Transformer Is Inherently a Causal Learner

Anonymous Author(s)
Affiliation
Address
email

Abstract

We reveal that decoder-only 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 laws where causal 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 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 proposing 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].

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 the 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,

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

2002]. Optimization approaches require careful tuning to achieve the right balance between likelihood and structural regularization [Zheng et al., 2018, Ng et al., 2020, 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

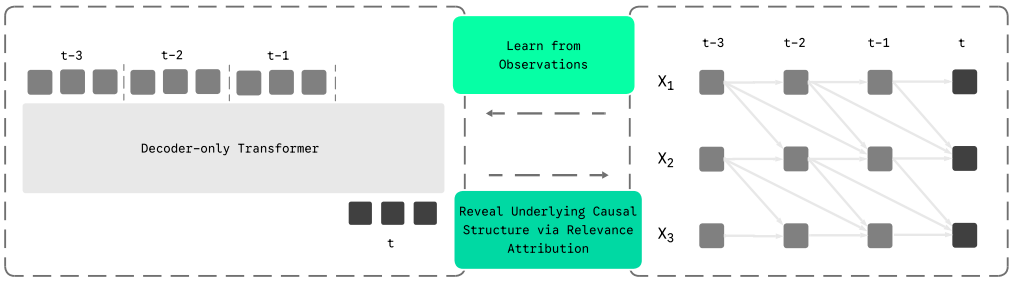

Figure 1: **Data generation and transformer-based causal discovery. Left:** A decoder-only transformer trained for next-step prediction. Tokens are lagged observations from $t-L$ to $t-1$; the model predicts X_t from $X_{t-1:t-L}$. **Right:** A lagged data-generating process with $N=3$ and window $L=3$. 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 closed to us 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 $\tilde{G}^{(\ell)}_{j,i} := \mathbb{E}[\,|R^{(\ell)}_{ij}(X)|\,]$, 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 adjust masking rules and combining traditional causal discovery methods as post-processing procedures. See Appendix §A.1 for the identifiability proof and Appendix §A.2 for the LRP–gradient connection.

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 predicting $X_{i,t}$ 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) an 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). We optimize:

$$\widehat{\theta} = \arg\min_{\theta} \; -\frac{1}{T-L} \sum_{t=L+1}^{T} \log p_\theta(X_t \,|\, X_{t-1:t-L}) \; + \; \lambda\, \Omega(\theta), \tag{1}$$

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 (MSE objective), and $\Omega(\theta)$ is optional (e.g., sparsity or entropy regularization; by default we do not use structural penalties).

Sparsity 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 implicitly sparsify: finite capacity, weight decay compress high-dimensional observations into generalizable parameters; 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 exaction. 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 $R^{(\ell)}_{ij}$ 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. Note that 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 normalize each row of $\mathbf{R}$ to sum to one and 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. See Appendix §A.6.4 for a detailed comparison.

3 Experiments

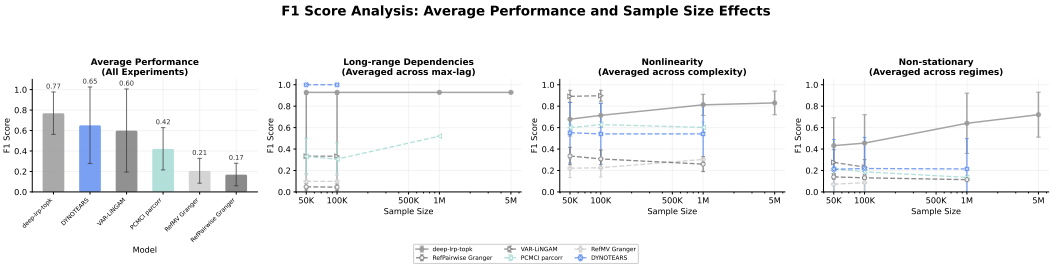

Figure 2: **F1 score analysis across regimes. (A)** Mean F1 across all experiments (averages exclude timeout cases). **(B)** Long-range dependencies: F1 averaged across max-lag vs. sample size. **(C)** Nonlinearity: F1 averaged across function complexity vs. sample size. **(D)** Non-stationarity: F1 averaged across regimes vs. sample size. Missing results indicate method timeouts due to computational limits.

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], and pairwise/multivariate Granger tests [Granger, 1969] across variations in nonlinearity, maximum lag, dimensionality, noise, and non-stationarity. After training, we extract edges with LRP and binarize with a per-target top-k rule.

General capability and complex dependencies. Transformer recovers lagged parents accurately and consistently across settings, achieving comparable or better performance to baseline methods (Figure 2A). It maintains strong performance under nonlinearity, long-term dependencies, large variable sizes, and non-stationarity (Figure 2B and C; see also Figure 4). 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 and analysis of transformer variants are provided in Appendix §A.6.

Capacity and scaling potential. The transformer effectively leverages additional data to improve causal structure modeling accuracy. 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 (Figure 2D). 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.

References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Annual Meeting of the Association for Computational Linguistics*, 2020. URL https://api.semanticscholar.org/CorpusID:218487351.

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.

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.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021.

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.

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.

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.

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.

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.

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.

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.

Ziyu Lu, Anika Tabassum, Shruti Kulkarni, Lu Mi, J Nathan Kutz, Eric Shea-Brown, and Seung-Hwan Lim. Attention for causal relationship discovery from biological neural dynamics. *arXiv preprint arXiv:2311.06928*, 2023.

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.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *ArXiv*, abs/1905.10650, 2019. URL `https://api.semanticscholar.org/CorpusID:166227946`.

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, 2020. URL `https://api.semanticscholar.org/CorpusID:219792014`.

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.

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.

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`.

Raanan Y Rohekar, Yaniv Gurwicz, and Shami Nisimov. Causal interpretation of self-attention in pre-trained transformers. *Advances in Neural Information Processing Systems*, 36:31450–31465, 2023.

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

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.

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

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.

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 $X = (X_1, \ldots, X_d)$ collect all covariates formed by stacking all variables over lags $1{:}L$ at time t, and let $Y := Y_t$. Write $S \subseteq \{1, \ldots, d\}$ 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

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

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 computation efficiency, we use the gradient-input formalization of LRP [Achtibat et al., 2024]. We backpropagate from a chosen scalar component z_i by setting a one-hot seed e_i at the output. Let J_ℓ denote the local Jacobian used in the backward pass at layer ℓ.

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

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 number of variables, maximum lag, noise type, non-stationarity, latent variables. For the variants, we only vary the interested property 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, 5e6).

Variables and lags. For a system with N observed variables and maximum lag K. We disable instantaneous effects and set the transition probability of 0.3. Latent and noise autoregression are set to 0 unless noted.

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: mixture of linear, piecewise linear, and monotonic (sum-of-sigmoids) functions (2) periodic: mixture of linear, piecewise linear, monotonic, and sinusoidal (periodic) 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].

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 random sampled graph.

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 embedding dimension 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, gradient clipping at 1.0.

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 Complete experiment results

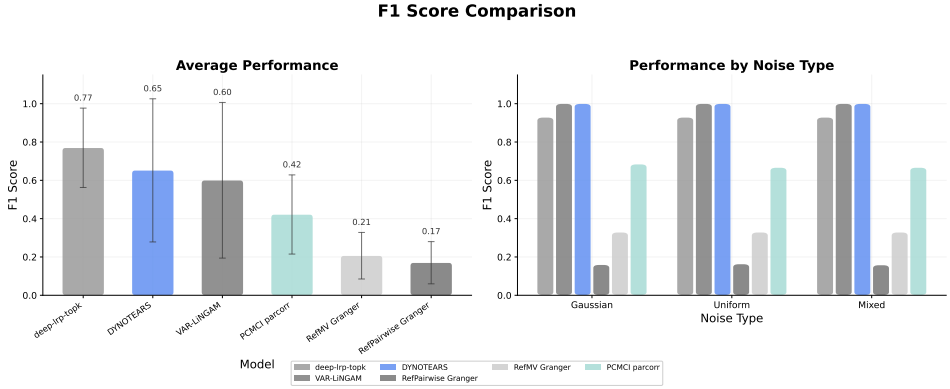

Figure 3: **Performance overview and comparison across different exogenous noise types. Left:** Average performance of transformer and baselines (timeout results are excluded). **Right:** Performance comparison across different exogenous noise types (Gaussian, Uniform, Mixed).

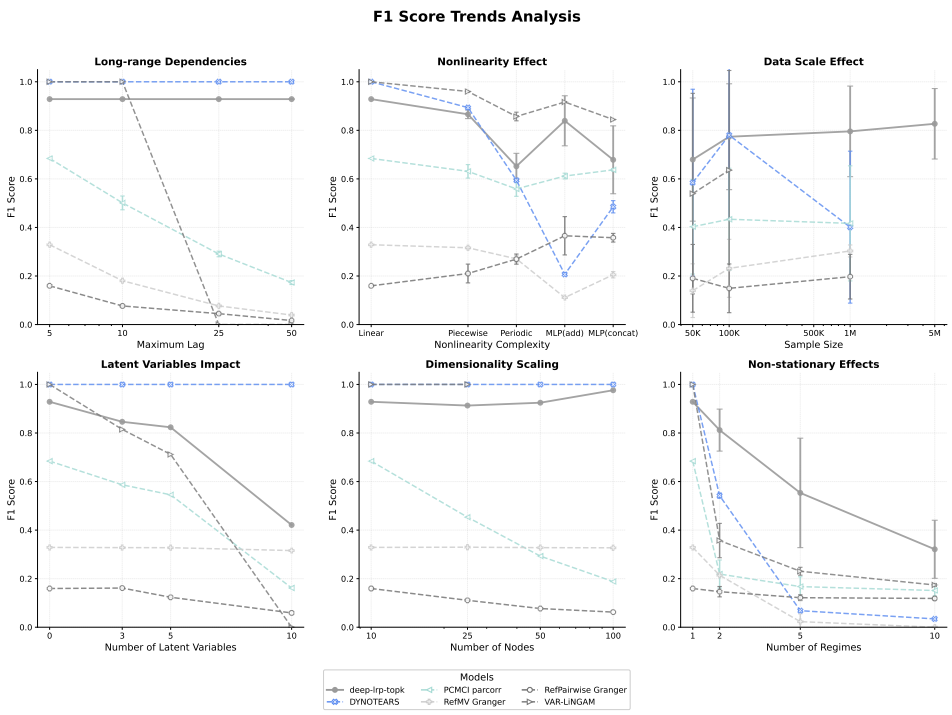

Figure 4: **More experiment results on different settings.** We report the F1 scores of our approach and baselines on different settings of multiple dimensions including maximum lag, nonlinearity, sample size, latent variables, variable size, and non-stationarity.

Overall, the line plots in Fig. 4 show that our Transformer+LRP (top-k) approach is accurate and stable across settings. Performance remains strong as the maximum lag and the number of nodes increase, whereas classical baselines (PCMCI, VAR-LiNGAM, Granger variants) degrade markedly. Under increased nonlinearity and non-additive fusion, our method shows only a modest dip for MLP(add) and recovers for MLP(concat); in contrast, the baselines drop sharply. Larger sample sizes further improve scores and reduce variance. Increasing the number of regimes in non-stationary data lowers all curves. We also observe limitations when data are scarce and latent confounders are present. Future work includes developing natural, implicit sparsity regularization to reduce data requirements, and explicit latent modeling.

A.6.1 Attention and Gradient Attribution

We also evaluate non-gradient proxies, such as raw attention scores, for recovering causal structure. Prior work reports mixed evidence: some positive results [Rohekar et al., 2023, Lu et al., 2023], but many studies find that attention weights alone are noisy and do not reliably capture token relationships [Jain and Wallace, 2019, Achtibat et al., 2024]. In our experiments, attention scores help only in the shallow (single-layer) transformer. This aligns with findings in the large language model literature: as depth increases, repeated attention routing and residual MLP updates mix token representations. A single layer's attention matrix reflects intra-layer routing rather than the final output's functional dependence on the original inputs. Cross-layer composition entangles paths through value vectors, and marginalizing these paths makes raw attention a poor proxy for causal influence; many heads are also redundant or prunable. Empirically, prior work reports weak correlations between attention weights and counterfactual importance, and shows that faithfulness improves only when accounting for cross-layer attention flow or using gradient-/relevance-based methods [Jain and Wallace, 2019, Serrano and Smith, 2019, **?**, Michel et al., 2019, Abnar and Zuidema, 2020, Chefer et al., 2021, Achtibat et al., 2024].

In short, attention scores work substantially better in the shallow transformer than in the deep one, consistent with prior findings. Although the deep model learns the data-generating process well,

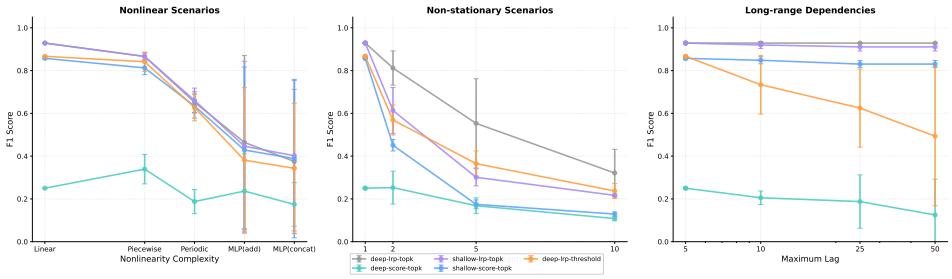

Figure 5: **Transformer variants performance comparison on challenging regimes. Left:** F1 scores on linear and nonlinear dynamics. **Middle:** F1 scores on non-stationary dynamics. **Right:** F1 scores on long-range dependencies.

attention alone does not reveal the causal structure. The mixing of token information across layers makes attention unreliable in either setting.

A.6.2 Effect of model depth

The depth of the transformer primarily affects capacity and the ability to capture complex long-term 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 slightly higher accuracy in recovering causal structure and show clear advantages on non-stationary dynamics. This highlights the potential of deep transformers for highly heterogeneous, long-range dynamics, echoing the success of pretrained large language and vision models.

A.6.3 Effect of sample size scaling

In nonlinear and non-stationary settings, we study how sample size affects causal discovery. As the sample size increases, the deep transformer more accurately recovers regime-specific causal structure and implicitly learns to route gradients to the appropriate regime. This trend aligns with the zero-shot generalization observed in large pretrained transformers. Such models are promising causal discoverers when fine-tuned or used as foundation models. We expect cross-domain pretraining to further improve the modeling of stable, mechanistic dynamics.

A.6.4 Effect of graph binarization

Different binarization rules can lead to different causal graphs. We compare thresholding and top-k. Thresholding performs comparably 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).

A.7 More Discussions

A.7.1 The role of prediction objective.

While a Gaussian likelihood (MSE) is a convenient objective and matches a Gaussian noise prior, richer likelihoods can better fit complex, heteroskedastic, or multi-modal dynamics and thereby sharpen attribution quality. Promising directions include flow-matching and diffusion-based objectives, as well as quantile and energy-based losses; these can improve calibration of gradients in scarce and highly heterogeneous regimes. In parallel, scalable implicit sparsity regularization techniques may further stabilize edge selection without sacrificing scalability. Together, improved objectives and sparsity control directly affect the fidelity of recovered causal structure—and may offer causal insights for mitigating memory limits and hallucinations in foundation models by steering representations toward stable, mechanistic dependencies.

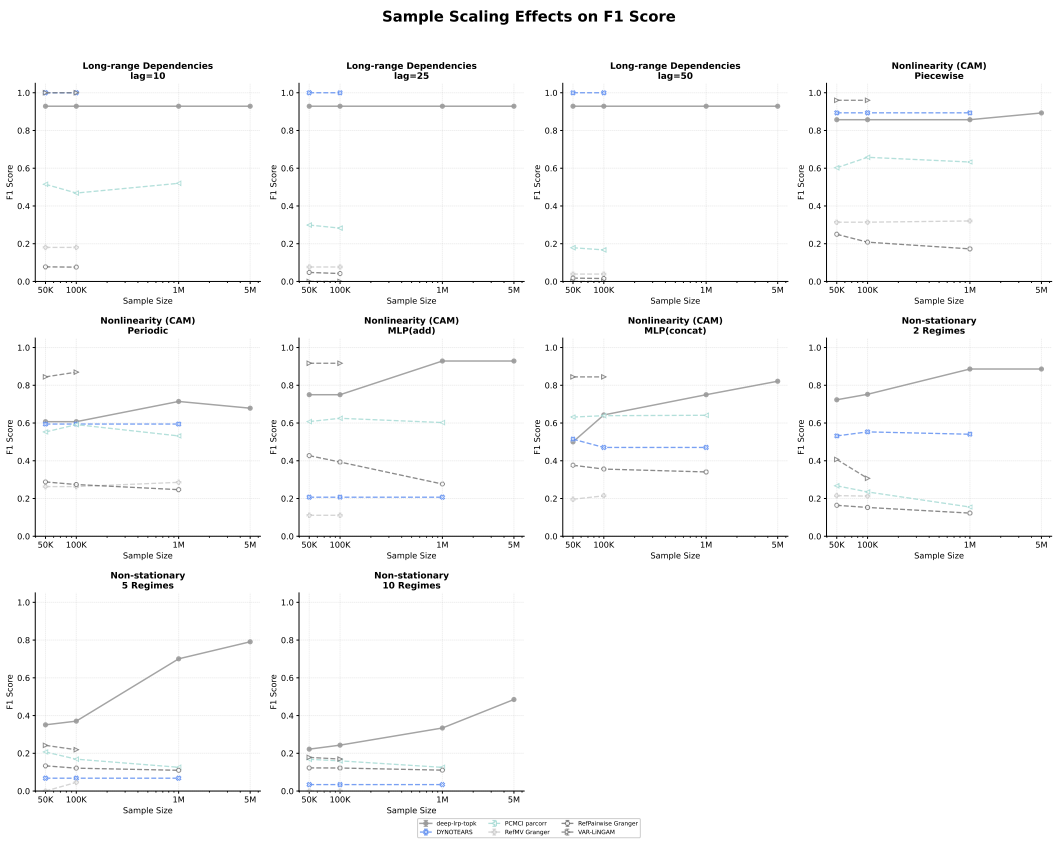

Figure 6: **Sample scaling effects on challenging regimes.** Sample scaling effects on three challenging regimes: nonlinear dynamics, long-term dependencies, and non-stationary dynamics. The transformer generally improves in these challenging scenarios with more data.

A.7.2 Limitations on small samples and latent confounders.

When data are scarce, attributions are noisy and edge selection becomes challenging in nonlinear and non-stationary settings (see Fig. 4). With latent confounders, the method can learn spurious links. It can be mitigated by methods that account for latent confounding; for example, one can post-process the learned Markov blanket using FCI [Spirtes et al., 2000]. Complementary remedies include explicit latent modeling and stronger structural priors within the forecasting-to-discovery pipeline.

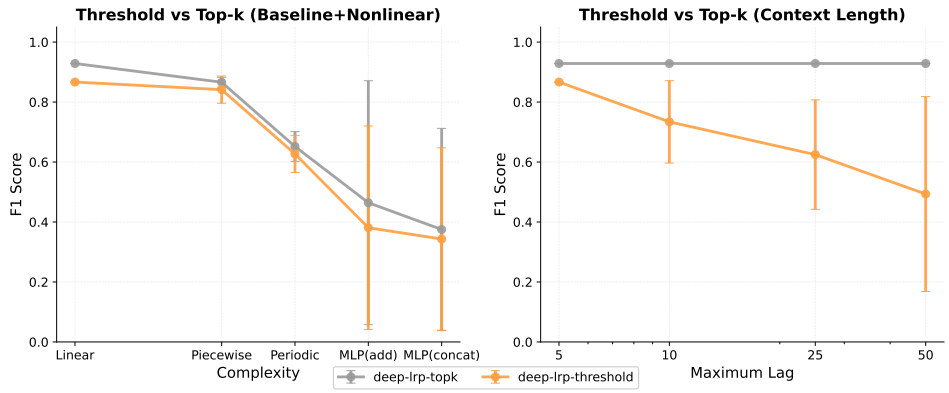

Figure 7: **Graph binarization methods comparison. Left:** F1 scores comparing threshold vs. top-k binarization across different complexity levels for baseline and nonlinear settings. **Right:** Performance degradation of threshold method vs. stable top-k performance as maximum lag increases, showing top-k's robustness to longer context lengths.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction align with our scope: theoretical identifiability for decoder-only transformers and empirical validation; see Section 2, Section 3 and supplementary results in Appendix.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss small-sample and latent-confounding limitations and mitigations in Appendix A ("Limitations on small samples and latent confounders").

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Assumptions A1–A4 and Theorem 1 are stated in §2.1; detailed lemmas and proofs are provided in Appendix ("Identifiability of the causal structure").

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe simulator settings, attribution/binarization, and evaluation protocols (Appendix A), as well as the model architecture and optimization objective in Appendix A.4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may

be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Currently we do not provide open access to the data and code, but we plan to release them with instructions in the supplemental material in the camera-ready version.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

Justification: We include the simulator design, candidate functions, noise types, regime setup in Appendix A.3, as well as the model architecture and optimization objective in Appendix A.4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report overall standard deviation and the error bars across different sample sizes given certain experimental conditions.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We report compute resources for our approach and baselines in Appendix A.5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The work uses only synthetic data and public baselines, involves no human subjects or sensitive data, and adheres to the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: There is no societal impact of the work performed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not release high-risk models or scraped datasets; the paper uses simulated data and a standard methodology.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all external assets and respect the license and terms of use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not introduce new datasets, models, or code artifacts in this submission.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs are not an important or non-standard component of the core methods in this research.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.