INTERVENTIONAL TIME SERIES PRIORS FOR CAUSAL FOUNDATION MODELS

Dennis Thumm
National University of Singapore
`dennis.thumm@u.nus.edu`

Ying Chen
Department of Mathematics
Centre for Quantitative Finance
Risk Management Institute
National University of Singapore
`matcheny@nus.edu.sg`

ABSTRACT

Prior-data fitted networks (PFNs) have emerged as powerful foundation models for tabular causal inference, yet their extension to time series remains limited by the absence of synthetic data generators that provide interventional targets. Existing time series benchmarks generate observational data with ground-truth causal graphs but lack the interventional data required for training causal foundation models. To address this, we propose **CausalTimePrior**, a principled framework for generating synthetic temporal structural causal models (TSCMs) with paired observational and interventional time series. Our prior supports configurable causal graph structures, nonlinear autoregressive mechanisms, regime-switching dynamics, and multiple intervention types (hard, soft, time-varying). We demonstrate that PFNs trained on CausalTimePrior can perform in-context causal effect estimation on held-out TSCMs, establishing a pathway toward foundation models for time series causal inference.

Track: Research

1 INTRODUCTION

Foundation models have transformed machine learning by enabling test-time inference without task-specific training. In tabular domains, prior-data fitted networks (PFNs) achieve this by pre-training transformers on synthetic datasets sampled from a prior distribution over data-generating processes (Müller et al., 2022; Hollmann et al., 2023). Recent work extends PFNs to causal inference: Do-PFN (Robertson et al., 2025) and CausalFM (Ma et al., 2025) train on synthetic structural causal models (SCMs) (Pearl, 2009) with interventional data, enabling in-context estimation of treatment effects from observational data alone.

However, extending causal PFNs to time series faces a fundamental obstacle: the lack of suitable synthetic data generators. While benchmarks like CausalTime (Cheng et al., 2024), TimeGraph (Ferdous et al., 2025), and CauseMe (Runge et al., 2019) provide time series with ground-truth causal graphs, they generate only *observational* data. Without interventional targets, one cannot train models to predict outcomes under interventions—the core task of causal inference.

We address this gap with **CausalTimePrior**[1], a framework for sampling temporal SCMs (TSCMs) together with paired observational and interventional time series (see Table 1). Our contributions are (1) a **practical prior over discrete-time dynamic SCMs** (Boeken & Mooij, 2024) that generates paired observational and interventional time series for training causal foundation models, (2) support for **regime-switching TSCMs** with Markov-driven structural breaks and interventional data—the first generator combining regime-switching dynamics (Balsells-Rodas et al., 2024) with interventional time series generation, and (3) **preliminary experiments** demonstrating that PFNs trained on CausalTimePrior can predict interventional outcomes given only observational context.

---

[1] `https://github.com/thummd/CausalTimePrior`

Table 1: Comparison of time series causal data generators. CausalTimePrior is the first to support regime-switching dynamics (changing causal structures over time).

Generator	Interventions	Nonlinear SCMs	Time-varying Interventions	Regime-switch
CausalTime	✗	✓	—	✗
TimeGraph	✗	✓	—	✗
CAnDOIT	✓ Hard	✓	✗	✗
TECDI	✓ Soft	✗	✗	✗
CaTSG	✓ do-calculus	✓	✗	✗
CausalTimePrior	✓ All	✓	✓	✓

2 RELATED WORK

Time series causal discovery benchmarks. CausalTime (Cheng et al., 2024) fits neural networks to real observations and extracts causal graphs via importance analysis, producing realistic data with known ground truth. TimeGraph (Ferdous et al., 2025) generates synthetic time series from linear and nonlinear SCMs with configurable graph properties. CauseMe (Runge et al., 2019) benchmarks (including Lorenz-96 systems) and CausalRivers (Stein et al., 2025), the largest real-world benchmark (1,160 stations), offer physical ground-truth graphs. In contrast to CausalTimePrior, these existing methods are limited to observational data without interventions.

Generators with interventional support. We identified only three frameworks supporting temporal interventions, each with limitations. **CAnDOIT** (Castri et al., 2024) generates time-lagged SCMs with hard interventions on *known, single-node* targets. It supports nonlinear mechanisms but only static intervention values. **TECDI/RealTCD** (Li et al., 2023; 2024) use structural VAR models with soft (TECDI) or hard (RealTCD) interventions. RealTCD handles unknown targets but is limited to linear mechanisms. **CaTSG** (Xia et al., 2025) implements do-calculus (Pearl et al., 2016) via diffusion models with a causal score function. While CaTSG implements interventions via learned diffusion models requiring training on specific datasets, CausalTimePrior generates interventional data analytically from explicit structural equations, enabling fast prior sampling without a separate generative model.

Causal PFNs for tabular data. Do-PFN (Robertson et al., 2025) pre-trains transformers on synthetic SCMs to predict conditional interventional distributions (CIDs) without knowing the causal graph. CausalFM (Ma et al., 2025) formalizes Bayesian priors over SCMs for back-door, front-door, and instrumental variable settings. Both demonstrate strong performance on i.i.d. tabular data but do not address temporal dependencies. Related work on counterfactual time series estimation includes CRN (Bica et al., 2020) and the Causal Transformer (Melnychuk et al., 2022), which estimate individualized treatment effects over time but require per-dataset training rather than in-context learning.

Time series foundation models. Recent work has explored zero-shot forecasting via synthetic pre-training. ForecastPFN (Dooley et al., 2023) and TimePFN (Taga et al., 2025) pre-train transformers on synthetic data for multivariate forecasting. TempoPFN (Moroshan et al., 2025) pre-trains linear Recurrent Neural Networks (RNNs) for univariate forecasting, using diverse synthetic generators including Stochastic Differential Equations (SDEs), Gaussian processes, and causal kernels (CauKer) (Xie et al., 2025). While CauKer generator produces multivariate SCM-based time series, it lacks temporal lag structures and intervention support. These models target prediction rather than causal inference.

Regime-switching dynamics. Switching Dynamical Systems (SDSs) with Markov Switching Models provide identifiability theory for regime-dependent causal discovery (Balsells-Rodas et al., 2024), but focus on inference from observational data without generating interventional datasets for training foundation models.

3 CAUSALTIMEPRIOR

We define a prior Π over temporal structural causal models that generates paired observational and interventional time series suitable for training causal foundation models (see Algorithm 1).

3.1 TEMPORAL STRUCTURAL CAUSAL MODELS

Following the Dynamic Structural Causal Model (DSCM) framework (Boeken & Mooij, 2024), we consider the discrete-time acyclic case. A temporal SCM (TSCM) $\mathcal{S} = (\mathcal{G}, \mathbf{F}, P_\epsilon)$ consists of:

- A **time-lagged DAG** $\mathcal{G} = (G_0, G_1, \ldots, G_K)$ where $G_0 \in \{0,1\}^{N \times N}$ encodes instantaneous (intra-slice) edges and G_k encodes edges from time $t - k$ to t for lags $k \in \{1, \ldots, K\}$.

- **Structural equations** $\mathbf{F} = \{f_i\}_{i=1}^N$ where:

$$X_t^{(i)} = f_i \left(\text{Pa}_{\mathcal{G}}(X_t^{(i)}) \right) + \epsilon_t^{(i)}, \quad \epsilon_t^{(i)} \sim P_\epsilon^{(i)} \tag{1}$$

  with parents $\text{Pa}_{\mathcal{G}}(X_t^{(i)}) = \{X_{t-k}^{(j)} : G_k[j,i] = 1, k \in \{0, \ldots, K\}\}$.

3.2 PRIOR DISTRIBUTION OVER TSCMS

Graph prior $\Pi_{\mathcal{G}}$. We sample the number of variables $N \sim \text{Uniform}(3, N_{\max})$, maximum lag $K \sim \text{Uniform}(1, K_{\max})$, and edge probability $p \sim \text{Beta}(\alpha, \beta)$. Instantaneous edges G_0 are sampled from an Erdős-Rényi model (Erdős & Rényi, 1960) with acyclicity enforced via topological ordering. Lagged edges G_k are sampled independently with probability decaying as $p \cdot \gamma^k$ for persistence factor $\gamma \in (0, 1]$.

Mechanism prior $\Pi_{\mathbf{F}}$. We sample mechanisms from multiple families. For simple mechanisms:

$$f_i(\mathbf{x}) = \sum_{j \in \text{Pa}(i)} w_{ij} \cdot \phi_{ij}(x_j) + b_i \tag{2}$$

where weights $w_{ij} \sim \mathcal{N}(0, \sigma_w^2)$, biases $b_i \sim \mathcal{N}(0, \sigma_b^2)$, and ϕ_{ij} is sampled uniformly from $\{\text{id}, \sin, \cos, \tanh, |\cdot|, (\cdot)^2, \exp(-|\cdot|)\}$. The diversity of activation functions ensures the prior covers a wide range of nonlinear temporal dynamics.

Noise prior Π_ϵ. Noise distributions are sampled per variable from $\{\mathcal{N}(0, \sigma^2), \text{Uniform}(-a, a), \text{Laplace}(0, b)\}$ with scale parameters drawn from suitable hyperpriors.

3.3 INTERVENTION TYPES

Given a sampled TSCM $\mathcal{S}$, we generate interventional data by modifying structural equations (Eberhardt & Scheines, 2007). Let $I \subseteq \{1, \ldots, N\}$ denote intervention targets and $t_I \subseteq \{1, \ldots, T\}$ the intervention times. We provide examples for each intervention type in Appendix C.

Hard interventions. (do-operator) Replace $X_t^{(i)} := c$ for $i \in I, t \in t_I$, severing incoming edges.

Soft interventions. Perturb the mechanism: $X_t^{(i)} = f_i(\text{Pa}(X_t^{(i)})) + \delta_i + \epsilon_t^{(i)}$ for shift $\delta_i \sim \mathcal{N}(\mu_\delta, \sigma_\delta^2)$.

Time-varying interventions. Set $X_t^{(i)} := c(t)$ where $c(t)$ follows a specified profile (step, ramp, sinusoidal, or sampled trajectory) (Hernán & Robins, 2020).

3.4 DATA GENERATION PIPELINE

For each training example, we:

1. Sample $\mathcal{S} \sim \Pi$ (graph, mechanisms, noise).

2. Sample intervention specification: targets I, times t_I, type, and values.

3. Generate **observational** series $\mathbf{X}_{1:T}^{\text{obs}}$ by forward simulation.

4. Generate **interventional** series $\mathbf{X}_{1:T}^{\text{int}}$ under $\text{do}(X_{t_I}^{(I)} = c)$.

5. Form training tuple: $(\mathbf{X}_{1:T}^{\text{obs}}, I, t_I, c, Y_\tau^{\text{int}})$ where Y_τ^{int} is the outcome variable at target time τ under intervention.

3.5 REGIME-SWITCHING PRIORS

Real-world time series often exhibit structural breaks where causal relationships change (Rahmani et al., 2025). We extend our prior to regime-switching TSCMs, following the Markov Switching Model framework (Balsells-Rodas et al., 2024):

$$X_t^{(i)} = f_i^{(d_t)} \left(\text{Pa}_{\mathcal{G}^{(d_t)}}(X_t^{(i)}) \right) + \epsilon_t^{(i)}, \quad d_t \sim \text{Markov}(\mathbf{P}) \tag{3}$$

where $d_t \in \{1, \ldots, R\}$ indexes the active regime with transition matrix $\mathbf{P}$. Each regime has its own causal graph $\mathcal{G}^{(r)}$ and mechanisms $\mathbf{F}^{(r)}$. Regime transitions follow a sticky Markov chain ($P_{ii} \approx 0.9$) to model persistent causal structures. In our experiments, 15% of training TSCMs are regime-switching with $R \in \{2, 3\}$ regimes. Combined with interventional data generation, this enables training PFNs that can reason about interventions under time-varying causal structures.

3.6 DESIGN CHOICES AND ASSUMPTIONS

Complexity, identifiability, and causality. Adding lags increases model complexity *linearly*, not exponentially: each lag k contributes $O(N)$ weight parameters per node, and edge density decays geometrically as $p_k = p \cdot \gamma^k$ ($\gamma = 0.7$), so distant lags are increasingly sparse. Identifiability holds per time step: within each t, variables are evaluated in topological order of G_0, so instantaneous parents are already computed, while lagged parents $X_{t-k}^{(j)}$ are fixed from prior steps. This naturally resolves instantaneous and temporal causality through a single forward simulation—G_0 captures same-timestep effects (akin to imputation), while $G_1, \ldots, G_K$ capture lagged effects (akin to forecasting), with temporal ordering preventing cycles by construction.

Noise assumptions and prediction horizon. Noise is *additive* and *exogenous*: $\epsilon_t^{(i)}$ is added after the nonlinear activation and sampled independently of parent values. It is also *Markovian*—i.i.d. across time—implying that the causal graph fully mediates all temporal dependencies; extending to non-Markovian (temporally correlated) noise would model latent confounders persisting across time and is an important future direction. At training time, the PFN queries outcomes 0–5 steps after intervention onset (70% downstream queries at 1–5 steps, 30% instantaneous), concentrating learning on the short-range causal propagation window; see Appendix B for further discussion.

4 EXPERIMENTS

Prior Validation. We validate CausalTimePrior by analyzing a dataset of 100K generated TSCMs. (1) **Structural diversity**: the prior generates TSCMs with $N \in [3, 10]$ variables, $K \in [1, 3]$ lags, and $T = 50$ time steps, including 70% diverse nonlinear TSCMs, 15% chain structures, and 15% regime-switching TSCMs. Erdős-Rényi sampling with varied edge probabilities implicitly produces canonical causal motifs (confounders, mediators, colliders) as subgraphs. (2) **Stability**: 0% divergence rate (no NaN/Inf values) across all 100K samples, achieved through value clipping and careful mechanism parameterization. (3) **Intervention coverage**: hard, soft, and time-varying interventions with mean effect size of 17.98 (std 53.93), demonstrating substantial variability in intervention magnitudes across types. (4) **Paired data quality**: observational and interventional series maintain

similar statistics (obs: $\mu = 46.78$, $\sigma = 242.85$; int: $\mu = 41.56$, $\sigma = 228.52$), confirming that interventions produce realistic counterfactual outcomes rather than out-of-distribution artifacts. Example visualization and full distributions of prior properties are shown in Appendix D and E.

Proof-of-Concept PFN. As a preliminary demonstration, we train a simple 2-layer GRU-based PFN (128 hidden dim, 11 min on CPU) on 100K TSCMs from CausalTimePrior and evaluate on 1,000 held-out TSCMs. The model learns to distinguish causal from non-causal queries: Pred/GT ratio of 0.95 for intervened queries vs. 0.46 for non-causal queries (Table 2), and achieves comparable RMSE to per-dataset VAR baselines without per-sample fitting (Table 3). Implementation details, full results, baselines, a shuffled-intervention control experiment, ablations, generalizations, and an example are in Appendix F, G, H, I, and J.

5 CONCLUSION

CausalTimePrior addresses a critical gap in time series causal inference: the absence of synthetic generators with interventional data for training foundation models. By combining diverse temporal generators with principled intervention logic, it yields a prior over TSCMs with diverse intervention types. Our preliminary results suggest that PFNs trained on this prior can perform in-context causal effect estimation, opening a pathway toward foundation models for time series causality.

Limitations and future work. The framework currently assumes Markovian noise and discrete-time dynamics; extensions to non-Markovian confounding and continuous-time processes are important future directions. Our Erdős-Rényi graph prior implicitly covers canonical causal structures (confounders, mediators, colliders) but does not explicitly stratify over them as Do-PFN does for tabular settings; adding structured temporal motifs could improve coverage. The prior has not been validated against real-world causal time series distributions. We plan to (1) scale training to larger models with explicit canonical structure sampling, (2) incorporate continuous-time dynamics, and (3) benchmark on semi-synthetic datasets derived from real observational data.

REFERENCES

Carles Balsells-Rodas, Yixin Wang, and Yingzhen Li. On the identifiability of switching dynamical systems. In *International Conference on Machine Learning*, pp. 2639–2672. PMLR, 2024.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 41–48, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

Ioana Bica, Ahmed M Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*, 2020.

Philip Boeken and Joris M. Mooij. Dynamic structural causal models. In *UAI 2024 Workshop on Causal Inference for Time Series (CI4TS)*, 2024.

Luca Castri, Sariah Mghames, Marc Hanheide, and Nicola Bellotto. CAnDOIT: Causal discovery with observational and interventional data from time series. *Advanced Intelligent Systems*, 2024.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Yuxiao Cheng, Ziqian Yang, Xu Chen, Jiecheng Li, and Junchi Yan. CausalTime: Realistically generated time-series for benchmarking of causal discovery. In *International Conference on Learning Representations*, 2024.

Samuel Dooley, Gurnoor Singh Khurana, Chirag Mohapatra, Siddartha V Naidu, and Colin White. Forecastpfn: Synthetically-trained zero-shot forecasting. *Advances in Neural Information Processing Systems*, 36:2403–2426, 2023.

Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of science*, 74(5):981–995, 2007.

Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publicationes Mathematicae*, 5: 17–61, 1960.

Muhammad Hasan Ferdous, Emam Hossain, and Md Osman Gani. Timegraph: Synthetic benchmark datasets for robust time-series causal discovery. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5425–5435, 2025.

Miguel A Hernán and James M Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020. ISBN 9780367583421. URL https://miguelhernan.org/whatifbook/.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations*, 2023.

Patrick Kidger, James Foster, Xuechen Li, and Terry Lyons. Neural SDEs as infinite-dimensional GANs. In *International Conference on Machine Learning*, pp. 5453–5463. PMLR, 2021.

Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer-Verlag, Berlin, 1992.

Peiwen Li, Yuan Meng, Xin Wang, Fang Shen, Yue Li, Jialong Wang, and Wenwu Zhu. Causal discovery in temporal domain from interventional data. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 1306–1315, 2023.

Peiwen Li, Xin Wang, Zeyang Zhang, Yuan Meng, Fang Shen, Yue Li, Jialong Wang, Yang Li, and Wenwu Zhu. Realtcd: Temporal causal discovery from interventional data with large language model. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 4669–4677, 2024.

Yuchen Ma, Dennis Frauen, Emil Javurek, and Stefan Feuerriegel. Foundation models for causal inference via prior-data fitted networks. *arXiv preprint arXiv:2506.10914*, 2025.

Georg Manten, Cecilia Casolo, Emilio Ferrucci, Søren Wengel Mogensen, Cristopher Salvi, and Niki Kilbertus. Signature kernel conditional independence tests in causal discovery for stochastic processes. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Nx4PMtJ1ER.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *International Conference on Machine Learning*, pp. 15293–15329. PMLR, 2022.

Vladyslav Moroshan, Julien Siems, Arber Zela, Timur Carstensen, and Frank Hutter. TempoPFN: Synthetic pre-training of linear RNNs for zero-shot time series forecasting. In *NeurIPS 2025 Workshop on AI for Tabular Data*, 2025.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do Bayesian inference. In *International Conference on Learning Representations*, 2022.

Judea Pearl. *Causality: Models, reasoning, and inference*. Cambridge University Press, 2009.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 2014.

Hamed Rahmani et al. FANTOM: Temporal causal discovery with regime-switching normalizing flows. In *International Conference on Learning Representations*, 2025.

Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. Do-pfn: In-context learning for causal effect estimation. In *1st ICML Workshop on Foundation Models for Structured Data*, 2025.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pp. 1388–1397, 2020.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in Earth system sciences. *Nature Communications*, 10(1):2553, 2019.

Gideon Stein, Maha Shadaydeh, Jan Blunk, Niklas Penzel, and Joachim Denzler. Causalrivers-scaling up benchmarking of causal discovery for real-world time-series. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ege Onur Taga, Muhammed Emrullah Ildiz, and Samet Oymak. Timepfn: Effective multivariate time series forecasting with synthetic data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20761–20769, 2025.

Dennis Thumm and Luis Ontaneda Mijares. Towards causal market simulators. In *ICAIF 2025 Workshop on Rethinking Financial Time-Series*, Singapore, 2025. URL `https://icaif-25-rtfs.github.io/`.

Yutong Xia, Chang Xu, Yuxuan Liang, Qingsong Wen, Roger Zimmermann, and Jiang Bian. Causal time series generation via diffusion models. *arXiv preprint arXiv:2509.20846*, 2025.

Shifeng Xie, Vasilii Feofanov, Marius Alonso, Ambroise Odonnat, Jianfeng Zhang, Themis Palpanas, and Ievgen Redko. Cauker: classification time series foundation models can be pretrained on synthetic data only. *CoRR*, abs/2508.02879, August 2025. URL `https://doi.org/10.48550/arXiv.2508.02879`.

A   CAUSALTIMEPRIOR ALGORITHM

Algorithm 1 formalizes the CausalTimePrior sampling procedure for generating paired observational and interventional time series from TSCMs.

Algorithm 1 CausalTimePrior Sampling

1: **Input:** Prior hyperparameters $\Pi = (\Pi_\mathcal{G}, \Pi_\mathbf{F}, \Pi_\epsilon)$, sequence length T
2: **Output:** Observational series $\mathbf{X}_{1:T}^{\text{obs}}$, interventional series $\mathbf{X}_{1:T}^{\text{int}}$, intervention spec I, t_I, c
3:
4: // Sample TSCM
5: $N \sim \text{Uniform}(3, N_{\max})$, $K \sim \text{Uniform}(1, K_{\max})$
6: $p \sim \text{Beta}(\alpha, \beta)$
7: Sample $G_0 \in \{0, 1\}^{N \times N}$ (acyclic via topological ordering)
8: **for** $k = 1$ to K **do**
9:     Sample G_k with edge probability $p \cdot \gamma^k$
10: **end for**
11: $\mathcal{G} \leftarrow (G_0, G_1, \ldots, G_K)$
12:
13: **for** $i = 1$ to N **do**
14:     Sample mechanism $f_i \sim \Pi_\mathbf{F}$ (nonlinear autoregressive)
15:     Sample noise $P_\epsilon^{(i)} \sim \Pi_\epsilon$
16: **end for**
17: $\mathcal{S} \leftarrow (\mathcal{G}, \{f_i\}_{i=1}^N, P_\epsilon)$
18:
19: // Sample intervention specification
20: Sample targets $I \subseteq \{1, \ldots, N\}$
21: Sample times $t_I \subseteq \{1, \ldots, T\}$
22: Sample type $\in \{\text{hard}, \text{soft}, \text{time-varying}\}$
23: Sample value(s) c or $c(t)$
24:
25: // Generate paired time series
26: $\mathbf{X}_{1:T}^{\text{obs}} \leftarrow$ Forward simulation of $\mathcal{S}$
27: $\mathbf{X}_{1:T}^{\text{int}} \leftarrow$ Forward simulation of $\mathcal{S}$ under $\text{do}(X_{t_I}^{(I)} = c)$
28:
29: **return** $\mathbf{X}_{1:T}^{\text{obs}}, \mathbf{X}_{1:T}^{\text{int}}, (I, t_I, c)$

Continuous-time extension. CausalTimePrior currently generates discrete-time SCMs (Algorithm 1), but our autoregressive mechanisms have a natural continuous-time interpretation via the Euler-Maruyama discretization (Kloeden & Platen, 1992). Recent work on causal discovery in continuous-time SDEs (Manten et al., 2025) motivates extending CausalTimePrior to generate interventional data from SDE-based causal models. Consider a causal Ornstein-Uhlenbeck process $dX_t = \theta(\mu - X_t)\,dt + \sigma_w\,dW_t$; applying Euler-Maruyama with step Δt yields:

$$x_{t+1} = \underbrace{(1 - \theta\Delta t)}_{c_2}\, x_t + \underbrace{\theta\mu\Delta t}_{c_1} + \underbrace{\sigma_w\sqrt{\Delta t}}_{c_3}\, Z_t, \quad Z_t \sim \mathcal{N}(0, 1) \tag{4}$$

which is precisely the AR(1) form our mechanism prior generates. This means each sampled discrete-time SCM can be viewed as an Euler-Maruyama discretization of a continuous-time causal SDE system (Thumm & Mijares, 2025). A natural extension is to sample continuous-time mechanisms directly—e.g., via Neural ODEs (Chen et al., 2018) or Neural SDEs (Kidger et al., 2021)—and discretize at variable time steps, enabling the prior to generate irregularly-sampled interventional time series.

B   PRIOR ASSUMPTIONS AND LIMITATIONS

This section details the modeling assumptions underlying CausalTimePrior and their implications for identifiability and generalization.

Effect propagation and decay. Causal effects in the prior decay through three mechanisms: (1) the maximum lag $K \in \{1, 2, 3\}$ bounds direct causal influence to at most 3 steps back, (2) geometric edge decay ($p_k = p \cdot 0.7^k$) makes distant lag connections increasingly sparse, and (3) bounded activations (e.g., $\tanh$, $\sin$) naturally dampen signal propagation. There is no explicit spectral radius constraint on the weight matrices; instead, stability is enforced empirically via value clipping to $[-1000, 1000]$ at each simulation step and divergence rejection (any trajectory with $|x| > 10^6$ or NaN/Inf is discarded and resampled). A 50-step burn-in period is discarded to approximate the stationary distribution.

This design implies that causal effects attenuate rapidly—typically within 1–5 steps after intervention onset. While this is appropriate for systems with short-range temporal dependencies, it may underweight long-range causal effects (e.g., economic policy impacts over months). A PFN trained on this prior would inherit this short-range bias.

Markovian noise and identifiability. The noise $\epsilon_t^{(i)}$ is sampled i.i.d. at each time step— independent across both time and variables. This Markovian assumption implies that the causal graph $\mathcal{G}$ fully mediates all temporal dependencies: there are no hidden common causes acting across time beyond what the lagged edges encode. Identifiability within the modeled lag window follows from the additive noise model (ANM) structure (Peters et al., 2014): each mechanism has the form $\sigma(\sum w_j x_j + b) + \epsilon$, and the use of non-Gaussian noise families (Uniform, Laplace) for non-root nodes supports distinguishing causal from non-causal associations under standard ANM identifiability results.

An extension to *non-Markovian confounding*—where $\text{Cov}(\epsilon_t^{(i)}, \epsilon_{t-s}^{(j)}) \neq 0$ for some $s > 0$—would model latent confounders that persist over time (e.g., unobserved trends or regime states not captured by the graph). This would break the assumption that interventional distributions depend only on the explicit causal structure and is an important direction for increasing the prior's realism.

Interventional vs. counterfactual semantics. CausalTimePrior generates *interventional* paired data, not *counterfactual* paired data. Concretely, when `generate_pair` produces an observational series $\mathbf{X}^{\text{obs}}$ and an interventional series $\mathbf{X}^{\text{int}}$, the two simulations draw **independent noise realizations** $\{\epsilon_t^{(i)}\}$ and $\{\tilde{\epsilon}_t^{(i)}\}$. This means the trajectories may differ even *outside* the intervention window—not because of any causal effect, but because they are driven by different exogenous noise.

This corresponds to the **interventional** query $P(\mathbf{X} \mid \text{do}(X_{t_I}^{(i)} = c))$: "what would a *new* draw from the system look like under this intervention?" The alternative is the **counterfactual** query $P(\mathbf{X}^{CF} \mid \mathbf{X}^{\text{obs}}, \text{do}(X_{t_I}^{(i)} = c))$: "given *this specific* observational realization, what *would have happened* if we had intervened?" Counterfactual inference requires the three-step abduction-action-prediction procedure (Pearl, 2009): (1) infer the noise $\epsilon = \mathbf{X}^{\text{obs}} - f(\text{Pa})$ (abduction), (2) modify the structural equation for the intervened variable (action), and (3) re-simulate with the *same* noise (prediction).

In our framework, generating counterfactual pairs would require sharing the pre-sampled noise tensor between the observational and interventional simulation runs, so that the two trajectories are identical before the intervention onset and diverge only through causal propagation of the intervention effect. This is a natural extension that would enable training PFNs for counterfactual estimation—a strictly harder task than interventional prediction, since it requires reasoning about unit-level rather than population-level effects.

Noise model details. All noise is additive (added after the nonlinear activation) and exogenous (independent of parent values). Root nodes (no parents in G_0) receive Gaussian noise with std $\sim$ ShiftedExp(rate $= 1.0$, shift $= 0.1$), providing larger driving noise. Non-root nodes receive noise from one of three variance-matched families (each with probability $\frac{1}{3}$): Gaussian $\mathcal{N}(0, \text{std}^2)$, Uniform $U(-a, a)$ with $a = \text{std}\sqrt{3}$, or Laplace $\text{Lap}(0, b)$ with $b = \text{std}/\sqrt{2}$, where std $\sim$ ShiftedExp(rate $= 10.0$, shift $= 0.01$). The smaller non-root noise ensures that most of a variable's variance comes from its causal parents rather than its own noise term.

Stationarity scope. Within a single SCM instance, mechanisms are stationary: weights, biases, activation functions, and noise distributions are fixed across all time steps. Non-stationarity en-

ters only through regime-switching SCMs (15% of samples), where 2–3 regimes—each with its own causal graph and mechanisms—alternate according to a sticky Markov chain. The current framework does not support continuously time-varying coefficients, one-time structural breaks, or heteroscedastic noise. These extensions would broaden the prior's coverage of real-world non-stationary dynamics.

PFN prediction horizon. During training, 70% of queries target a different variable 1–5 steps after intervention onset (downstream causal effects), while 30% query the intervention target itself at the intervention time (instantaneous effects). The PFN architecture accepts arbitrary (intervention_time, query_time) pairs as normalized inputs and can in principle predict at any horizon, but accuracy is expected to degrade beyond the 0–5 step training concentration. Extending the training distribution to longer horizons—potentially with curriculum learning (Bengio et al., 2009)—could improve long-range causal effect estimation.

C   INTERVENTION TYPE EXAMPLES

Figure 1 illustrates the three intervention types supported by CausalTimePrior, applied to the same 10-variable TSCM. Each row shows one type: (1) **Hard intervention** ($\mathrm{do}(X_0 := 3.0)$) replaces the variable's mechanism entirely with a constant, severing all incoming causal edges during the intervention window—visible as a flat interventional trajectory. (2) **Soft intervention** ($\delta = 1.5$) adds an additive shift to the mechanism output while preserving parental influence—the interventional trajectory tracks the observational one but is displaced. (3) **Time-varying intervention** ($\mathrm{do}(X_0 := 2\sin(2\pi t/L))$) replaces the mechanism with a time-dependent function, producing a sinusoidal interventional trajectory.

Figure 2 shows the four time-varying intervention profiles in detail, each applied to the same SCM and variable. The *step* profile jumps from -2 to $+2$ at the midpoint; the *ramp* linearly interpolates from -2 to $+2$; the *sinusoidal* profile follows $c(t) = 2\sin(2\pi t/L)$; and the *sampled trajectory* draws independent $c(t) \sim \mathcal{N}(0, 4)$ at each time step. The dashed line shows the intervention signal $c(t)$ on the secondary axis.

D   EXAMPLE VISUALIZATIONS

Figure 3 shows an example of paired observational and interventional time series generated from CausalTimePrior. The hard intervention on Variable 4 between $t = 20$ and $t = 80$ causes a clear divergence between the observational (blue) and interventional (red) trajectories during and after the intervention period. Figure 4 displays all 6 variables in the sampled TSCM, with Variable 4 (the intervention target) highlighted. The propagation of intervention effects through the causal graph is visible in downstream variables, while non-causally connected variables remain unaffected.

E   PRIOR PROPERTY DISTRIBUTIONS

Figure 5 shows the distributions of key properties across 100K TSCMs sampled from CausalTimePrior. (a) Graph sizes are approximately uniform over $N \in [3, 10]$. (b) Intervention types are split 50% hard, 30% soft, and 20% time-varying. (c) Intervention effect magnitudes span several orders of magnitude (median 1.4) on a log scale, ensuring the prior covers both subtle and large causal effects. (d) Intervention start times are concentrated in the second half of the sequence to allow sufficient observational context. (e) Edge probabilities follow a Beta(2,5) prior (mean 0.29), producing mostly sparse graphs. (f) Intervention values are approximately Gaussian-distributed around zero.

F   IMPLEMENTATION DETAILS

We implement a simple proof-of-concept architecture using a 2-layer GRU encoder (128 hidden dim). The **temporal encoder** processes the observational time series $\mathbf{X}^{\mathrm{obs}}_{1:T}$, the **intervention encoder** embeds the intervention specification $\mathrm{do}(X_t^{(i)} = c)$ (which variable, when, what value), and
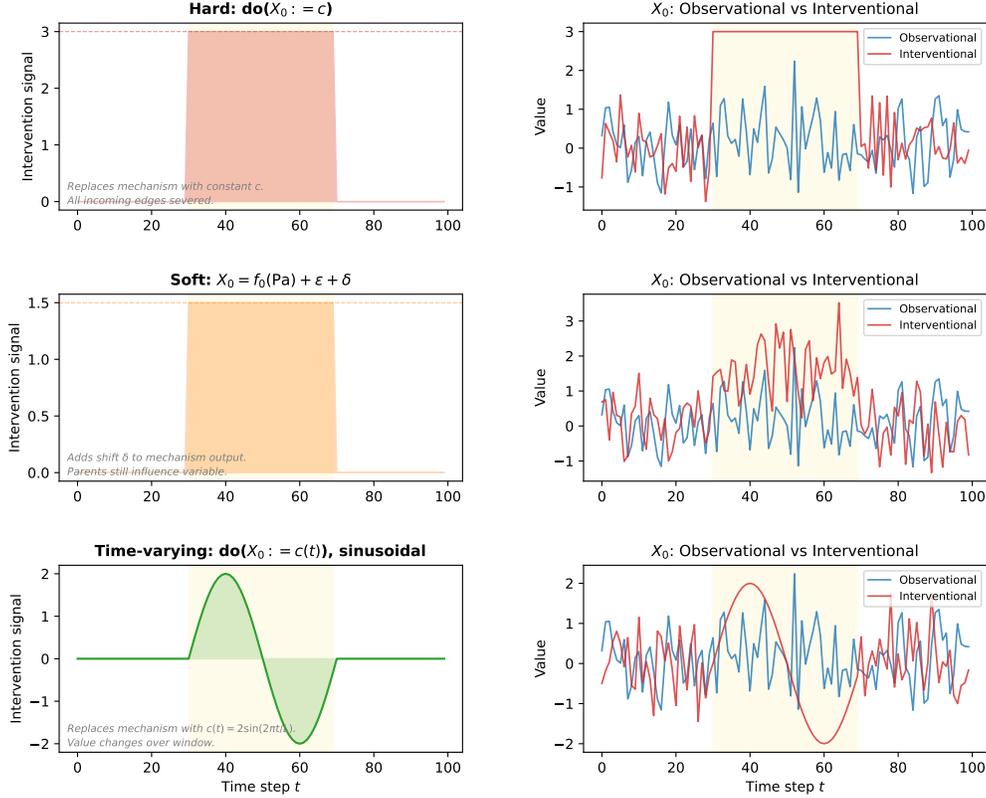
CausalTimePrior: Intervention Types

Figure 1: The three intervention types in CausalTimePrior. Left column: the intervention signal applied to X_0. Right column: resulting observational (blue) vs. interventional (red) trajectories. Yellow shading marks the intervention window $[30, 70)$.

the **query encoder** embeds the prediction query (which variable to predict, when). The **prediction head** combines these encodings to output $P(Y_\tau^{\text{int}}|\text{do}(X_t^{(i)} = c), \mathbf{X}_{1:T}^{\text{obs}})$ as a Gaussian distribution with predicted mean and standard deviation. The training objective is:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathcal{S} \sim \Pi} \mathbb{E}_{\mathbf{X}_{1:T}^{\text{obs}} \sim P_{\text{obs}}^{\mathcal{S}}} \mathbb{E}_{(x_t, y_\tau) \sim P_{\text{int}}^{\mathcal{S}}} \left[-\log q_\theta(y_\tau | \text{do}(x_t), \mathbf{X}_{1:T}^{\text{obs}}) \right] \tag{5}$$

where q_θ is a Gaussian with predicted mean and variance, $\mathbf{X}_{1:T}^{\text{obs}}$ is the observational time series context, $\text{do}(x_t)$ is the temporal intervention query at time t, and y_τ is the interventional outcome at target time τ.

Prior hyperparameters. $N_{\max} = 10$, $K_{\max} = 3$, $\alpha = 2$, $\beta = 5$ (sparse graphs), $\gamma = 0.7$ (lag decay), $\sigma_w = 1.0$, $\sigma_b = 0.5$.

Training. Adam optimizer, learning rate 10^{-4}, batch size 32, sequence length 50, 15 epochs on 100K TSCMs. Training takes approximately 11 minutes on CPU (Intel/AMD with AVX2). The model checkpoint requires ~330KB of storage.

Architecture. We use a simplified 2-layer GRU encoder with 128 hidden dimensions and Gaussian prediction head (mean + standard deviation). We chose a GRU over transformer architectures (as used in Do-PFN and TimePFN) for computational simplicity in this proof-of-concept: the GRU processes variable-length sequences efficiently and its recurrent state naturally captures temporal
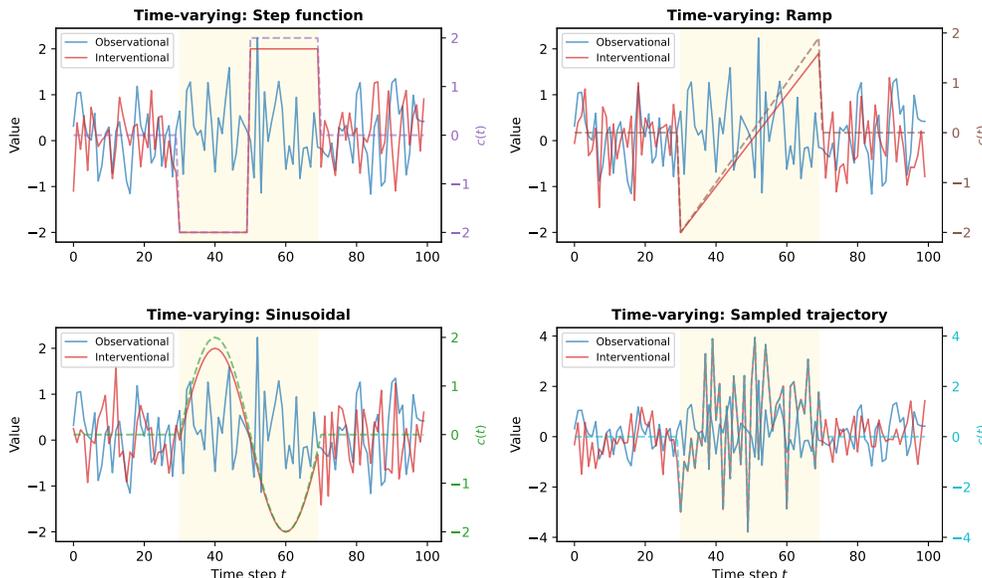
Figure 2: Four time-varying intervention profiles. Blue: observational trajectory. Red: interventional trajectory. Dashed: intervention signal $c(t)$ (right axis). All four profiles are applied to the same variable in the same TSCM.
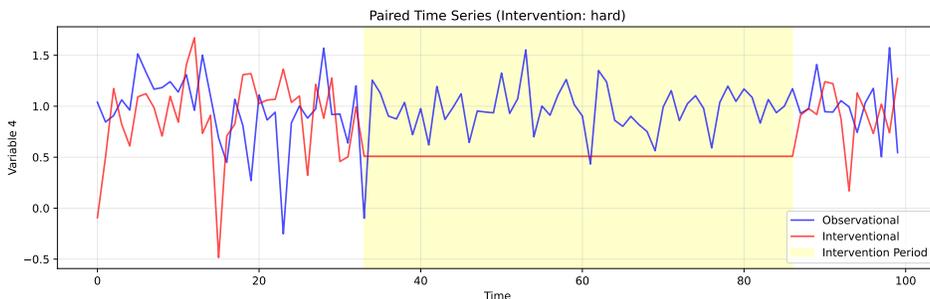

Figure 3: Paired observational and interventional time series for the intervention target variable. The yellow shaded region indicates the intervention period. The divergence between blue (observational) and red (interventional) trajectories demonstrates the causal effect of the intervention.

dependencies. This choice prioritizes fast iteration (∼11 min training on CPU) over architectural optimality; scaling to transformer or GatedDeltaProduct architectures (Moroshan et al., 2025) is an important next step.

Implementation. CausalTimePrior is implemented from scratch, drawing conceptual inspiration from TempoPFN's diverse generator design and intervention logic from Do-PFN. The core is a `TemporalSCM` class that supports both `sample_observational(T)` and `sample_interventional(T, intervention)` methods, enabling paired data generation from the same underlying causal structure with time-lagged dependencies and multiple intervention types.
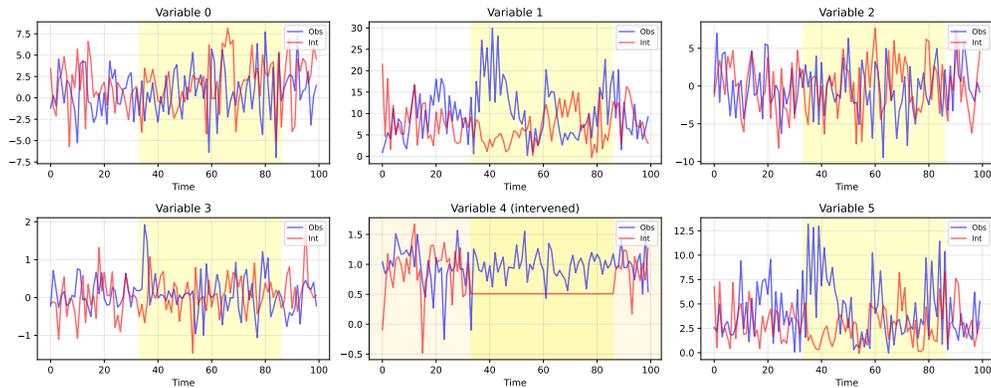
Figure 4: All variables in a sampled 6-variable TSCM with a hard intervention on Variable 4. The intervention target is highlighted with a yellow background. Causal effects propagate through the graph structure, affecting downstream variables while leaving non-causally connected variables unchanged.
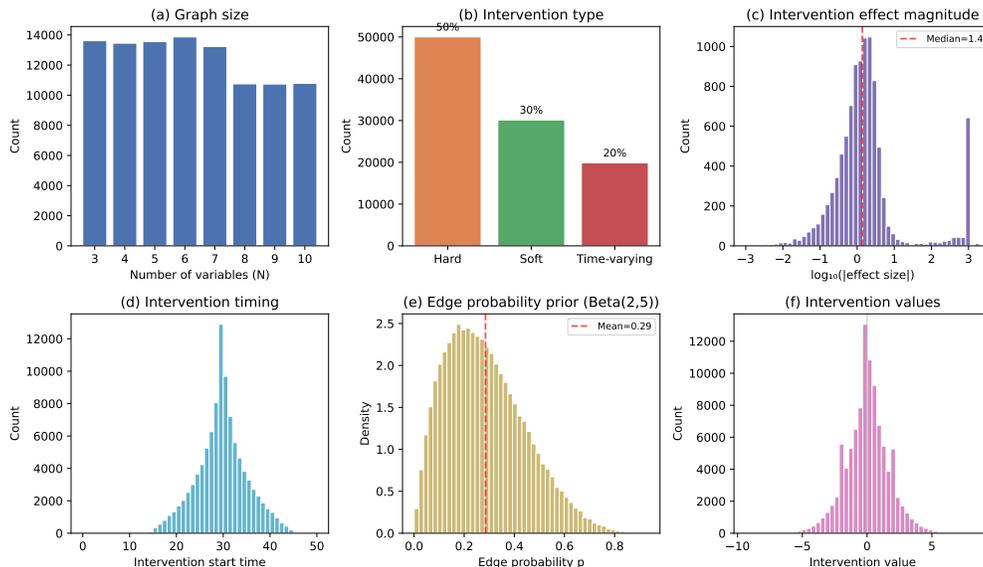

Figure 5: Distributions of prior properties across 100K sampled TSCMs from CausalTimePrior. The prior produces diverse graph structures, intervention types, and effect magnitudes.

G RESULTS

We train a 2-layer GRU encoder (128 hidden dim) for 15 epochs on 100K TSCMs and evaluate on 1,000 held-out TSCMs using three query types: (1) **Intervened** (query the intervention target itself), (2) **Downstream** (query a variable causally reachable from the intervention), and (3) **Non-causal** (query a variable with no causal path from the intervention).

Table 2 shows the model's predictions are most accurate for intervened queries (Pred/GT = 0.95), reasonable for downstream queries (0.85), and substantially biased downward for non-causal queries (0.46). The low non-causal ratio reflects the model correctly predicting near-zero causal effect for non-causal variables, combined with nonzero ground-truth values at those time steps due to the variables' own dynamics. The model predicts $\sim 2\times$ larger effects for causally connected queries compared to non-causal queries (33.91 vs 15.66 mean prediction).

13

Baselines. We compare against a Vector Autoregression (VAR-OLS) fitted per-dataset, PCMCI+ (Runge, 2020) which discovers causal graphs via conditional independence tests, and a mean prediction baseline (Table 3). SimpleCausalPFN achieves comparable RMSE to VAR-OLS (176.4 vs 176.5) while requiring no per-dataset fitting. PCMCI+ achieves lower overall RMSE (161.4) by leveraging discovered causal structure, but requires expensive per-sample graph discovery.

Shuffled-intervention control. To test whether the model's query-type sensitivity reflects learned causal structure rather than distributional artifacts, we evaluate with randomly shuffled intervention targets. Under shuffling, predictions for intervened queries change substantially (+33% mean prediction) and Pred/GT degrades from 0.95 to 1.26, indicating the model is sensitive to intervention target information. However, non-causal predictions remain low (+13%), suggesting the model has also partially learned distributional properties of variable positions, motivating richer architectures that more explicitly encode causal graph structure.

Table 2: Three-way evaluation on held-out TSCMs. NMSE (Normalized MSE = MSE/Var(GT); NMSE<1 is better than predicting the mean). The model's Pred/GT ratio is highest for intervened queries (0.95) and lowest for non-causal queries (0.46), suggesting learned causal understanding. Overall NMSE $\approx$ 1.0 indicates limited absolute prediction quality.

Query Type	Queries	RMSE ↓	NMSE ↓	Mean Pred	Mean GT	Pred/GT ↑↓= 1
Intervened	573	226.39	1.41	33.91	35.68	**0.95**
Downstream	270	231.31	0.67	57.50	67.70	0.85
Non-causal	157	**143.15**	**0.66**	15.66	33.99	0.46
Overall	1000	216.87	0.99	37.41	44.06	0.85

Table 3: Comparison with baselines on held-out TSCMs. SimpleCausalPFN achieves comparable RMSE to VAR-OLS while requiring no per-dataset fitting. PCMCI+ achieves the lowest RMSE but requires per-sample causal graph discovery.

Method	RMSE ↓	MAE ↓	Effect Dir. Acc. ↑	Effect Size Corr. ↑
Oracle	0.00	0.00	100.0%	1.000
PCMCI+	**161.38**	**28.67**	74.1%	0.784
SimpleCausalPFN	176.45	34.92	70.4%	**0.821**
VAR-OLS	176.45	33.86	**93.7%**	**0.821**
Mean Prediction	256.18	73.26	60.4%	0.286

H INTERVENTION TYPE ABLATION

We investigate whether training on diverse intervention types (hard, soft, time-varying) improves performance compared to training only on hard interventions. Table 4 compares the mixed-intervention model (trained on 100K TSCMs with all intervention types) against a hard-only model (trained on 10K TSCMs with only hard interventions) on the same three-way test set.

Table 4: Intervention type ablation. The mixed-intervention model achieves higher effect direction accuracy and effect size correlation compared to the hard-only model.

Model	RMSE ↓	MAE ↓	Effect Dir. Acc. ↑	Effect Size Corr. ↑
Mixed (100K)	216.87	81.68	**70.4%**	**0.821**
Hard-only (10K)	**212.16**	**69.68**	63.9%	0.691

The mixed model achieves higher effect direction accuracy (70.4% vs 63.9%) and effect size correlation (0.821 vs 0.691), suggesting that diversity in intervention types during training improves the model's ability to reason about causal effects. While the hard-only model achieves slightly lower

overall RMSE (212.16 vs 216.87), this is likely due to the simpler prediction task when only hard interventions are present.

I   OUT-OF-DISTRIBUTION GENERALIZATION

We evaluate whether the model generalizes to TSCMs with structural properties outside the training distribution. We generate an out-of-distribution (OOD) test set of 1,000 TSCMs with: (1) larger graphs ($N \in [8, 10]$ vs training mean $\sim$6), (2) maximum lag $K = 3$, (3) denser graphs (edge probability ≥ 0.3 vs training mean $\sim$0.29), and (4) only complex nonlinear mechanisms (sin, cos, square, tanhReLU).

Table 5: Out-of-distribution generalization. Performance degrades on OOD TSCMs with larger, denser graphs and complex mechanisms, but the model retains basic causal structure (downstream RMSE > intervened RMSE).

Test Set	RMSE $\downarrow$	NMSE $\downarrow$	MAE $\downarrow$	Effect Dir. Acc. $\uparrow$	Effect Size Corr. $\uparrow$
In-distribution	**216.87**	0.99	**81.68**	**70.4%**	**0.821**
OOD	265.97	**0.72**	163.25	62.7%	0.599

As expected, performance degrades on OOD data: RMSE increases from 216.87 to 265.97 and effect size correlation drops from 0.821 to 0.599. The lower OOD NMSE (0.72 vs 0.99) reflects higher ground-truth variance in the OOD test set rather than better prediction quality. However, the model retains some causal understanding, with intervened queries (RMSE=237.01) outperforming downstream queries (RMSE=313.92), consistent with the in-distribution pattern.

J   CAUSAL VS. CORRELATION-BASED PREDICTION

To illustrate how the PFN leverages causal structure rather than correlations, consider a concrete test case. In sample 626 from our test set, a 3-variable temporal TSCM is intervened on variable 2 at $t = 25$, and we query variable 0 at $t = 28$. There is no causal path from variable 2 to variable 0, but the observational time series exhibit a correlation of -0.49 between them. The ground truth interventional value (-0.056) is nearly identical to the observational baseline (-0.094), confirming a near-zero causal effect (0.039). The PFN predicts -0.050 (error 0.005), correctly recognizing the absence of causal influence despite the spurious correlation. In contrast, VAR-OLS—which relies on learned correlations without distinguishing causal from non-causal associations—predicts -0.992 (error 0.936), a $177\times$ larger prediction error. This pattern generalizes: across 157 non-causal queries, the PFN achieves lower prediction error than VAR-OLS in 45% of cases, with the largest gains precisely on samples exhibiting high spurious correlations ($|\rho| > 0.3$).