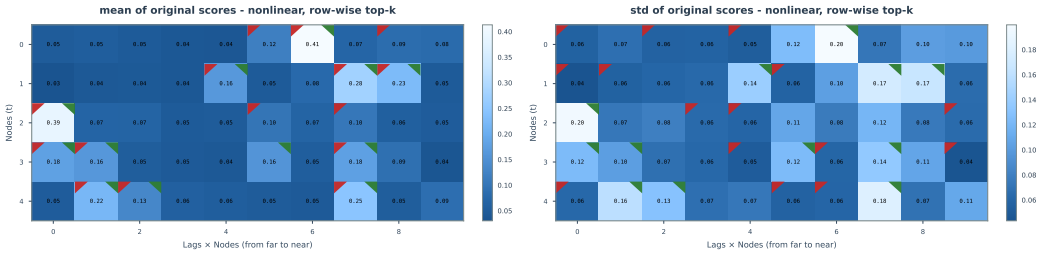


Figure 16: **Mean and standard deviation of relevance scores in the nonlinear setting: (A)** Heatmap showing the mean of edge attributions across all samples. **(B)** Heatmap showing the standard deviation of edge attributions across all samples. The top-left red triangle means that model predicts there is a causal edge and top-right green triangle means that there is a true edge between the two variables.

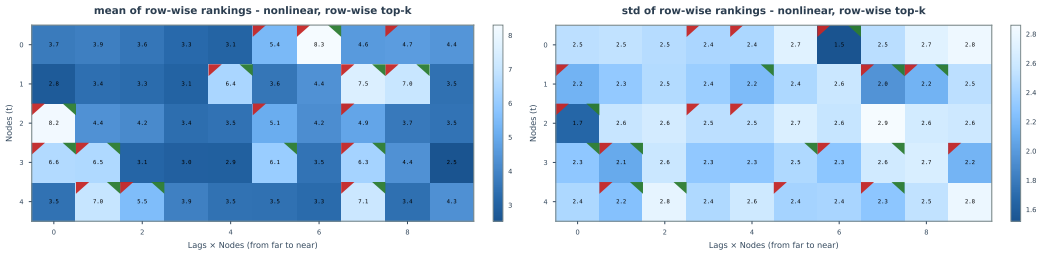


Figure 17: **Mean and standard deviation of row-wise rankings in the nonlinear setting: (A)** Heatmap showing the mean of the ranking of the edge attributions across all samples. **(B)** Heatmap showing the standard deviation of the ranking of the edge attributions across all samples. The top-left red triangle means that model predicts there is a causal edge and top-right green triangle means that there is a true edge between the two variables.

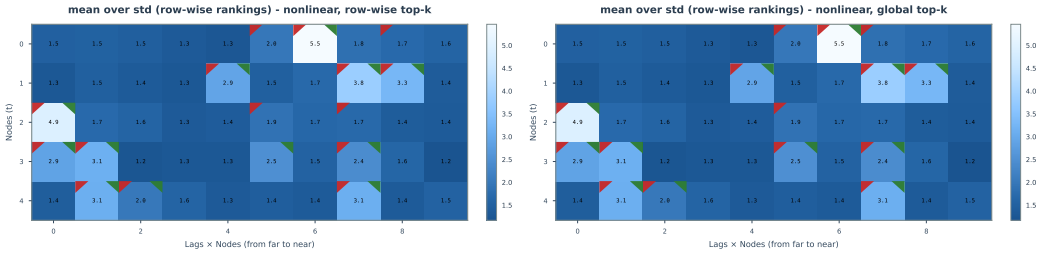


Figure 18: **Row-wise top-k and global top-k in the nonlinear setting: (A)** Heatmap showing the mean over standard deviation of the ranking of the edge attributions across all samples. **(B)** Heatmap showing the standard deviation of the ranking of the edge attributions across all samples. The global top-k select more accurate causal edges than the row-wise top-k. The top-left red triangle means that model predicts there is a causal edge and top-right green triangle means that there is a true edge between the two variables.

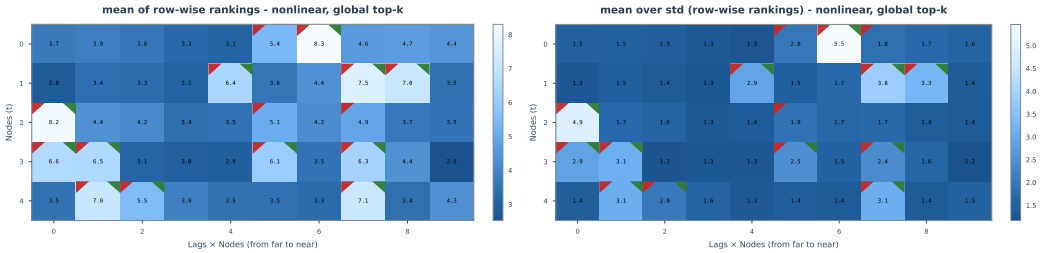


Figure 19: **Global top-k based on mean of rankings and mean over standard deviation of rankings in the nonlinear setting: (A)** Heatmap showing the mean of the ranking of the edge attributions across all samples and predictions selected by the global top-k. **(B)** Heatmap showing the mean over standard deviation of the ranking of the edge attributions across all samples and predictions selected by the global top-k. The top-left red triangle means that model predicts there is a causal edge and top-right green triangle means that there is a true edge between the two variables.

We show the histograms of edge strengths in both linear and nonlinear settings. Predictions in both linear and nonlinear settings miss a small part of edges with low strengths. In nonlinear settings, the prediction strengths are less uniform, and this over-concentration makes it wrongly identify some edges and miss some true ones. We also find that the ratio of medium-strength edges (0.1 - 0.3) in lag 1 is higher than in the other lags. It shows that the transformer is prone to assign more weights on the last time stamps and pay less attention to the further part, even though they might be true parents. It is largely due to the self-attention mechanism and causal masking that the transformer is more likely to attend to the last time stamps.

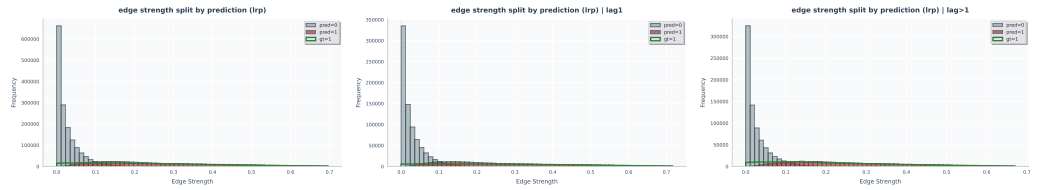


Figure 20: **Histograms of edge strengths in the linear setting: (A)** Histogram of edge strengths in the linear setting. **(B)** Histogram of edge strengths in the linear setting with lag 1. **(C)** Histogram of edge strengths in the linear setting with lag > 1.

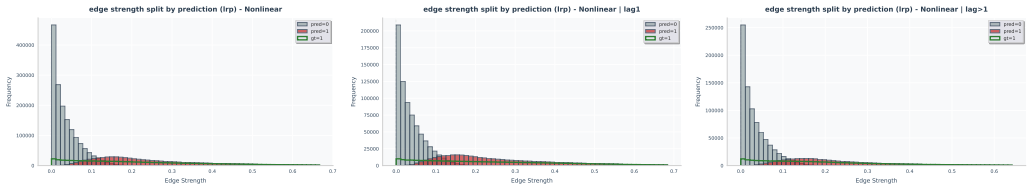


Figure 21: **Histograms of edge strengths in the nonlinear setting:** **(A)** Histogram of edge strengths in the nonlinear setting. **(B)** Histogram of edge strengths in the nonlinear setting with lag 1. **(C)** Histogram of edge strengths in the nonlinear setting with lag > 1.