Estimating Interventional Outcomes over Time with Causal Normalizing Flow

Yoonseok Yeom[*1] **Jonghwan Kim**[*1] **Taehui Yun**[1] **Juhyun Lyu**[2] **Jung-Hee Kim**[2] **Sangmin Lee**[2]
Jinseok Yang[2] **Hyemin Jung**[2] **Woohyung Lim**[†2] **Sanghack Lee**[†1,3]

[1] Graduate School of Data Science, Seoul National University, South Korea
[2] LG AI Research, South Korea
[3] SNU-LG AI Research Center, South Korea

Abstract

Estimating outcome distributions under time-varying treatments is an essential task for personalized decision-making, particularly in domains such as healthcare. Most prior work in this area focuses on point predictions, which fail to capture the inherent variability in outcomes. Recent efforts in causal inference have begun integrating generative models to address this limitation by estimating interventional distributions. However, existing approaches—including causal normalizing flows—are generally restricted to static settings and are not well suited to sequential, time-dependent data. In this work, we propose a novel framework that extends causal normalizing flows to time-series, enabling simulation-based interventional density estimation over time. Our method learns representations of treatment and covariate history that capture temporal dependencies. Conditioned on these representations and guided by a causal graph, our flow-based model generates interventional samples, allowing for the simulation of outcome trajectories under alternative treatment strategies. We evaluate our approach on both linear and non-linear synthetic time-series as well as on a simulated tumor growth dataset, demonstrating that it achieves performance competitive with state-of-the-art baselines, while accommodating a broader spectrum of causal queries.

1 INTRODUCTION

Causal inference in time-series plays a critical role in many real-world applications, such as personalized healthcare and policy evaluation [Robins et al., 2000, van der Laan et al., 2005, Lim et al., 2018, Li et al., 2021]. These domains often require understanding the consequences of interventions over time, making causal reasoning under temporal dynamics particularly important. To address this, a number of methods—such as g-computation [Robins, 1986] and marginal structural models (MSMs) [Robins et al., 2000]—have been proposed for causal inference in time-varying settings. While these methods have led to significant progress, most of them primarily focus on point estimation, e.g., estimating the average treatment effect (ATE) under time-varying treatments. However, in practice, point estimates are often insufficient, as many decision-making scenarios demand a more complete understanding of uncertainty and variability. This highlights the need to estimate the distribution under an intervention. Such distributions allow practitioners to account for variations beyond the ATE, capturing statistical patterns like skewness or multimodality [Kennedy et al., 2023, Melnychuk et al., 2023].

Recent research has started to explore how generative models can support causal inference [Kocaoglu et al., 2017, Pawlowski et al., 2020, Xia et al., 2021, Khemakhem et al., 2021, Javaloy et al., 2023]. In particular, normalizing flows (NFs) [Rezende and Mohamed, 2015]—a class of expressive, invertible generative models—have shown great promise for this purpose. Recent works such as Khemakhem et al. [2021] and Javaloy et al. [2023] have applied autoregressive normalizing flows (ANFs) [Papamakarios et al., 2017] to causal inference tasks, leveraging the models' bijective and autoregressive structure. These methods demonstrate the potential of causal and generative modeling beyond point estimation. Nevertheless, existing flow-based methods have largely focused on static data, and few are designed to model multivariate time-series, where causal dependencies evolve over time and across variables. This limits their applicability in many real-world temporal settings.

To bridge this gap, we develop the **Time-Series Causal Normalizing Flow (TSCNF)**, which generalizes causal nor-

[*]Equal contribution.

[†]Corresponding authors: w.lim@lgresearch.ai, sanghack@snu.ac.kr.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:7856–7895.

malizing flows [Javaloy et al., 2023] to handle temporal causal dependencies in multivariate time-series data. With this capability, TSCNF produces multi-step-ahead distributions across various hypothetical interventions. A distinctive feature of our model is its ability to handle continuous treatments and simultaneously manage multiple treatments and outcomes, providing a degree of flexibility in variable assignment that is rarely offered by existing time-series causal methods.

Contributions We propose TSCNF, a unified framework for causal inference in multivariate time-series. 1) We develop a model architecture that decomposes temporal causal dependencies into time-lagged and contemporaneous components, enabling a principled extension of causal normalizing flows to time-series data. 2) We design a learning and inference framework that supports multi-step-ahead interventional distribution estimation under hypothetical treatment strategies. 3) We demonstrate that TSCNF achieves competitive performance across diverse experimental settings, while—unlike prior models—generalizing across various query types without requiring query-specific designs.

2 PRELIMINARIES

We use the following notation in this paper. Random variables are denoted by uppercase letters (e.g., X) and their realizations by lowercase letters (e.g., x). Boldface letters, $\mathbf{X}$ and $\mathbf{x}$, represent sets of random variables and their realizations, respectively. We consider multivariate time-series with discrete time steps $t \in \{1, \dots, T\}$, and let X_t^j denote the j-th variable at time t. For a time interval $t : K$, we write $X_{t:K}^j = (X_t^j, \dots, X_K^j)$ for the sequence of the j-th component. Likewise, the joint sequence over the interval is $\mathbf{X}_{t:K} = (\mathbf{X}_t, \dots, \mathbf{X}_K)$. The history of X_t^j up to time t is written as $\bar{X}_t^j = X_{1:t}^j = (X_1^j, \dots, X_t^j)$, while its future from time t onward is denoted by $\underline{X}_t^j = X_{t:T}^j = (X_t^j, \dots, X_T^j)$. Similarly, $\bar{\mathbf{X}}_t = \mathbf{X}_{1:t} = (\mathbf{X}_1, \dots, \mathbf{X}_t)$ denotes the joint history of all variables up to time t, and $\underline{\mathbf{X}}_t = \mathbf{X}_{t:T} = (\mathbf{X}_t, \dots, \mathbf{X}_T)$ denotes the joint future.

Causal Framework A structural causal model (SCM) $\mathcal{M}$ [Pearl, 2009] is a 4-tuple $(\mathbf{Z}, \mathbf{X}, \mathbf{F}, P(\mathbf{Z}))$[1] characterizing the data-generating process. Here, $\mathbf{Z}$ denotes a set of exogenous variables with a joint distribution $P(\mathbf{Z})$, and $\mathbf{X} = (X^1, \dots, X^d)$ denotes a set of endogenous variables. The set $\mathbf{F} = (f^1, \dots, f^d)$ consists of deterministic mappings that define the structural assignments for each endogenous variable. These assignments induce a causal graph $\mathcal{G}$ that encodes the causal dependencies among the variables. According to this graph $\mathcal{G}$, each mapping f^j specifies how the variable X^j is generated from a subset of other endogenous variables and its corresponding exogenous variables:

$$X^j = f^j(\mathbf{Pa}^j, \mathbf{Z}^j), \quad \text{for } j = 1, \dots, d, \tag{1}$$

where $\mathbf{Pa}^j \subseteq \mathbf{X} \setminus \{X^j\}$ denotes the set of endogenous variables that directly affect X^j.

The causal graph $\mathcal{G}$ is composed of vertices and (bi)directed edges, where each endogenous variable X^j corresponds to a vertex. A directed edge $(X^j \to X^i)$ exists if X^j is a parent of X^i, i.e., $X^j \in \mathbf{Pa}^i$. A bidirected edge $(X^j \leftrightarrow X^i)$ may arise when the corresponding exogenous variables are dependent; we rule out this possibility by restricting our analysis to Markovian models. Under the Markovian assumption, the causal graph $\mathcal{G}$ is a directed acyclic graph (DAG), and the exogenous variables are mutually independent, i.e., $P(\mathbf{Z}) = \prod_{j=1}^{d} P(\mathbf{Z}^j)$. An SCM provides a unified framework for understanding causality by incorporating the notion of intervention at its core. Pearl [1995] formalized interventions using the *do*-operator, denoted as $do(X^i = a)$, which replaces the structural equation for X^i with a constant a and leaves all other equations unchanged. This formulation allows us to evaluate interventional distributions such as $P(Y \mid do(X^i = a))$, describing how the distribution of Y changes when X^i is set externally to a.

2.1 NORMALIZING FLOWS

Normalizing flows [Rezende and Mohamed, 2015, Papamakarios et al., 2021] are a class of generative models that express a complex distribution over observed variables $\mathbf{X}$ by transforming a simple latent distribution $P(\mathbf{Z})$ through a sequence of invertible mappings. Using the change-of-variables formula, they yield an exact and tractable density $P(\mathbf{X})$. Given $\mathbf{z} \sim P(\mathbf{Z})$ (e.g., standard normal) and $\mathbf{X} = f_\theta(\mathbf{Z})$, the density of $\mathbf{X}$ is:

$$P_{\mathbf{X}}(\mathbf{x}) = P_{\mathbf{Z}}(f_\theta^{-1}(\mathbf{x})) \left| \det \nabla_{\mathbf{x}} f_\theta^{-1}(\mathbf{x}) \right|.$$

The function f_θ must be diffeomorphic, meaning it must be invertible and both f_θ and f_θ^{-1} must be differentiable. In practice, f_θ is typically parameterized as a composition of K invertible transformations to model complex densities, i.e., $f_\theta = f_K \circ f_{K-1} \circ \cdots \circ f_1$.

Autoregressive Normalizing Flows (ANFs) [Kingma et al., 2016, Papamakarios et al., 2017, Huang et al., 2018] are a special class of normalizing flows that introduces an autoregressive structure into the transformation f_θ. With a permutation (or ordering) π over observed variables, each output value is given by the transformation f_θ as:

$$x^j = f_\theta^j(z^j; \mathbf{x}^{<\pi(j)}), \quad \text{for } j = 1, \dots, d,$$

where f_θ^j is monotone in z^j and parameterized by the preceding variables $\mathbf{x}^{<\pi(j)}$ in the ordering π. This autoregressive structure yields a triangular Jacobian matrix for the

[1] An SCM typically uses $\mathbf{U}, \mathbf{V}$ [Pearl, 2009], whereas we use $\mathbf{Z}, \mathbf{X}$ to follow normalizing flow and time-series conventions.

inverse transformation f_θ^{-1}. As a result, computing the Jacobian determinant becomes tractable as it now reduces to the product of the diagonal entries: $\left|\det \nabla_{\mathbf{x}} f_\theta^{-1}(\mathbf{x})\right| = \prod_{j=1}^{d} \left|\partial z^j / \partial x^j\right|$. Khemakhem et al. [2021] pointed out shared characteristics between SCMs and autoregressive flows, particularly in their use of latent variables and variable ordering, albeit only for bivariate *affine* autoregressive flows. They show that an SCM $\mathcal{M}$ can be modeled using an autoregressive flow when the variable ordering π aligns with the causal ordering of $\mathcal{G}$ and remains fixed throughout the transformations. Based on this idea, several studies [Balgi et al., 2022, Javaloy et al., 2023] extend ANF to support causal inference tasks.

2.2 CAUSAL NORMALIZING FLOWS

CNF [Javaloy et al., 2023] builds on the foundational work of Khemakhem et al. [2021], which explored the connection between ANFs and SCMs. The goal is to recover SCM $\mathcal{M}$ by designing and training an ANF, assuming a causal ordering π and the fully-specified causal graph $\mathcal{G}$. For this purpose, they derived identifiability of exogenous variables of an SCM $\mathcal{M}$ using recent advances in non-linear independent component analysis (ICA) [Xi and Bloem-Reddy, 2023]. Given a class of SCMs with diffeomorphic data-generating processes, a causal graph, and no hidden confounding, they proved that the ANF and the SCM differ only up to an invertible, component-wise transformation of exogenous variables $\mathbf{Z}$.

$$x^j = f_\theta^j(z^j; \mathbf{Pa}^j), \quad \mathbf{z} \sim P(\mathbf{Z})$$

The identifiability result of CNF underpins *causal consistency*, meaning that functional dependencies of ANF align with the causal graph $\mathcal{G}$ from SCM $\mathcal{M}$. Javaloy et al. [2023] address this property in practice by using an abductive single-flow-layer design or by adding a causal-consistency regularizer when stacking multiple layers, which can be viewed as a restrictive design choice. To relax this, Zhou et al. [2025] extend CNFs to allow multiple hidden layers while preserving causal consistency.

CNFs implement interventions by modifying the exogenous variables associated with the intervened variables, such that the resulting density matches the target interventional distribution. This is made feasible by the Markovian assumption, which posits the absence of unmeasured confounders. The procedure begins by drawing a sample from the observational distribution and then replacing the values of the intervened variables with the desired values. The corresponding exogenous variables are inferred by applying the inverse transformation of the normalizing flow. These exogenous variables are then propagated forward through the model, ensuring that the downstream (descendant) variables reflect the effect of the intervention.

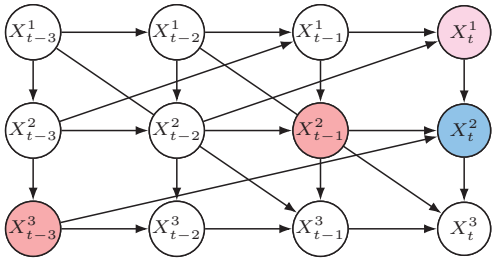


Figure 1: An example of a window causal graph with $\tau = 3$ highlighting contemporaneous (pink) and time-lagged (red) parents of a target variable (blue).

3 PROBLEM SETUP

Assumptions We assume that the data generating process (DGP) of multivariate time-series data admits a description by an SCM satisfying the following: the SCM is Markovian—its exogenous variables are mutually independent, ruling out hidden confounding—and *causally stationary*, in that the causal mechanisms and the exogenous marginals remain invariant over time, i.e., $f_t^j = f^j$ and $P(Z_t^j) = P(Z^j)$ for $j = 1, \dots, d$. In addition, each causal mechanism is diffeomorphic and strictly increasing in its exogenous argument, and the exogenous variables are independent of the observed history, $\mathbf{Z}_t \perp\!\!\!\perp \bar{\mathbf{X}}_{t-1}$ for $t \geq 2$. This SCM induces an acyclic *full time-series causal graph*, where the vertex set comprises variables indexed over discrete time steps $t \in \{1, \dots, T\}$ and each directed edge $X_{t-i}^p \to X_t^q$ indicates that X^p at time $t - i$ causally influences X^q at time t with lag i. Since the SCM is causally stationary, this global causal structure repeats over time, and can be represented via a fixed-size local representation, termed the *window causal graph*.

Definition 1 (Window causal graph; Assaad et al. 2022)**.** *Let $\mathcal{X}$ be a multivariate discrete-time stochastic process and $\mathcal{G} = (V, E)$ the associated window causal graph for a window of size τ. The set of vertices in that graph consists of the set of components $X^1, \dots, X^d$ at each time $t, \dots, t + \tau$. The edges E of the graph are defined as follows: variables X_{t-i}^p and X_t^q are connected by a lag-specific directed link $X_{t-i}^p \to X_t^q$ in graph $\mathcal{G}$ pointing forward in time if and only if X^p causes X^q at time t with a time lag of $0 \leq i \leq \tau$ for $p \neq q$ and with a time lag of $0 < i \leq \tau$ for $p = q$.*

An example of a window causal graph is shown in Fig. 1. We adopt the *window causal graph* as the underlying graphical representation of the causal mechanism in multivariate time-series data.[2] We assume this graph is given.[3]

[2]While we focus on causally-stationary time-series, our framework could in principle be extended to non-stationary settings via change point detection. However, modeling non-stationarity is beyond our current scope.

[3]In practice, such a graph may be supplied by prior do-

Target estimand Our goal is to estimate the distribution of outcome variable $Y_t \in \mathbf{X}_t$, under a hypothetical sequence of interventions on a treatment variable $A_t \in \mathbf{X}_t$, conditioned on the observed history, if necessary. Although we focus on a single outcome and treatment for simplicity, our framework *naturally generalizes* to multiple outcomes and interventions, with any variable in $\mathbf{X}_t$ capable of serving as either outcome or treatment. We denote the counterfactual variable of Y_t under a sequence of interventions by $Y_t(\underline{a}_{m+1})$—the value that Y_t would take if $\underline{A}_{m+1}$ were fixed to a set of constant values $\underline{a}_{m+1}$ through time t. Our target estimand is therefore:

$$P\left(Y_t(\underline{a}_{m+1}) \mid \bar{\mathbf{x}}_m\right) \quad \text{for } t \geq m+1, \tag{2}$$

where $\bar{\mathbf{x}}_m = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ denotes the observed history.[4] This interventional estimand reflects what the outcome variable Y_t would be under the treatment strategy $\underline{a}_{m+1}$, given the trajectory up to time m. It is particularly relevant in domains such as personalized healthcare, where a patient's past treatment assignments and covariate measurements inform future clinical decisions. The estimand encompasses many common probabilistic and causal queries: without intervention, it reduces to a conditional observational distribution; without conditioning on observed history, it becomes a marginal interventional distribution.

Causal Effect Identifiability Assuming the DGP follows a Markovian SCM with a known causal graph over multivariate time-series, the causal effect of an intervention sequence $\underline{a}_{m+1}$ on future outcomes is identifiable via a graphical adjustment criterion. Specifically, we can express the interventional distribution by adjusting for the parent sets in the graph. For identifiability, we require a positivity condition on the intervention value a_k relative to the local mechanism $P(a_k \mid \mathrm{Pa}(A_k))$, namely, $P(a_k \mid \mathrm{Pa}(A_k)) > 0$ whenever $P(\mathrm{Pa}(A_k)) > 0$, which guarantees that all considered interventions are supported by the observed data and therefore feasible. It is analogous to the standard positivity assumption in causal inference [Robins et al., 2000, van der Laan et al., 2005, Petersen et al., 2012] and aligns with the naturalness constraint in causal generative model literature [Hao et al., 2024]. Under these assumptions, our target estimand is identifiable within the SCM framework. This is an estimand-level identification statement, while model-level recovery of the target estimand requires a separate identifiability result, as discussed later in Sec. 4.4.

main knowledge or time-series causal-discovery pipelines such as TiMINo [Peters et al., 2013] and DYNOTEARS [Pamfil et al., 2020] [see also Sun et al., 2023, Faruque et al., 2024]. In this work, we focus on causal inference and density estimation given an input window causal graph; an empirical study of graph misspecification is provided in Appendix D.9.

[4]While we adopt this counterfactual notation (a Level 3 concept in Pearl's hierarchy [Pearl, 2009]) for notational compactness, it is important to note that our target estimand is fundamentally a conditional interventional distribution—a Level 2 quantity.

4 TIME-SERIES CAUSAL NORMALIZING FLOWS

Building upon the interventional estimand defined in Sec. 3, we now present Time-Series Causal Normalizing Flow (TSCNF), a CNF-based architecture extended to handle multivariate time-series. The practical estimation of such interventional distributions, however, poses several challenges: conditioning on the observed joint history is difficult, as it may be high-dimensional and exhibit complex temporal dependencies; and existing methods lack sampling schemes specifically designed for generating multi-step-ahead interventional trajectories. To address these limitations, TSCNF decouples the modeling of past and present variables via a *causal history encoder* and a *causal conditional normalizing flow* to capture temporal and contemporaneous dependencies in a principled manner. In the following, we illustrate how this architecture is trained and supports multi-step interventional sampling.

4.1 SEPARATING PAST AND PRESENT

To extend the CNF framework—originally developed for static settings—to time-series, it is crucial to distinguish lagged and contemporaneous causal dependencies. In TSCNF, lagged dependencies provide the temporal conditioning context that determines the system's state, whereas contemporaneous dependencies define the instantaneous structural mechanisms within the current time slice. This separation does not treat temporal dependencies as non-causal; rather, it assigns lagged and contemporaneous causes to different components of the model.

To formalize this separation, we first assume that the multivariate time-series $\mathcal{X} \in \mathbb{R}^{T \times d}$ is governed by a *window causal graph* $\mathcal{G}$ (Fig. 1). In this graph $\mathcal{G}$, each variable X_t^j has lagged parents $\mathrm{Pa}_{\mathcal{G}}^{<}(X_t^j)$ from previous time steps and contemporaneous parents $\mathrm{Pa}_{\mathcal{G}}^{=}(X_t^j)$ at time t. This structure induces the following factorization:

$$\begin{aligned} P(\mathcal{X}) &= \textstyle\prod_{t=1}^{T} P(\mathbf{X}_t \mid \mathbf{X}_{t-\tau:t-1}) \\ &= \textstyle\prod_{t=1}^{T} \prod_{j=1}^{d} P(X_t^j \mid \mathrm{Pa}_{\mathcal{G}}^{=}(X_t^j), \mathrm{Pa}_{\mathcal{G}}^{<}(X_t^j)). \end{aligned}$$

TSCNF leverages this decomposition by assigning lagged dependencies to a *causal history encoder*, and contemporaneous ones to a *causal conditional normalizing flow* (i.e., CNF conditioned on history encoding), applied at each step.

4.2 MODEL ARCHITECTURE AND TRAINING

Our model, TSCNF, comprises two main components: a *causal history encoder* that summarizes the observed history into a latent representation capturing relevant causal information, and a *causal conditional normalizing flow* that models present-time causal mechanisms conditioned on this

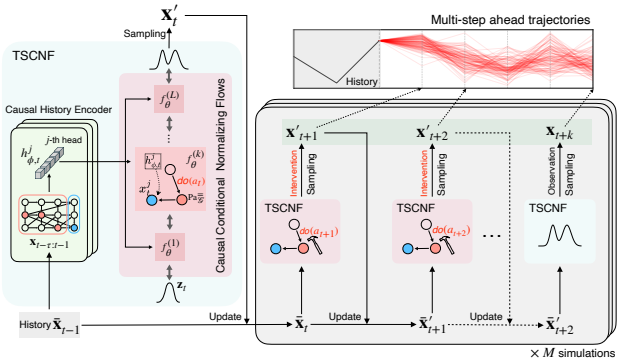


Figure 2: Overview of multi-step interventional trajectory sampling with TSCNF and its underlying architecture.

representation. As illustrated in Fig. 2, this architecture enables TSCNF to properly process multivariate time-series and estimate the desired interventional distribution.

Causal history encoder To extract meaningful history features for each variable, we design an encoder E_ϕ that focuses on *causally relevant* past inputs of each variable:

$$h^j_{\phi,t} = E^j_\phi \left(\mathbf{M}^j \odot \mathbf{X}_{t-\tau:t-1} \right), \tag{3}$$

where E^j_ϕ is the encoder for X^j_t, and $\mathbf{M}^j$ is a feature-wise causal mask selecting its time-lagged parents $\mathrm{Pa}^{<}_{\mathcal{G}}(X^j_t)$ based on the *window causal graph*. We use this feature-wise causal masking together with feature-wise conditioning throughout TSCNF, and evaluate both components in the ablation study of Sec. 5. This use of $h^j_{\phi,t}$ as a learned summary of the explicit lagged parents is a modeling approximation: the encoder is intended to preserve the information in $\mathrm{Pa}^{<}_{\mathcal{G}}(X^j_t)$ that is relevant for the local conditional mechanism. The encoder type depends on the window size τ: When τ is small relative to the sequence length T, we employ lightweight 1D convolutions to capture recent history:

$$h^j_{\phi,t} = \bar{\mathbf{W}}^j \cdot \mathrm{ReLU}\left(\mathbf{W}^j * \left(\mathbf{M}^j \odot \mathbf{X}_{t-\tau:t-1} \right) \right), \tag{4}$$

where $\mathbf{W}^j \in \mathbb{R}^{H \times \tau \times d}$ and $\bar{\mathbf{W}}^j \in \mathbb{R}^{o \times H}$ with convolutional channels H and output embedding dimension o. For larger windows where long-range dependencies matter ($\tau \approx T$), we adopt encoders with wider receptive fields, such as multi-layer dilated convolutions or self-attention mechanisms [Vaswani et al., 2017]. See Appendix D.10 for details on the encoder choices and ablation studies on their performance under different τ settings.

Causal Conditional Normalizing Flows We consider the CNF [Javaloy et al., 2023] for modeling contemporaneous dependencies among variables at time t. To properly adapt this model to the time-series setting, we integrate temporal context by conditioning the model on history embeddings $h^j_{\phi,t}$, obtained from a causal history encoder. These embeddings serve as additional conditioning inputs, enabling the model to account for both present and past influences. With this conditioning, each CNF transformation f^j_θ is parameterized not only by the variable's contemporaneous parents in the causal graph $\mathcal{G}$, but also by its own historical embedding $h^j_{\phi,t}$, as follows:

$$x^j_t = f^j_\theta(z^j_t; \mathrm{Pa}^{=}_{\mathcal{G}}(X^j_t), h^j_{\phi,t}), \quad z^j_t \overset{\text{i.i.d.}}{\sim} P_z \quad \forall j \tag{5}$$

We condition our model on the history embeddings by concatenating them to the model inputs, following the practice established in prior work on conditional normalizing flows [Rasul et al., 2021, Winkler et al., 2019]. Unlike prior work that relies on a global or shared context vector, our method applies feature-wise conditioning. Specifically, each variable X^j_t is modeled with its own history embedding $h^j_{\phi,t}$, while irrelevant historical signals are masked based on the *window causal graph*. This design enables fine-grained, causally-aware conditioning that respects the specific dependencies of each variable within the model. In addition, we adopt multiple NF layers as a flexible architectural choice, which distinguishes our approach from CNF. Further discussion is provided in Appendix B.4, with empirical evidence in Appendices D.11 and D.12.

Training objectives TSCNF is trained by maximizing the log-likelihood of the multivariate time-series $\mathcal{X}$, factorized according to the causal graph and historical encodings:

$$\log P(\mathcal{X}) \approx \textstyle\sum_{t=1}^{T} \sum_{j=1}^{d} \log P_\theta(X^j_t \mid \mathrm{Pa}^{=}_{\mathcal{G}}(X^j_t), h^j_{\phi,t})$$

Model parameters $\boldsymbol{\Theta} = \{\theta, \phi\}$ are optimized via:

$$\boldsymbol{\Theta}^* = \arg\max_{\boldsymbol{\Theta}} \textstyle\sum_{t=1}^{T} \sum_{j=1}^{d} \log P_\theta(X^j_t \mid \mathrm{Pa}^{=}_{\mathcal{G}}(X^j_t), h^j_{\phi,t}).$$

This objective encourages the encoder to capture causally relevant historical dependencies and guides the CNF to model contemporaneous causal interactions faithfully, enabling reliable density estimation for time-series data.

4.3 ESTIMATION OF INTERVENTIONAL DISTRIBUTION

We now describe the estimation of multi-step-ahead distribution under a hypothetical intervention sequence and observed history. Under assumptions outlined in Sec. 3, the joint interventional distribution of future variables $\mathbf{X}_t$ under the treatment strategy $a_{m+1:t-1}$, given joint history $\bar{\mathbf{x}}_m$, i.e., $P\left(\mathbf{X}_t(a_{m+1:t-1}) \mid \bar{\mathbf{x}}_m\right)$, is expressed as:

$$\textstyle\int_{\mathbf{x}'_{m+1:t-1}} P\left(\mathbf{X}_t \mid \bar{\mathbf{x}}_{t-1}\right) \cdot \prod_{k=m+1}^{t-1} P\left(\mathbf{x}'_k \mid do(a_k), \bar{\mathbf{x}}_{k-1}\right),$$

where the integration is over $\mathbf{x}'_{m+1:t-1}$ (i.e., the sequence of all variables except treatments $A_{m+1:t-1}$). This integration is equivalent to $\int_{\mathbf{x}'_{m+1}} \cdots \int_{\mathbf{x}'_{t-1}}$. See Appendix B.1 for the full derivation. While the expression defines the joint interventional distribution, the target quantity, such

Algorithm 1 Monte Carlo Estimation of $P(\mathbf{X}_{t+k} \mid do(a_t), \ldots, do(a_{t+k-1}), \mathbf{h})$

Input: History $\mathbf{h} = \bar{\mathbf{x}}_{t-1}$, interventions $\{a_t, \ldots, a_{t+k-1}\}$, number of samples M
Output: Empirical distribution over $\mathbf{X}_{t+k}$

1: Initialize empty sample list $\mathcal{S} \leftarrow [\,]$
2: **for** $m = 1$ to M **do**
3: Initialize $\mathbf{h}_0 \leftarrow \mathbf{h}$
4: **for** $j = 0$ to $k-1$ **do**
5: Sample $\mathbf{x}'_{t+j} \sim P(\mathbf{X}'_{t+j} \mid do(a_{t+j}), \mathbf{h}_j)$
6: Update $\mathbf{h}_{j+1} \leftarrow [\mathbf{h}_j \,\|\, [a_{t+j}, \mathbf{x}'_{t+j}]]$
7: Sample $\mathbf{x}_{t+k} \sim P(\mathbf{X}_{t+k} \mid \mathbf{h}_k)$
8: Append $\mathbf{x}_{t+k}$ to $\mathcal{S}$
9: **return** empirical distribution from $\mathcal{S}$

as $Y_t(a_{m+1:t-1})$, is obtained by marginalizing irrelevant variables.[5] Even though the target estimand admits a closed-form expression, its analytical computation is intractable. Hence, we approximate it via Monte Carlo simulation, following g-computation methods [Robins, 1986, Li et al., 2021]. Below, we propose an algorithm that samples from the desired distribution using TSCNF.

Sampling Algorithm Alg. 1 presents the full procedure for sampling from the target distribution defined by our estimand using a sequential sampling strategy, as illustrated in Fig. 2. The algorithm begins by initializing an empty list to collect samples of target variables $\mathbf{X}_{t+k}$ (line 1). It repeats the sampling process M times to form an empirical distribution (lines 2–8). In each iteration, the observed history is copied into a working variable (line 3). Then, for each step $j = 0$ to $k-1$, a set of intermediate variables $\mathbf{x}'_{t+j}$ is sampled from the conditional interventional distribution given $do(a_{t+j})$ and the current history (line 5), and the history is updated accordingly (line 6). This creates a forward chain of causal influence across k time steps (lines 4–6). After the chain, target variables $\mathbf{X}_{t+k}$ are sampled from the model conditioned on the final updated history (line 7), and added to the sample list (line 8). Once all samples are collected, the algorithm returns their empirical distribution as an approximation to the target (line 9). A detailed instantiation of this procedure for TSCNF, including its application to various interventional queries, is provided in Appendix C.

4.4 THEORETICAL DISCUSSION

Model Identifiability We now address the model identifiability of TSCNF to establish whether a TSCNF trained on observational data recovers the exogenous variables required to implement the interventions conditioned on the observed history, as described above. We show that, under the assumptions of Sec. 3 and provided that each observed history is sufficiently encoded by the history embedding $\mathbf{h}_{\phi,t}$, matching the observational distribution isolates each exogenous variable up to a single invertible, component-wise transformation that is invariant across time and observed histories. We formalize this result in the following theorem, whose proof is provided in Appendix B.2.

Theorem 1 (Identifiability of TSCNF)**.** *Suppose we fit TSCNF $f_\theta(\cdot \mid \mathbf{h}_{\phi,t})$ with fixed base distribution P_z. At the initial time step $t = 1$, the conditioning history is empty, so that TSCNF reduces to a static causal NF (f_θ, P_z). If this initial static model induces the same initial distribution as the SCM of the initial time step $(f_{\text{init}}, P(\mathbf{Z}_1))$, and, for almost every history $\bar{\mathbf{x}}_{t-1}$ that is sufficiently encoded by $\mathbf{h}_{\phi,t}$, the TSCNF $(f_\theta(\cdot \mid \mathbf{h}_{\phi,t}), P_z)$ induces the same conditional distribution as the history-conditioned contemporaneous SCM $(f_{\bar{\mathbf{x}}}, P(\mathbf{Z}_t))$, then there exists a component-wise invertible mapping*

$$\mathbf{a}(\mathbf{z}_t) = \big(a_1(z_t^1), \ldots, a_d(z_t^d)\big),$$

where $\mathbf{a}$ is independent of time and history, such that

$$f_\theta^{-1}\big(f_{\bar{\mathbf{x}}}(\mathbf{z}_t) \mid \mathbf{h}_{\phi,t}\big) = \mathbf{a}(\mathbf{z}_t)$$

for almost every history, for all $t = 2, \ldots, T$ (where the mapping at $t = 1$ reduces to $f_\theta^{-1}\big(f_{\text{init}}(\mathbf{z}_1)\big) = \mathbf{a}(\mathbf{z}_1)$).

Proof Sketch The proof builds upon the model identifiability of CNF, extending it to the temporal setting where the data-generating process is conditioned on the observed history. We first establish that, under a fixed history, both the true history-conditioned contemporaneous SCM (which governs the causal relations among variables at the current step given the past) and the conditional causal NF model (which transforms current variables to latents using the history as context) belong to the family of triangular monotone increasing (TMI) maps. Applying CNF identifiability then yields a unique, history-invariant mapping between latents and exogenous variables. Finally, since the causally masked TSCNF satisfies this conditional framework under sufficient history encoding, its identifiability follows.

Theorem 1 implies that the component-wise transformation $\mathbf{a}(\cdot)$ links each latent of TSCNF to a single exogenous variable. This allows us to implement conditional interventions by modifying only the latents of the intervened variables. Crucially, since this mapping is invariant across time and history, the latent-exogenous relation remains stable during the multi-step rollouts in Algorithm 1. Consequently, TSCNF can safely estimate sequential interventional distributions and successfully recover the target estimand at the model level.

[5] Here the outcome Y_t is sampled observationally; instead, applying an intervention at this step yields the target estimand $P(Y_t(\underline{a}_{m+1}) \mid \bar{\mathbf{x}}_m)$. Our comparison experiments follow this convention to match the baselines.

5 EMPIRICAL EVALUATION

The goal of our empirical evaluation is twofold: (1) to validate the architectural and modeling choices behind TSCNF via ablation studies, and (2) to demonstrate its competitiveness with state-of-the-art models via comparative evaluations. These experiments are designed to assess whether our modeling principles are empirically grounded and to highlight the effectiveness of TSCNF in standard benchmarks for causal inference under time-varying treatments. Additional experimental results are in Appendix D.8.

5.1 EXPERIMENT SETUP

Overall Setup Experiments are conducted on fully synthetic time-series datasets generated from both linear and non-linear data generating processes (DGPs), as well as on a simulated PK–PD tumor-growth benchmark; the corresponding causal structures are illustrated in Appendix D. We evaluate model performance using three complementary metrics: Maximum Mean Discrepancy (MMD) [Gretton et al., 2012], 1-D Wasserstein distance [Ramdas et al., 2017] to assess distributional similarity at a specific time step, and Root Mean Square Error (RMSE) [Li et al., 2021] to evaluate multi-step prediction accuracy across horizons up to k. Metric definitions are provided in Appendix D. Across all experiments, we consider two types of interventional queries: *marginal*, which target population-level effects under hypothetical treatment sequences (i.e., $P\left(Y_t(\underline{a}_{m+1})\right)$), and *history-adjusted*, which condition on each subject's covariate and treatment history to estimate individualized causal effects (i.e., $P\left(Y_t(\underline{a}_{m+1}) \mid \bar{\mathbf{x}}_m\right)$).

Ablation Setup To assess the contribution of core architectural components, we compare three variants of TSCNF: (1) the full model with both feature-wise causal masking via $\mathbf{M}^j$ and feature-wise conditioning via $h^j_{\phi,t}$ (Sec. 4); (2) a variant without feature-wise conditioning, relying solely on global context in the flow; and (3) a minimal version without either component, equivalent to a standard conditional normalizing flow.

The experiments are conducted on a synthetic dataset derived from the causal DAG shown in Fig. 6 (see Appendix D.2), which includes six variables at each time step. We select X^2_t and X^4_t as treatment variables and consider both single and joint interventions. For each intervention scenario, we examine the induced changes in the distributions of all remaining variables, allowing us to assess how effectively each model component contributes to capturing inter-variable dependencies. All ablation studies are conducted with fixed hyperparameters to ensure consistency across configurations. Additional experimental details are provided in Appendix D.5.

Comparison Setup on Synthetic Datasets We evaluate TSCNF against baselines that are specifically designed for either marginal or history-adjusted queries. Following Schomaker et al. [2024], we consider static treatment strategies in our experiments, where a fixed treatment value is applied at every time step. Specifically, we consider five such sequences, corresponding to the 25th, 37th, 50th, 63rd, and 75th percentiles of the empirical treatment distribution. For example, the sequence representing the 25th percentile assigns the 25th percentile treatment value at each time step. This design allows us to evaluate performance across a wide range of hypothetical interventions.

For marginal queries, we vary the size of the causal graph by setting $\tau \in \{1, 3, 5\}$ and assess model performance across the five treatment sequences. We compare TSCNF to MSCVAE [Wu et al., 2024] and G-Net [Li et al., 2021] (adapted to this setting by averaging predictions over covariate-treatment histories). In line with prior work, we report the mean distance instead of RMSE. For history-adjusted queries, we consider two evaluation regimes designed to probe different forms of time-varying confounding. The first adopts a DGP with $\tau = 30$ (Fig. 7a), where treatment depends on the full covariate and treatment history; the second uses $\tau = 5$ (Fig. 7b in Appendix D.2), where only a few recent covariates influence treatment. These settings highlight the model's ability to generalize across different temporal graphs, examining whether it can flexibly attend to causally relevant information depending on the underlying graph structure. In both settings, we compare TSCNF to G-Net. Details of the experiments are in Appendix D.6.

Comparison Setup on the Tumor Growth Dataset We evaluate TSCNF on the PK–PD tumor growth dataset [Geng et al., 2017], which simulates tumor progression under treatment and exhibits time-dependent confounding. The dataset includes continuous tumor volume measurements and binary treatment assignments (chemotherapy and radiotherapy). We compare TSCNF with baselines previously evaluated on this dataset for estimating the history-adjusted expected outcome under a hypothetical treatment sequence, $\mathbb{E}\left[Y_t(\underline{a}_{m+1}) \mid \bar{\mathbf{x}}_m\right]$: RMSN [Lim et al., 2018], G-Net [Li et al., 2021], Causal CPC [El Bouchattaoui et al., 2024], and DoFlow [Wu et al., 2026].

To assess performance under diverse time-varying confounding structures, we construct separate datasets for different levels of time-dependent confounding, controlled by $(\gamma_c, \gamma_d) \in \{(5, 5), (5, 0), (0, 5)\}$. For each setting, we evaluate on four static treatment strategies: no treatment, chemotherapy only, radiotherapy only, and both treatments. For TSCNF and G-Net, which are distributional models, we obtain point estimates by averaging 1,000 Monte Carlo samples per patient. We report Normalized RMSE (nRMSE, %), defined as the RMSE expressed as a percentage of the maximum tumor volume, for each prediction horizon $k \leq 5$ (i.e.,

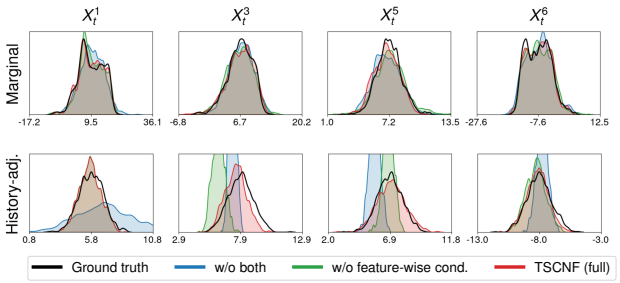


Figure 3: Marginal and history-adjusted interventional densities under $do(X_t^2, X_t^4)$ for TSCNF and its ablations in the non-linear dataset ($\tau = 3$).

Metric	Method	$do(X_t^2)$	$do(X_t^4)$	$do(X_t^2, X_t^4)$
Mean Dist.	w/o both	1.98 (1.32)	2.03 (1.10)	1.61 (1.16)
	w/o feature-wise cond.	1.48 (1.05)	1.30 (0.88)	1.18 (0.86)
	TSCNF	**0.72 (0.57)**	**0.66 (0.61)**	**0.51 (0.47)**
MMD	w/o both	0.53 (0.29)	0.46 (0.23)	0.48 (0.25)
	w/o feature-wise cond.	0.38 (0.25)	0.21 (0.18)	0.40 (0.23)
	TSCNF	**0.11 (0.14)**	**0.08 (0.12)**	**0.13 (0.21)**
Wass.	w/o both	2.19 (1.26)	2.39 (1.03)	1.79 (1.10)
	w/o feature-wise cond.	1.57 (1.03)	1.42 (0.87)	1.26 (0.85)
	TSCNF	**0.74 (0.57)**	**0.74 (0.63)**	**0.53 (0.47)**

Table 1: Ablation study of TSCNF for history-adjusted interventional distributions under $do(X_t^2)$, $do(X_t^4)$ or $do(X_t^2, X_t^4)$ in the non-linear dataset.

for the k-step-ahead outcome Y_{t+k}), matching prior baseline evaluations of this benchmark. Additional experimental details are provided in Appendix D.7.

5.2 RESULTS

Ablation Study Results Fig. 3 and Table 1 (see also Tables 6 and 7 in Appendix D.8) show that the full model consistently performs best. Specifically, applying causal graph masking alone leads to a notable improvement over the baseline. By filtering the input history to retain only causally relevant variables, the model avoids relying on spurious signals, resulting in cleaner and more informative representations. Adding feature-wise conditioning on top of this further improves performance. This mechanism enables each flow dimension to condition solely on its causal parents, promoting faithful causal modeling. The full TSCNF model, which incorporates both mechanisms, achieves the best performance, demonstrating the complementary nature of causal graph masking and feature-wise conditioning.

Comparison on Synthetic Datasets Table 2 (see also Table 8 in Appendix D.8) presents model comparison results under the marginal interventional setting. Both TSCNF and G-Net consistently outperform MSCVAE. Although MSCVAE is tailored for this task, it relies on inverse propensity weighting, multiplying treatment probabilities across time to

	Metric	Model	$\tau = 1$	$\tau = 3$	$\tau = 5$
Marginal	Mean Dist.	MSCVAE	0.13 (0.02)	0.52 (0.04)	1.97 (0.03)
		G-Net	**0.09 (0.04)**	**0.05 (0.03)**	0.09 (0.04)
		TSCNF	**0.09 (0.01)**	0.07 (0.02)	**0.05 (0.02)**
	MMD ($\times 10^{-2}$)	MSCVAE	0.95 (0.11)	2.05 (0.15)	20.15 (0.47)
		G-Net	**0.08 (0.05)**	0.12 (0.05)	0.13 (0.04)
		TSCNF	0.09 (0.02)	**0.06 (0.02)**	**0.05 (0.02)**
	Wasserstein	MSCVAE	0.24 (0.01)	0.70 (0.03)	2.04 (0.03)
		G-Net	**0.10 (0.03)**	0.14 (0.02)	0.16 (0.02)
		TSCNF	**0.10 (0.01)**	**0.11 (0.01)**	**0.11 (0.01)**
	Metric	**Model**	$k = 1$	$k = 2$	$k = 3$
History-adjusted	RMSE	G-Net	0.06 (0.00)	0.06 (0.00)	0.06 (0.00)
		TSCNF	**0.05 (0.00)**	**0.05 (0.00)**	**0.05 (0.00)**
	MMD ($\times 10^{-2}$)	G-Net	0.14 (0.00)	0.13 (0.00)	**0.12 (0.00)**
		TSCNF	**0.11 (0.01)**	**0.11 (0.01)**	**0.12 (0.00)**
	Wasserstein	G-Net	0.07 (0.00)	0.07 (0.00)	0.07 (0.00)
		TSCNF	**0.06 (0.00)**	**0.06 (0.00)**	**0.06 (0.00)**

Table 2: (Top) Marginal interventional density estimation at the final step across various window sizes τ. (Bottom) History-adjusted interventional density estimation across prediction horizons k in the non-linear dataset ($\tau = 30$).

construct pseudo-outcomes. This procedure is often numerically unstable and sensitive to the training data distribution, particularly over long horizons. In contrast, our model and G-Net follow a g-computation-style approach that avoids explicit propensity modeling. Though not inherently superior, g-computation can be more statistically efficient than MSMs when models are correctly specified [Daniel et al., 2013, Li et al., 2021].

For history-adjusted queries, Table 2 (see also Table 8 in Appendix D.8) shows that our model achieves comparable or superior performance relative to G-Net across both linear and non-linear experiments under two temporal graph settings. While results are generally close, our model remains competitive even in graphical settings where G-Net is specifically designed to perform well. The performance gap becomes more pronounced in experiments involving sparser time-lagged structures (see Fig. 4 and Table 9 in Appendix D.8). Here, G-Net incorporates all past covariates and treatments without filtering irrelevant history, which introduces noise under sparse graphs. In contrast, TSCNF adapts its modeling to the given graph by selectively attending to relevant time-lagged parents, allowing for more faithful estimation under varying structural assumptions. Unlike G-Net, which is tailored to a fixed graphical structure, TSCNF is inherently flexible and can accommodate any given causal graph as input, broadening its applicability across diverse settings.

Comparison on the Tumor Growth Dataset Table 3 shows that TSCNF performs best in most settings and remains consistently strong across confounding structures. In particular, when one treatment has no time-dependent con-

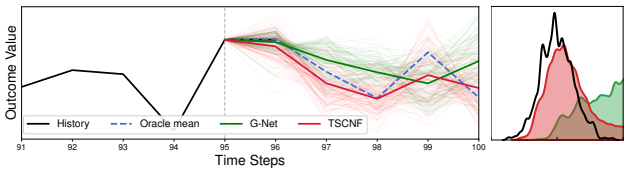


Figure 4: Simulated 5-step-ahead history-adjusted interventional trajectories and the final step density plot in the sparse temporal dependency setting.

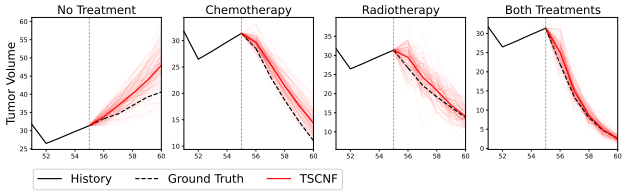


Figure 5: Tumor volume trajectories sampled from TSCNF for a single patient under various treatment strategies.

founding (e.g., $\gamma_c = 0$ or $\gamma_d = 0$), TSCNF maintains low error, demonstrating its flexibility in handling various input causal structures. Fig. 5 further shows that TSCNF captures tumor dynamics across all treatment strategies, including differences between single and combined treatments, while generating trajectory samples that reflect simulated outcome variability.

6 CONCLUSION

In this work, we introduced Time-Series Causal Normalizing Flow (TSCNF), a generalization of causal normalizing flows (CNFs) to multivariate time-series. Our method enables principled modeling of time-dependent data and supports the estimation of multi-step-ahead interventional distributions given observed histories and hypothetical intervention sequences. Through various experimental settings, including a clinically grounded simulation, we demonstrated that TSCNF performs competitively against existing baselines while accommodating a wider range of interventional queries. We hope that this work will inspire future efforts to advance flow-based approaches for causal decision-making in time-series settings.

Acknowledgements

This work was supported by LG AI Research. We thank all reviewers for constructive comments to improve the manuscript. This work was also supported by the NRF (RS-2023-00211904) grant funded by the Korean government.

	Model	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
$\gamma_c=5,\gamma_d=5$	RMSN	0.63 (0.11)	0.74 (0.12)	0.79 (0.11)	0.83 (0.11)	0.88 (0.11)
	G-Net	**0.52 (0.03)**	0.64 (0.04)	0.72 (0.06)	0.79 (0.09)	0.92 (0.24)
	Causal CPC	0.70 (0.10)	0.77 (0.10)	0.81 (0.10)	0.84 (0.12)	0.88 (0.14)
	DoFlow	1.01 (0.05)	1.54 (0.06)	1.90 (0.06)	2.19 (0.07)	2.40 (0.09)
	TSCNF	0.55 (0.09)	**0.58 (0.10)**	**0.63 (0.10)**	**0.67 (0.10)**	**0.71 (0.11)**
$\gamma_c=5,\gamma_d=0$	RMSN	0.33 (0.04)	0.38 (0.04)	0.41 (0.04)	0.43 (0.04)	0.45 (0.04)
	G-Net	0.30 (0.04)	0.34 (0.03)	0.37 (0.03)	0.43 (0.21)	0.51 (0.47)
	Causal CPC	0.32 (0.04)	0.36 (0.04)	0.38 (0.04)	0.41 (0.04)	**0.43 (0.04)**
	DoFlow	0.37 (0.04)	0.44 (0.05)	0.53 (0.05)	0.61 (0.06)	0.69 (0.06)
	TSCNF	**0.26 (0.07)**	**0.31 (0.06)**	**0.36 (0.05)**	**0.40 (0.06)**	0.44 (0.08)
$\gamma_c=0,\gamma_d=5$	RMSN	0.04 (0.02)	0.04 (0.02)	0.04 (0.02)	0.05 (0.02)	0.06 (0.02)
	G-Net	0.02 (0.01)	0.04 (0.01)	0.07 (0.09)	0.13 (0.26)	0.19 (0.30)
	Causal CPC	0.12 (0.04)	0.12 (0.04)	0.13 (0.05)	0.15 (0.07)	0.18 (0.08)
	DoFlow	0.34 (0.04)	0.35 (0.05)	0.38 (0.04)	0.42 (0.05)	0.45 (0.07)
	TSCNF	**0.02 (0.02)**	**0.02 (0.01)**	**0.03 (0.01)**	**0.03 (0.01)**	**0.04 (0.02)**

Table 3: Normalized RMSE (%) of the treatment effect on tumor volume for different (γ_c, γ_d) settings at each prediction horizon k (best in bold, second best underlined).

References

Charles K. Assaad, Emilie Devijver, and Éric Gaussier. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73: 767–819, 2022.

Sourabh Balgi, Jose M Pena, and Adel Daoud. Personalized public policy analysis in social sciences using causal-graphical normalizing flows. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11810–11818, 2022.

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.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Martina Cinquini, Isacco Beretta, Salvatore Ruggieri, and Isabel Valera. A practical approach to causal inference over time. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(14):14832–14839, 2025. doi: 10.1609/aaai.v39i14.33626.

Rhian M Daniel, Sara N Cousens, Bianca L De Stavola, Michael G Kenward, and Jonathan AC Sterne. Methods for dealing with time-dependent confounding. *Statistics in medicine*, 32(9):1584–1618, 2013.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In *Advances in Neural Information Processing Systems*, volume 32, pages 7511–7522, 2019.

Mouad El Bouchattaoui, Myriam Tami, Benoit Lepetit, and Paul-Henry Cournède. Causal contrastive learning for counterfactual regression over time. *Advances in Neural Information Processing Systems*, 37:1333–1369, 2024.

Omar Faruque, Sahara Ali, Xue Zheng, and Jianwu Wang. TS-CausalNN: Learning temporal causal relations from non-linear non-stationary time series data. *arXiv preprint arXiv:2404.01466*, 2024.

Changran Geng, Harald Paganetti, and Clemens Grassberger. Prediction of treatment response for combined chemo-and radiation therapy for non-small cell lung cancer patients using a bio-mathematical model. *Scientific reports*, 7(1): 13542, 2017.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Guang-Yuan Hao, Jiji Zhang, Biwei Huang, Hao Wang, and Kun Zhang. Natural counterfactuals with necessary backtracking. *Advances in Neural Information Processing Systems*, 37:14962–14995, 2024.

Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron C. Courville. Neural autoregressive flows. In *International Conference on Machine Learning*, 2018.

Adrián Javaloy, Pablo Sánchez-Martín, and Isabel Valera. Causal normalizing flows: from theory to practice. In *Advances in Neural Information Processing Systems*, volume 36, pages 58833–58864, 2023.

Edward H Kennedy, Sivaraman Balakrishnan, and LA Wasserman. Semiparametric counterfactual density estimation. *Biometrika*, 110(4):875–896, 2023.

Ilyes Khemakhem, Ricardo Monti, Robert Leech, and Aapo Hyvarinen. Causal autoregressive flows. In *International conference on artificial intelligence and statistics*, pages 3520–3528. PMLR, 2021.

Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improving variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, 2016.

Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. CausalGAN: Learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023*, 2017.

Rui Li, Stephanie Hu, Mingyu Lu, Yuria Utsumi, Prithwish Chakraborty, Daby M Sow, Piyush Madan, Jun Li, Mohamed Ghalwash, Zach Shahn, et al. G-Net: a recurrent network approach to g-computation for counterfactual prediction under a dynamic treatment regime. In *Machine Learning for Health*, pages 282–299. PMLR, 2021.

Bryan Lim, Ahmed M. Alaa, and Mihaela van der Schaar. Forecasting treatment responses over time using recurrent marginal structural networks. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 15293–15329. PMLR, 2022.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Normalizing flows for interventional density estimation. In *International Conference on Machine Learning*, pages 24361–24397. PMLR, 2023.

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.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *Advances in neural information processing systems*, 33:857–869, 2020.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

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

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Maya L Petersen, Kristin E Porter, Susan Gruber, Yue Wang, and Mark J Van Der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research*, 21(1):31–54, 2012.

Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

Kashif Rasul, Abdul-Saboor Sheikh, Ingmar Schuster, Urs Bergmann, and Roland Vollgraf. Multivariate probabilistic time series forecasting via conditioned normalizing flows. In *The Ninth International Conference on Learning Representations (ICLR)*, 2021. Spotlight.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

James M Robins, Sander Greenland, and Fu-Chang Hu. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association*, 94(447):687–700, 1999.

James M Robins, Miguel Ángel Hernán, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

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

Michael Schomaker, Helen McIlleron, Paolo Denti, and Iván Díaz. Causal inference for continuous multiple time point interventions. *Statistics in Medicine*, 43(28):5380–5400, 2024.

Xiangyu Sun, Oliver Schulte, Guiliang Liu, and Pascal Poupart. NTS-NOTEARS: Learning nonparametric DBNs with prior knowledge. In *International Conference on Artificial Intelligence and Statistics*, pages 1942–1964. PMLR, 2023.

Mark J van der Laan, Maya L Petersen, and Marshall M Joffe. History-adjusted marginal structural models and statically-optimal dynamic treatment regimens. *The International Journal of Biostatistics*, 1(1), 2005.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*, 2019.

Dongze Wu, Feng Qiu, and Yao Xie. DoFlow: Flow-based generative models for interventional and counterfactual forecasting on time series. In *The Fourteenth International Conference on Learning Representations*, 2026.

Shenghao Wu, Wenbin Zhou, Minshuo Chen, and Shixiang Zhu. Counterfactual generative models for time-varying treatments. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3402–3413, 2024.

Quanhan Xi and Benjamin Bloem-Reddy. Indeterminacy in generative models: Characterization and strong identifiability. In *International Conference on Artificial Intelligence and Statistics*, pages 6912–6939. PMLR, 2023.

Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. *Advances in Neural Information Processing Systems*, 34:10823–10836, 2021.

Qingyang Zhou, Kangjie Lu, and Meng Xu. Causally consistent normalizing flow. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 22974–22981, 2025.

Estimating Interventional Outcomes over Time with Causal Normalizing Flow Supplementary Material

Yoonseok Yeom[*1] **Jonghwan Kim**[*1] **Taehui Yun**[1] **Juhyun Lyu**[2] **Jung-Hee Kim**[2] **Sangmin Lee**[2]
Jinseok Yang[2] **Hyemin Jung**[2] **Woohyung Lim**[†2] **Sanghack Lee**[†1,3]

[1] Graduate School of Data Science, Seoul National University, South Korea
[2] LG AI Research, South Korea
[3] SNU-LG AI Research Center, South Korea

A APPENDIX FOR PRELIMINARY

A.1 EXTENDED RELATED WORKS

A.1.1 Causal Inference with Time-Series

On the use of *Counterfactual* terminology and Notation Throughout the literature on causal inference with time-varying treatments, the terms *counterfactual outcome, counterfactual prediction, and counterfactual regression* have been predominantly used within the Potential Outcome (PO) framework [Rubin, 1974]. However, viewed through the lens of Pearl's Structural Causal Model (SCM) [Pearl, 2009] and the Ladder of Causation, these tasks fundamentally correspond to Level 2 interventional queries rather than true Level 3 counterfactuals. In this paper, we adopt the SCM as our primary framework. While we utilize counterfactual notation (a Level 3 concept) for mathematical compactness when denoting hypothetical treatment trajectories, our task is strictly framed as modeling conditional interventional distributions (a Level 2 concept) in a sequential setting. We review the following related works with this terminological distinction in mind.

Early statistical approaches Robins' seminal work on g-computation [Robins, 1986] laid the foundation for causal inference in longitudinal settings with time-varying treatments. This approach formalized the use of potential outcomes under hypothetical interventions in sequential treatment strategies. Building on this, Marginal Structural Models (MSMs) [Robins et al., 2000] were introduced to estimate population-level causal effects by reweighting samples using inverse probability of treatment weighting (IPTW) across time. However, MSMs are limited in that they only account for population-level averages, making them more appropriate for policy-level evaluation than for personalized treatment decisions. To address these limitations, history-adjusted MSMs (HA-MSMs) [van der Laan et al., 2005] were proposed as a generalization that conditions on individual-level histories of covariates and treatments. HA-MSMs thus enable estimation of personalized effects of treatments and better reflect clinical decision-making scenarios. Nevertheless, both MSMs and HA-MSMs typically rely on restrictive parametric assumptions which can be easily violated in practice, especially in high-dimensional and non-linear settings.

Deep models for time-varying treatments Restrictive parametric assumptions motivated a shift toward flexible, learning-based approaches. Recurrent Marginal Structural Networks (RMSN) [Lim et al., 2018] were among the first to incorporate deep learning into the HA-MSM framework. RMSN uses recurrent neural networks (RNNs) to estimate time-varying treatment probabilities and outcome models, offering a more expressive alternative to traditional implementations. CRN [Bica et al., 2020] extended this direction by introducing balanced representation learning via adversarial domain training. Instead of using observed covariates directly, CRN learns representations that are invariant to treatment assignment, effectively mitigating bias from time-varying confounders and enabling valid counterfactual estimation under joint histories. G-Net [Li et al., 2021] took a different approach by developing a simulation-based model of g-computation using RNNs. Unlike most prior works, G-Net models the conditional distribution of potential outcomes through Monte Carlo simulation, thereby supporting density estimation. Causal Transformer [Melnychuk et al., 2022] further advanced counterfactual regression by adopting the Transformer architecture [Vaswani et al., 2017] to capture long-range temporal dependencies. Like CRN, it

employs domain-invariant representations to deconfound the learned embeddings, and introduces a counterfactual domain confusion loss for the learning process. Building on recent advances, Causal CPC [El Bouchattaoui et al., 2024] employs contrastive learning to tackle counterfactual regression over time, facilitating more efficient learning. Recently, Cinquini et al. [2025] proposed a practical approach to causal inference over time based on vector autoregressive processes; in contrast to the methods discussed above, it targets causal effects at equilibrium in linear settings rather than outcomes under time-varying treatments.

From point estimates to distribution modeling Despite recent advances, most prior methods—including RMSN, CRN, Causal Transformer, and Causal CPC—focus on point estimation of counterfactual outcomes and do not model predictive distributions of potential outcomes. A recent exception is MSCVAE [Wu et al., 2024], which introduces a generative modeling framework for causal inference over time using variational autoencoders (VAE). Grounded in the MSM formalism, MSCVAE estimates distributions over potential outcomes under time-varying treatment sequences. However, its scope remains limited to marginal (population-level) quantities and does not support history-adjusted counterfactuals. In contrast, our proposed model fills this gap by combining the advantages of generative modeling with the flexibility to support both marginal and history-adjusted queries. Leveraging normalizing flows, our model provides full density estimates of potential outcomes while respecting the causal structure across time. This enables a unified framework capable of addressing a wide range of interventional queries without relying on restrictive parametric assumptions or being tied to particular graphical structures.

B APPENDIX FOR THEORETICAL RESULTS

B.1 DERIVATION OF TARGET ESTIMAND

We provide the derivation of the target estimand as follows. Let $\mathbf{X}'_t = \mathbf{X}_t \setminus \{A_t\}$. $\mathbf{H} = \mathbf{H}_{t-1} = \mathbf{X}_{t-l:t-1}$, where l is a lag such that $l \geq \tau$. To marginalize over all intermediate variables, we write

$$P(\mathbf{X}_{t+k} \mid do(a_t), \ldots, do(a_{t+k-1}), \mathbf{h}) = \int_{\mathbf{x}'_t} \cdots \int_{\mathbf{x}'_{t+k-1}} P(\mathbf{X}_{t+k}, \mathbf{x}'_{t+k-1}, \ldots, \mathbf{x}'_t \mid do(a_t), \ldots, do(a_{t+k-1}), \mathbf{h}). \quad (6)$$

By applying the chain rule, the joint term inside the integral can be expressed as

$$P(\mathbf{X}_{t+k}, \mathbf{x}'_{t+k-1}, \ldots, \mathbf{x}'_t \mid do(a_t), \ldots, do(a_{t+k-1}), \mathbf{h}) = \left(\prod_{j=0}^{k-1} P(\mathbf{x}'_{t+j} \mid \mathbf{x}'_t, \ldots, \mathbf{x}'_{t+j-1}, do(a_t), \ldots, do(a_{t+k-1}), \mathbf{h}) \right) \cdot P(\mathbf{X}_{t+k} \mid \mathbf{x}'_t, \ldots, \mathbf{x}'_{t+k-1}, do(a_t), \ldots, do(a_{t+k-1}), \mathbf{h}). \quad (7)$$

Under the rules of do-calculus [Pearl, 2009], each factor in the product simplifies as

$$P(\mathbf{x}'_{t+j} \mid \mathbf{x}'_t, \ldots, \mathbf{x}'_{t+j-1}, do(a_t), \ldots, do(a_{t+k-1}), \mathbf{h}) = P(\mathbf{x}'_{t+j} \mid do(a_{t+j}), \mathbf{h}_{t+j-1}). \quad (8)$$

The final term outside the product becomes

$$P(\mathbf{X}_{t+k} \mid \mathbf{x}'_t, \ldots, \mathbf{x}'_{t+k-1}, do(a_t), \ldots, do(a_{t+k-1}), \mathbf{h}) = P(\mathbf{X}_{t+k} \mid \mathbf{h}_{t+k-1}). \quad (9)$$

Combining all these terms, we arrive at the general form:

$$P(\mathbf{X}_{t+k} \mid do(a_t), \ldots, do(a_{t+k-1}), \mathbf{h}) = \int_{\mathbf{x}'_t} \cdots \int_{\mathbf{x}'_{t+k-1}} \left(\prod_{j=0}^{k-1} P(\mathbf{x}'_{t+j} \mid do(a_{t+j}), \mathbf{h}_{t+j-1}) \right) \cdot P(\mathbf{X}_{t+k} \mid \mathbf{h}_{t+k-1}). \quad (10)$$

B.2 PROOF OF THM. 1

This section states the assumptions formally and proves Thm. 1.

Assumption 1 (Acyclicity)**.** *The causal graph at the initial time step and the contemporaneous graph at each time step are DAGs with known causal orderings.*

Assumption 2 (Markovianity)**.** *The corresponding exogenous variables are mutually independent. Equivalently, the SCM has no hidden confounding.*

Assumption 3 (Causally stationary SCM)**.** *The data-generating process is governed by a causally stationary SCM. Specifically, the structural assignments $F = (f^1, \dots, f^d)$, and the marginal distributions of exogenous noises $P(Z_t^j) = P(Z^j)$ do not depend on t. In particular, for $t \geq 2$ and $j = 1, \dots, d$,*

$$X_t^j = f^j\big(Z_t^j; \mathrm{Pa}_{\mathcal{G}}^{=}(X_t^j), \mathrm{Pa}_{\mathcal{G}}^{<}(X_t^j)\big),$$

with the initial time step generated by

$$\mathbf{X}_1 = f_{\mathrm{init}}(\mathbf{Z}_1).$$

Assumption 4 (Diffeomorphism and monotonicity)**.** *The SCM of the initial time step and each history-conditioned contemporaneous SCM are diffeomorphic, with each structural assignment strictly increasing in its exogenous argument, for all values of the parents of the corresponding variable.*

Assumption 5 (History-independent exogenous variables)**.** *For every $t \geq 2$,*

$$\mathbf{Z}_t \perp\!\!\!\perp \bar{\mathbf{X}}_{t-1}.$$

Assumption 6 (Fixed base distribution)**.** *TSCNF uses a fixed factorized base distribution. In the notation of TSCNF,*

$$z_t^j \overset{\text{i.i.d.}}{\sim} P_z, \qquad j = 1, \dots, d,$$

and this base distribution is independent of t and of the observed history. To obtain a single history-invariant reparameterization, we assume that the one-dimensional monotone transports from $P(Z^j)$ to P_z are unique; for instance, their distribution functions are continuous and strictly increasing.

Assumption 7 (Sufficient history encoding)**.** *The TSCNF history encoder is feature-wise causally masked:*

$$h_{\phi,t}^j = E_\phi^j\big(\mathbf{M}^j \odot \mathbf{X}_{t-\tau:t-1}\big),$$

where $\mathbf{M}^j$ is the feature-wise causal mask selecting the lagged parents $\mathrm{Pa}_{\mathcal{G}}^{<}(X_t^j)$. We assume that $h_{\phi,t}^j$ is sufficient for the local conditional mechanism of X_t^j; equivalently, replacing the explicit lagged parents $\mathrm{Pa}_{\mathcal{G}}^{<}(X_t^j)$ by $h_{\phi,t}^j$ preserves the corresponding local mechanism. If the model conditions directly on the explicit lagged parents, this assumption is automatic.

Theorem 2 (Identifiability of CNF; Javaloy et al. 2023)**.** *Let $\mathcal{T}$ be the family of triangular monotone increasing maps and let $\mathcal{P}_z$ be the family of fully factorized distributions. If two elements of $\mathcal{T} \times \mathcal{P}_z$ induce the same observational distribution, then the two data-generating processes differ by an invertible component-wise transformation of the exogenous variables. In particular, if $(f, P_{\mathcal{M}}) \in \mathcal{T} \times \mathcal{P}_z$ is the true SCM representation and a causal NF $(f_\theta, P_z) \in \mathcal{T} \times \mathcal{P}_z$ induces the same observational distribution, then there exists*

$$\mathbf{a}(\mathbf{z}) = \big(a_1(z^1), \dots, a_d(z^d)\big)$$

with each a_j invertible such that

$$f_\theta^{-1}(f(\mathbf{z})) = \mathbf{a}(\mathbf{z}).$$

Lemma 1 (History-conditioned SCM as a TMI map)**.** *Fix an observed history $\bar{\mathbf{x}}_{t-1}$. The induced contemporaneous SCM over the current slice is*

$$X_t^j = f_{\bar{\mathbf{x}}}^j\big(Z_t^j; \mathrm{Pa}_{\mathcal{G}}^{=}(X_t^j)\big) := f^j\big(Z_t^j; \mathrm{Pa}_{\mathcal{G}}^{=}(X_t^j), \bar{\mathbf{x}}_{t-1}\big).$$

This history-conditioned contemporaneous SCM admits a representation

$$(f_{\bar{\mathbf{x}}}, P(\mathbf{Z}_t)) \in \mathcal{T} \times \mathcal{P}_z.$$

Proof of Lem. 1. Once an observed history is fixed, all lagged parents become constant inputs, and the remaining system is the contemporaneous SCM over $(X_t^1, \dots, X_t^d)$, whose exogenous law is unaffected by the conditioning: by Assump. 5, $P(\mathbf{Z}_t \mid \bar{\mathbf{x}}_{t-1}) = P(\mathbf{Z}_t)$. By Assumps. 1, 2, and 4, this history-conditioned contemporaneous SCM is acyclic, diffeomorphic, and Markovian, with a known causal ordering. Hence, by the SCM-to-triangular-map construction used in Thm. 2, it admits a triangular monotone increasing representation with a factorized exogenous distribution. □

Lemma 2 (Conditional CNF as a TMI map)**.** *Let $f_\theta(\cdot \mid \bar{\mathbf{x}}_{t-1})$ be a conditional CNF such that, for the ordering π,*

$$x_t^{\pi(k)} = f_\theta^{\pi(k)}\big(z_t^{\pi(k)}; \mathbf{x}_t^{\pi(<k)}, \bar{\mathbf{x}}_{t-1}\big), \qquad z_t^{\pi(k)} \overset{i.i.d.}{\sim} P_z, \qquad k = 1, \dots, d,$$

where each transformation is strictly increasing in $z_t^{\pi(k)}$. Then, for any fixed observed history $\bar{\mathbf{x}}_{t-1}$, the map $\mathbf{z}_t \mapsto f_\theta(\mathbf{z}_t \mid \bar{\mathbf{x}}_{t-1})$ belongs to $\mathcal{T}$, and $\big(f_\theta(\cdot \mid \bar{\mathbf{x}}_{t-1}), P_z\big) \in \mathcal{T} \times \mathcal{P}_z$.

Proof of Lem. 2. For fixed $\bar{\mathbf{x}}_{t-1}$, the coordinate $x_t^{\pi(k)}$ depends only on $z_t^{\pi(k)}$ and the previously generated variables $\mathbf{x}_t^{\pi(<k)}$, which themselves depend only on $\mathbf{z}_t^{\pi(<k)}$. Hence $x_t^{\pi(k)}$ depends only on $\mathbf{z}_t^{\pi(\leq k)}$, so the Jacobian is triangular in the order π. Strict monotonicity gives positive diagonal entries, and therefore the map is triangular monotone increasing. □

Proposition 1 (Identifiability of Conditional CNF)**.** *Let $\mathcal{T}$ be the family of triangular monotone increasing maps and let $\mathcal{P}_z$ be the family of fully factorized distributions. If, for almost every history $\bar{\mathbf{x}}_{t-1}$, a conditional causal NF $(f_\theta(\cdot \mid \bar{\mathbf{x}}_{t-1}), P_z)$ induces the same conditional distribution as the history-conditioned contemporaneous SCM $(f_{\bar{\mathbf{x}}}, P(\mathbf{Z}_t))$, then there exists a component-wise invertible map*

$$\mathbf{a}(\mathbf{z}_t) = \big(a_1(z_t^1), \dots, a_d(z_t^d)\big),$$

independent of time and history, such that

$$f_\theta^{-1}\big(f_{\bar{\mathbf{x}}}(\mathbf{z}_t) \mid \bar{\mathbf{x}}_{t-1}\big) = \mathbf{a}(\mathbf{z}_t)$$

for almost every history.

Proof of Prop. 1. Fix $\bar{\mathbf{x}}_{t-1}$ outside a null set. By Lem. 1, $(f_{\bar{\mathbf{x}}}, P(\mathbf{Z}_t)) \in \mathcal{T} \times \mathcal{P}_z$. By Lem. 2, $(f_\theta(\cdot \mid \bar{\mathbf{x}}_{t-1}), P_z) \in \mathcal{T} \times \mathcal{P}_z$. Since the induced conditional distributions match, the identifiability result of static CNF (Thm. 2) yields a component-wise invertible map $\mathbf{a}_{\bar{\mathbf{x}}}$ such that

$$f_\theta^{-1}\big(f_{\bar{\mathbf{x}}}(\mathbf{z}_t) \mid \bar{\mathbf{x}}_{t-1}\big) = \mathbf{a}_{\bar{\mathbf{x}}}(\mathbf{z}_t).$$

It remains to show that $\mathbf{a}_{\bar{\mathbf{x}}}$ does not depend on the fixed history.

Since $f_{\bar{\mathbf{x}}} \in \mathcal{T}$ and $f_\theta(\cdot \mid \bar{\mathbf{x}}_{t-1}) \in \mathcal{T}$, and since the class of triangular monotone increasing maps is closed under inversion and composition, the map $\mathbf{a}_{\bar{\mathbf{x}}} = f_\theta^{-1}(\cdot \mid \bar{\mathbf{x}}_{t-1}) \circ f_{\bar{\mathbf{x}}}$ is itself triangular monotone increasing. Thm. 2 implies that $\mathbf{a}_{\bar{\mathbf{x}}}$ is component-wise. Hence, each coordinate $a_{\bar{\mathbf{x}},j}$ is a one-dimensional strictly increasing map transporting the exogenous marginal $P(Z_t^j)$ to the fixed base marginal P_z. By Assump. 3 (causal stationarity), $P(Z_t^j) = P(Z^j)$ for all t. Thus each $a_{\bar{\mathbf{x}},j}$ is a one-dimensional monotone transport from $P(Z^j)$ to P_z, and by the uniqueness of such transports (Assump. 6),

$$a_{\bar{\mathbf{x}},j} = F_{P_z}^{-1} \circ F_{P(Z^j)} \qquad \text{a.e.}$$

This expression depends only on the marginal distribution $P(Z^j)$ of the exogenous variable and the fixed base marginal P_z, and therefore is independent of t and of the fixed history. Thus $\mathbf{a}_{\bar{\mathbf{x}}} = \mathbf{a}$ for almost every history. □

Theorem 1 (Identifiability of TSCNF)**.** *Suppose we fit TSCNF $f_\theta(\cdot \mid \mathbf{h}_{\phi,t})$ with fixed base distribution P_z. At the initial time step $t = 1$, the conditioning history is empty, so that TSCNF reduces to a static causal NF (f_θ, P_z). If this initial static model induces the same initial distribution as the SCM of the initial time step $(f_{\text{init}}, P(\mathbf{Z}_1))$, and, for almost every history $\bar{\mathbf{x}}_{t-1}$ that is sufficiently encoded by $\mathbf{h}_{\phi,t}$, the TSCNF $(f_\theta(\cdot \mid \mathbf{h}_{\phi,t}), P_z)$ induces the same conditional distribution as the history-conditioned contemporaneous SCM $(f_{\bar{\mathbf{x}}}, P(\mathbf{Z}_t))$, then there exists a component-wise invertible mapping*

$$\mathbf{a}(\mathbf{z}_t) = \big(a_1(z_t^1), \dots, a_d(z_t^d)\big),$$

where $\mathbf{a}$ is independent of time and history, such that

$$f_\theta^{-1}\big(f_{\bar{\mathbf{x}}}(\mathbf{z}_t) \mid \mathbf{h}_{\phi,t}\big) = \mathbf{a}(\mathbf{z}_t)$$

for almost every history, for all $t = 2, \dots, T$ (where the mapping at $t = 1$ reduces to $f_\theta^{-1}\big(f_{\text{init}}(\mathbf{z}_1)\big) = \mathbf{a}(\mathbf{z}_1)$).

Proof of Thm. 1. For $t = 1$: since the conditioning history is empty, f_θ reduces to a static causal NF $(f_\theta, P_z) \in \mathcal{T} \times \mathcal{P}_z$. By the static CNF identifiability (Thm. 2), the recovered latent mapping is coordinate-wise given by $F_{P_z}^{-1} \circ F_{P(Z_1^j)}$. Since

Assump. 3 stipulates that the initial noise marginal matches the subsequent stationary noise marginal, i.e., $P(Z_1^j) = P(Z^j)$, this recovered mapping is exactly $\mathbf{a}$ where each component is $a_j = F_{P_z}^{-1} \circ F_{P(Z^j)}$. Thus $f_\theta^{-1}\big(f_{\text{init}}(\mathbf{z}_1)\big) = \mathbf{a}(\mathbf{z}_1)$.

For $t \geq 2$: by Assumps. 2, 3, and 5, the history-conditioned contemporaneous SCM depends on the history only through the lagged parents $\mathrm{Pa}_{\mathcal{G}}^{<}(X_t^j)$. Under Assump. 7 (sufficient history encoding), the history embedding $\mathbf{h}_{\phi,t}$ serves as a sufficient representation of these lagged parents, which implies that the history-conditioned flow $f_\theta(\cdot \mid \mathbf{h}_{\phi,t})$ is a conditional CNF for the history-conditioned contemporaneous SCM, with $(f_\theta(\cdot \mid \mathbf{h}_{\phi,t}), P_z) \in \mathcal{T} \times \mathcal{P}_z$; the distributional match is provided by hypothesis. Prop. 1 then yields the component-wise invertible map $\mathbf{a}$, independent of time and history, with exactly the stated identity for almost every history. □

B.3 CAUSAL CONSISTENCY AT THE MODEL LEVEL

Corollary 1 (Causal consistency of CNF; Javaloy et al. 2023)**.** *If a causal NF $(f_\theta, P_z) \in \mathcal{T} \times \mathcal{P}_z$ isolates the exogenous variables of an SCM $\mathcal{M}$, then $\nabla_{\mathbf{x}} f_\theta^{-1}(\mathbf{x}) \equiv \mathbf{I} - \mathbf{A}$ and $\nabla_{\mathbf{z}} f_\theta(\mathbf{z}) \equiv \mathbf{I} + \sum_{n=1}^{\mathrm{diam}(\mathbf{A})} \mathbf{A}^n$, where $\mathbf{A}$ is the causal adjacency matrix of $\mathcal{M}$ and $\equiv$ denotes equality of Jacobian sparsity patterns. In other words, f_θ is causally consistent with the true data-generating process, $\mathcal{M}$.*

Corollary 2 (Causal consistency of TSCNF)**.** *Under the conditions of Theorem 1, for almost every history, a TSCNF is causally consistent with the contemporaneous structure of the window causal graph $\mathcal{G}$ in the following sense, where $\mathbf{A}^{=}$ denotes the contemporaneous adjacency matrix (the lagged structure $\mathbf{A}^{<}$ is respected by construction; see Remark 1):*

$$\nabla_{\mathbf{x}_t} f_\theta^{-1}\big(\mathbf{x}_t \mid \mathbf{h}_{\phi,t}\big) \equiv \mathbf{I} - \mathbf{A}^{=}, \qquad \nabla_{\mathbf{z}_t} f_\theta\big(\mathbf{z}_t \mid \mathbf{h}_{\phi,t}\big) \equiv \mathbf{I} + \textstyle\sum_{n=1}^{\mathrm{diam}(\mathbf{A}^{=})} \big(\mathbf{A}^{=}\big)^n.$$

Proof of Cor. 2. Fixing the history, TSCNF $f_\theta(\cdot \mid \mathbf{h}_{\phi,t})$ is a conditional CNF for the history-conditioned contemporaneous SCM $(f_{\bar{\mathbf{x}}}, P(\mathbf{Z}_t))$, with both in $\mathcal{T} \times \mathcal{P}_z$ (Lems. 1 and 2). By Thm. 1, the trained TSCNF isolates the exogenous variables of this system up to the time- and history-invariant component-wise map $\mathbf{a}$, i.e., $f_\theta^{-1}\big(f_{\bar{\mathbf{x}}}(\mathbf{z}_t) \mid \mathbf{h}_{\phi,t}\big) = \mathbf{a}(\mathbf{z}_t)$. Applying Cor. 1 to this static history-conditioned system, whose causal adjacency is the contemporaneous adjacency $\mathbf{A}^{=}$, yields the stated contemporaneous identities. □

Remark 1. *While the contemporaneous identities of Cor. 2 follow from identifiability, consistency with the lagged structure $\mathbf{A}^{<}$ is enforced architecturally. TSCNF accesses the history only through feature-wise masked embeddings*

$$h_{\phi,t}^j = E_\phi^j\big(\mathbf{M}^j \odot \mathbf{X}_{t-\tau:t-1}\big),$$

where $\mathbf{M}^j$ selects $\mathrm{Pa}_{\mathcal{G}}^{<}(X_t^j)$. Hence, for any lag $i \in \{1, \dots, \tau\}$ and any $(p, q) \in [d] \times [d]$,

$$\frac{\partial f_\theta^q(\mathbf{z}_t \mid \mathbf{h}_{\phi,t})}{\partial X_{t-i}^p} = \sum_{j=1}^{d} \frac{\partial f_\theta^q(\mathbf{z}_t \mid \mathbf{h}_{\phi,t})}{\partial h_{\phi,t}^j} \frac{\partial h_{\phi,t}^j}{\partial X_{t-i}^p},$$

and the second factor is zero whenever $(\mathbf{A}_i^{<})_{jp} = 0$, where $\mathbf{A}_i^{<}$ is the lag-i block of $\mathbf{A}^{<}$. Thus, any reduced-form lagged dependence of X_t^q on X_{t-i}^p must be routed through an embedding whose mask contains the corresponding lagged edge. Moreover, by feature-wise conditioning, the local mechanism for X_t^q depends on the history directly only through $h_{\phi,t}^q$. Therefore its direct lagged inputs are restricted to $\mathrm{Pa}_{\mathcal{G}}^{<}(X_t^q)$, while additional lagged influence may arise only through contemporaneous causal propagation within the current time slice.

Remark 2. *The contemporaneous identities in Cor. 2 hold for the idealized flow $f_\theta \in \mathcal{T}$, i.e., a single effective TMI map. As in static CNFs [Javaloy et al., 2023], such causal consistency can be ensured by design via an abductive single-flow-layer construction, or promoted via a causal-consistency regularizer as a practical safeguard when multiple flow layers are stacked. The practical multi-layer case is discussed at the estimand level in Appendix B.4.*

B.4 CAUSAL CONSISTENCY AT THE ESTIMAND LEVEL

TSCNF may use multiple NF layers to flexibly model complex data distributions. In static CNFs, stacking flow layers can introduce shortcut dependencies from ancestors to descendants, which violates the causal-consistency property of Cor. 1. Javaloy et al. [2023] address this in practice with an abductive single-layer construction or a causal-consistency regularizer. We view that regularizer as a practical safeguard for learned multi-layer flows, not as something contradicted by our approach.

Instead, we provide a complementary estimand-level observation: when the input graph is known, redundant forward edges that respect the same topological order need not change the causal effect computed from the same observational distribution via truncated-factorization [Pearl, 2009]. We formalize this in the following proposition.

Proposition 2. *Let $\mathcal{G}$ be a causal graph with causal sufficiency, and let $\mathcal{G}'$ be a DAG obtained by adding only forward edges that are consistent with a topological order of $\mathcal{G}$. If the observational distribution $P_{\mathcal{G}}$ is Markov with respect to $\mathcal{G}$ and satisfies positivity, then the truncated-factorization causal effect $P(\mathbf{X} \setminus \mathbf{A} \mid do(\mathbf{a}))$ computed from the same observational distribution under $\mathcal{G}'$ agrees with that under $\mathcal{G}$:*

$$P_{\mathcal{G}}(\mathbf{X} \setminus \mathbf{A} \mid do(\mathbf{a})) = P_{\mathcal{G}'}(\mathbf{X} \setminus \mathbf{A} \mid do(\mathbf{a}))$$

Proof of Prop. 2. Let $\prec$ be a topological order consistent with $\mathcal{G}$, and let $\mathcal{G}'$ be obtained by adding only edges that point forward in this order. Then $\mathbf{Pa}^i_{\mathcal{G}} \subseteq \mathbf{Pa}^i_{\mathcal{G}'} \subseteq \{X^k : X^k \prec X^i\}$ for every variable X^i. Since $P_{\mathcal{G}}$ is Markov with respect to $\mathcal{G}$, the ordered local Markov property gives

$$X^i \perp \{X^k : X^k \prec X^i\} \setminus \mathbf{Pa}^i_{\mathcal{G}} \mid \mathbf{Pa}^i_{\mathcal{G}}. \tag{11}$$

Therefore the additional parents in $\mathbf{Pa}^i_{\mathcal{G}'} \setminus \mathbf{Pa}^i_{\mathcal{G}}$ do not change the conditional distribution induced by the same observational law:

$$P_{\mathcal{G}}(X^i \mid \mathbf{Pa}^i_{\mathcal{G}'}) = P_{\mathcal{G}}(X^i \mid \mathbf{Pa}^i_{\mathcal{G}}). \tag{12}$$

Applying the truncated-factorization [Pearl, 2009] and evaluating any intervened parents at their assigned values $\mathbf{a}$, we obtain

$$P_{\mathcal{G}'}(\mathbf{X} \setminus \mathbf{A} \mid do(\mathbf{a})) = \prod_{X^i \in \mathbf{X} \setminus \mathbf{A}} P_{\mathcal{G}}(X^i \mid \mathbf{Pa}^i_{\mathcal{G}'})\big|_{\mathbf{A}=\mathbf{a}} \tag{13}$$

$$= \prod_{X^i \in \mathbf{X} \setminus \mathbf{A}} P_{\mathcal{G}}(X^i \mid \mathbf{Pa}^i_{\mathcal{G}})\big|_{\mathbf{A}=\mathbf{a}} \tag{14}$$

$$= P_{\mathcal{G}}(\mathbf{X} \setminus \mathbf{A} \mid do(\mathbf{a})). \tag{15}$$

Thus, redundant same-order forward edges do not change the causal estimand computed from the same observational distribution via the truncated-factorization. □

Remark 3. *Prop. 2 establishes estimand-level identifiability when the given graph is a supergraph of the ground-truth causal graph that preserves the causal order. In particular, redundant forward edges do not change the g-formula estimand once the observational distribution and causal order are fixed. This complements the model-level causal-consistency guarantees of Javaloy et al. [2023]. Empirical evidence is provided in Appendices D.11 and D.12.*

C APPENDIX FOR METHODS

C.1 DETAILS OF INFERENCE WITH TSCNF

In this subsection, we describe how our model answers various types of probabilistic and causal queries through Monte Carlo sampling. We first introduce the notations and the main queries to discuss. Then, we provide detailed sampling procedures for each inference type supported by TSCNF.

Notations Let $\mathbf{x}_{1:t_0-1} \in \mathbb{R}^{M \times (t_0-1) \times d}$ denote a batch of past trajectories, where d is the number of variables and M is the batch size. We denote the learned causal history encoder by E_ϕ, which produces a context embedding $\mathbf{h}_{\phi,t_0}$. The learned causal conditional normalizing flow is denoted by $f_\theta(\cdot \mid \mathbf{h}_{\phi,t_0})$, with a tractable inverse f_θ^{-1}, and we write $\hat{\mathbf{x}}_{t_0} = f_\theta(\mathbf{z}_{t_0} \mid \mathbf{h}_{\phi,t_0})$ for the sample generated by the flow. We write $\mathbf{z}_{t_0} \sim P_z$ to denote samples from the base distribution. Interventions on variable i are denoted as $do(X^i_{t_0} = a)$. Lastly, we use superscripts $(\cdot)$ for indexing within a batch, e.g., $\mathbf{h}^{(m)}_{\phi,t_0}$ denotes the context embedding for the m-th trajectory up to $t_0 - 1$ time steps.

Below, we present six sampling procedures: marginal observational, marginal interventional, history-adjusted observational, history-adjusted interventional inference, autoregressive inference, and inference with multiple interventions.

Marginal observational inference To estimate $P(\mathbf{X}_{t_0})$, we first encode each trajectory's recent history using a causal mask that captures variable-wise temporal dependencies, producing context vectors $\mathbf{h}_{\phi,t_0}^{(m)}$. For each $m = 1, \ldots, M$, we sample latent variables $\mathbf{z}_{t_0}^{(m)}$ from the base distribution and generate samples via the conditional flow using the corresponding context. These samples are then appended to the history and can be used recursively for autoregressive inference. The full procedure is shown in Algorithm 2.

Marginal interventional inference To estimate $P(\mathbf{X}_{t_0} \setminus \{X_{t_0}^i\} \mid do(X_{t_0}^i = a))$, we begin by encoding each trajectory's recent history to obtain context vectors $\mathbf{h}_{\phi,t_0}^{(m)}$ with a causal mask capturing variable-specific temporal dependencies. Subsequently, we sample noise from the base distribution and decode through the conditional flow with the context embedding to generate observational samples. We then overwrite the i-th variable with the intervention value a, treating it as externally assigned. This modified sample is inverted through the conditional flow to compute the corresponding latent. To preserve the original randomness, we retain the intervened latent dimension and substitute all non-intervened dimensions with their original noise counterparts. Finally, we decode the updated latent to produce interventional samples. See Algorithm 3 for details.

History-adjusted observational inference In some scenarios, we are interested in modeling the conditional distribution $P(\mathbf{X}_{t_0} \mid \mathbf{h}_{\phi,t_0})$ given a fixed observed history. In this case, we first encode a single observed trajectory with a causal mask into a fixed context vector $\mathbf{h}_{\phi,t_0}$ and then sample multiple $\mathbf{z}_{t_0}^{(m)}$ to generate corresponding samples via the conditional flow. These samples form an empirical approximation of the conditional distribution. The method is outlined in Algorithm 4.

History-adjusted interventional inference To estimate $P(\mathbf{X}_{t_0} \setminus \{X_{t_0}^i\} \mid do(X_{t_0}^i = a), \mathbf{h}_{\phi,t_0})$, we first encode the observed history with a causal mask capturing variable-specific temporal dependencies, to obtain a fixed context embedding. Then, we sample noise from the base distribution and decode it through the conditional flow—conditioned on the history embedding—to generate observational samples. We then apply the intervention by replacing the i-th variable with the desired value a. The modified sample is inverted through the conditional flow to obtain its corresponding latent. We preserve only the intervened component of this new latent and combine it with the original noise for other dimensions. Finally, we decode this adjusted latent under the same fixed context to produce interventional samples that reflect the effect of the intervention given the specific trajectory. See Algorithm 5.

Autoregressive inference All of the above procedures can be repeated in an autoregressive fashion. After each generation step (either observational or interventional), the new sample is appended to the current history. The encoder then processes the updated history to produce the context vector $\mathbf{h}_{\phi,t_0+1}$ for the next time step $t_0 + 1$. This allows for multi-step inference under arbitrary intervention strategies.

Inference with multiple interventions Our framework supports both single-variable interventions and simultaneous interventions on multiple variables as detailed in Algorithm 6. For valid causal inference, it is important that interventions respect the *topological order* of the causal graph, proceeding from ancestor variables to their descendants. Executing interventions in this order guarantees that the causal influence of an ancestor's intervention is properly propagated before assigning values to its descendants, ensuring that every intervention remains valid and as intended. In particular, overwriting each target variable with its corresponding intervention value should proceed sequentially according to the topological order of the causal graph. Once a variable has been assigned its intervened latent state, it remains fixed in all subsequent operations, with already intervened dimensions excluded from noise substitution to preserve their assigned values. This procedure ensures that each intervention remains valid and faithfully represented in the generated outcomes.

C.2 DETAILS OF SAMPLING ALGORITHM

In this subsection, we provide a detailed instantiation of Alg. 1 based on the model architecture and flow operations of TSCNF. This procedure builds on the principles of single-step sampling operations introduced earlier, extending them to multi-step interventional inference over future time steps. We begin with an observed history $\mathbf{h} = \bar{\mathbf{x}}_{t-1}$ and simulate forward under a sequence of interventions $\{do(a_t), \ldots, do(a_{t+k-1})\}$. At each step $j = 0, \ldots, k-1$, the following operations are performed:

- **Causal context encoding:** At each step, the current history is encoded with a causal mask to produce a context embedding, capturing variable-specific time-lagged dependencies.

Algorithm 2 Marginal Observational Inference

Input: Batch $\mathbf{x}_{1:t_0-1} \in \mathbb{R}^{M\times(t_0-1)\times d}$
Output: Empirical distribution $\hat{P}(\mathbf{X}_{t_0})$

Initialize empty sample list $\mathcal{S} \leftarrow [\,]$
for $m = 1$ to M **do**
  $\mathbf{h}_{\phi,t_0}^{(m)} \leftarrow E_\phi(\mathbf{M} \odot \mathbf{x}_{t_0-\tau:t_0-1}^{(m)})$
  $\mathbf{z}_{t_0}^{(m)} \sim P_z$
  $\hat{\mathbf{x}}_{t_0}^{(m)} \leftarrow f_\theta(\mathbf{z}_{t_0}^{(m)} \mid \mathbf{h}_{\phi,t_0}^{(m)})$
  Append $\hat{\mathbf{x}}_{t_0}^{(m)}$ to sample set $\mathcal{S}$
return empirical distribution from $\mathcal{S}$

Algorithm 3 Marginal Interventional Inference

Input: Batch $\mathbf{x}_{1:t_0-1} \in \mathbb{R}^{M\times(t_0-1)\times d}$, Intervention $do(X_{t_0}^i = a)$
Output: Empirical distribution $\hat{P}(\mathbf{X}_{t_0} \setminus \{X_{t_0}^i\} \mid do(X_{t_0}^i = a))$

Initialize empty sample list $\mathcal{S} \leftarrow [\,]$
for $m = 1$ to M **do**
  $\mathbf{h}_{\phi,t_0}^{(m)} \leftarrow E_\phi(\mathbf{M} \odot \mathbf{x}_{t_0-\tau:t_0-1}^{(m)})$
  $\mathbf{z}_{t_0}^{(m)} \sim P_z$
  $\hat{\mathbf{x}}_{t_0}^{(m)} \leftarrow f_\theta(\mathbf{z}_{t_0}^{(m)} \mid \mathbf{h}_{\phi,t_0}^{(m)})$
  $\hat{\mathbf{x}}_{t_0}^{(m),i} \leftarrow a$
  $\tilde{\mathbf{z}}_{t_0}^{(m)} \leftarrow f_\theta^{-1}(\hat{\mathbf{x}}_{t_0}^{(m)} \mid \mathbf{h}_{\phi,t_0}^{(m)})$
  Replace $\tilde{\mathbf{z}}_{t_0}^{(m),j} \leftarrow \mathbf{z}_{t_0}^{(m),j}$ for $j \neq i$
  $\tilde{\mathbf{x}}_{t_0}^{(m)} \leftarrow f_\theta(\tilde{\mathbf{z}}_{t_0}^{(m)} \mid \mathbf{h}_{\phi,t_0}^{(m)})$
  Append $\tilde{\mathbf{x}}_{t_0}^{(m)}$ to sample set $\mathcal{S}$
return empirical distribution from $\mathcal{S}$

- **Intervention:** A latent sample is drawn from the base distribution and decoded. To apply the intervention, the relevant latent dimension is adjusted so that the generated variable matches the intervention target.
- **Sample generation and history update:** The modified latent is passed through the conditional flow to produce the full sample, which is then appended to the history along with the applied action.

After k such steps, the final context is computed from the updated history, and a sample of $\mathbf{x}_{t+k}$ is generated. Repeating this process M times yields an empirical approximation to the interventional distribution, as detailed in Algorithm 7 for the sequential single-variable intervention case.

A Unified Framework for Interventional Inference While Algorithm 7 focuses on sequential single-variable interventions, we can further generalize this procedure to encompass all previously discussed inference types into a single, comprehensive framework. Algorithm 8 presents this generalized Monte Carlo inference methodology, naturally subsuming all other algorithms as its special cases. At its core, it flexibly supports an arbitrary multi-step prediction horizon $k \geq 0$, allows conditioning on any initial history $\mathbf{h}$ (either a single specific trajectory or a population batch), and accommodates dynamically varying joint intervention sets $\mathcal{I}_{t_0+j} \subseteq [d]$ at each simulation step $j = 0, \ldots, k$, where the final-step set $\mathcal{I}_{t_0+k}$ defaults to $\varnothing$.

By systematically constraining these three axes—the horizon k, the intervention set $\mathcal{I}_{t_0+j}$, and the initial history—this unified algorithm elegantly reduces to all specialized queries introduced earlier. For instance, limiting the temporal horizon to $k = 0$ without applying any intervention ($\mathcal{I}_{t_0} = \varnothing$) immediately recovers standard observational inferences (Algorithms 2 and 4). Extending the intervention to a single target variable yields instantaneous single-intervention inferences (Algorithms 3 and 5), while intervening on multiple targets simultaneously corresponds to joint instantaneous inference (Algorithm 6). Finally, applying a predefined sequence of single-variable interventions across consecutive future steps with the final step left observational ($\mathcal{I}_{t_0+k} = \varnothing$) directly mirrors the multi-step prediction strategy detailed above (Algorithm 7). In summary, Algorithm 8 provides a highly versatile methodology capable of executing arbitrarily complex, time-varying interventional strategies over extended horizons.

Algorithm 4 History-Adjusted Observational Inference

Input: Single history $\mathbf{x}_{1:t_0-1} \in \mathbb{R}^{1\times(t_0-1)\times d}$
Output: Empirical distribution $\hat{P}(\mathbf{X}_{t_0} \mid \mathbf{h}_{\phi,t_0})$

Initialize empty sample list $\mathcal{S} \leftarrow [\,]$
$\mathbf{h}_{\phi,t_0} \leftarrow E_\phi(\mathbf{M} \odot \mathbf{x}_{t_0-\tau:t_0-1})$
for $m = 1$ to M **do**
  $\mathbf{z}_{t_0}^{(m)} \sim P_z$
  $\hat{\mathbf{x}}_{t_0}^{(m)} \leftarrow f_\theta(\mathbf{z}_{t_0}^{(m)} \mid \mathbf{h}_{\phi,t_0})$
  Append $\hat{\mathbf{x}}_{t_0}^{(m)}$ to sample set $\mathcal{S}$
return empirical distribution from $\mathcal{S}$

Algorithm 5 History-Adjusted Interventional Inference

Input: Single history $\mathbf{x}_{1:t_0-1}$, Intervention $do(X_{t_0}^i = a)$
Output: Empirical distribution $\hat{P}(\mathbf{X}_{t_0} \setminus \{X_{t_0}^i\} \mid do(X_{t_0}^i = a), \mathbf{h}_{\phi,t_0})$

Initialize empty sample list $\mathcal{S} \leftarrow [\,]$
$\mathbf{h}_{\phi,t_0} \leftarrow E_\phi(\mathbf{M} \odot \mathbf{x}_{t_0-\tau:t_0-1})$
for $m = 1$ to M **do**
  $\mathbf{z}_{t_0}^{(m)} \sim P_z$
  $\hat{\mathbf{x}}_{t_0}^{(m)} \leftarrow f_\theta(\mathbf{z}_{t_0}^{(m)} \mid \mathbf{h}_{\phi,t_0})$
  $\hat{\mathbf{x}}_{t_0}^{(m),i} \leftarrow a$
  $\tilde{\mathbf{z}}_{t_0}^{(m)} \leftarrow f_\theta^{-1}(\hat{\mathbf{x}}_{t_0}^{(m)} \mid \mathbf{h}_{\phi,t_0})$
  Replace $\tilde{\mathbf{z}}_{t_0}^{(m),j} \leftarrow \mathbf{z}_{t_0}^{(m),j}$ for $j \neq i$
  $\tilde{\mathbf{x}}_{t_0}^{(m)} \leftarrow f_\theta(\tilde{\mathbf{z}}_{t_0}^{(m)} \mid \mathbf{h}_{\phi,t_0})$
  Append $\tilde{\mathbf{x}}_{t_0}^{(m)}$ to sample set $\mathcal{S}$
return empirical distribution from $\mathcal{S}$

D APPENDIX FOR EXPERIMENTS

D.1 METRICS

Root Mean Square Error (RMSE; Li et al. 2021) We compute RMSE to evaluate the point estimation performance. Given M Monte Carlo samples from a conditional distribution adjusted by a specific observed history $\mathbf{h}_{t,i}$, we define the point estimate at time t for individual i as the sample average:

$$\hat{\mathbf{x}}_{t,i} = \frac{1}{M}\sum_{k=1}^{M} \mathbf{x}_{t,i}^{(k)}, \quad \mathbf{x}_{t,i}^{(k)} \overset{\text{iid}}{\sim} P\left(\mathbf{X}_t \mid \mathbf{h}_{t,i}\right) \tag{16}$$

Given the observed history up to time m for each of the N trajectories, we evaluate RMSE over k-step-ahead prediction horizon ($t = m+1, \ldots, m+k$). Let the predicted and ground truth trajectories be represented as:

$$\mathbf{x}_{m+1:m+k} \in \mathbb{R}^{N\times k\times d} \tag{17}$$

Table 4: Relationship between specific inference algorithms and the generalized Monte Carlo framework (Algorithm 8). The generalized framework reduces to each specialized variant by imposing specific constraints on the prediction horizon length (k), the time-varying intervention set ($\mathcal{I}_{t_0+j}$), and the initial history ($\mathbf{h}$).

Algorithm	**Query Type**	**Horizon** (k)	**Intervention Set** ($\mathcal{I}_{t_0+j}$)	**Initial History** ($\mathbf{h}$)	**Special Case Reduction**
Algorithm 2 (Marginal Obs.)	Observational	$k=0$	$\varnothing$ $(j=0)$	Batch population	No intervention, single step estimation
Algorithm 3 (Marginal Interv.)	Interventional	$k=0$	$\{i\}$ $(j=0)$	Batch population	Instantaneous single-variable intervention
Algorithm 4 (HA Obs.)	Observational	$k=0$	$\varnothing$ $(j=0)$	Single specific trajectory	Conditioned on specific history, no intervention
Algorithm 5 (HA Interv.)	Interventional	$k=0$	$\{i\}$ $(j=0)$	Single specific trajectory	Specific history, single instantaneous intervention
Algorithm 6 (Joint Interv.)	Interventional	$k=0$	$\mathcal{I}_{t_0} \subseteq [d]$ $(j=0)$	Batch population	Instantaneous multiple-variable interventions
Algorithm 7 (Sequential Interv.)	Interventional	$k \geq 1$	$\{i\}$ $(j=0,\ldots,k-1)$, $\mathcal{I}_{t_0+k}=\varnothing$	Single specific trajectory	Sequential single-variable interventions ($\mathcal{I}_{t_0+j}=\{A\}$)
Algorithm 8 (General)	**Interventional**	$\mathbf{k \geq 0}$	$\mathcal{I}_{\mathbf{t_0+j}} \subseteq [\mathbf{d}]$ $(j=0,\ldots,k)$	**General**	**Unifies all constraints across time**

Algorithm 6 Joint Interventional Inference (Multiple Interventions)

Input: Batch $\mathbf{x}_{1:t_0-1} \in \mathbb{R}^{M\times(t_0-1)\times d}$,
Intervention set $\mathcal{I} \subseteq \{1,\dots,d\}$ with target values $\{a_i\}_{i\in\mathcal{I}}$,
Topological order $\pi(\mathcal{I}) = (i_1,\dots,i_{|\mathcal{I}|})$ over $\mathcal{I}$
Output: Empirical distribution $\hat{P}(\mathbf{X}_{t_0} \setminus \{X^i_{t_0}\}_{i\in\mathcal{I}} \mid do(\{X^i_{t_0} = a_i\}_{i\in\mathcal{I}}))$

Initialize empty sample list $\mathcal{S} \leftarrow [\,]$
for $m = 1$ to M **do**
  $\mathbf{h}^{(m)}_{\phi,t_0} \leftarrow E_\phi(\mathbf{M} \odot \mathbf{x}^{(m)}_{t_0-\tau:t_0-1})$
  $\mathbf{z}^{(m)}_{t_0} \sim P_z$
  $\hat{\mathbf{x}}^{(m)}_{t_0} \leftarrow f_\theta(\mathbf{z}^{(m)}_{t_0} \mid \mathbf{h}^{(m)}_{\phi,t_0})$
  $\mathcal{J} \leftarrow \varnothing$
  for each $i \in \pi(\mathcal{I})$ **do**
    $\hat{\mathbf{x}}^{(m),i}_{t_0} \leftarrow a_i$
    $\tilde{\mathbf{z}}^{(m)}_{t_0} \leftarrow f^{-1}_\theta(\hat{\mathbf{x}}^{(m)}_{t_0} \mid \mathbf{h}^{(m)}_{\phi,t_0})$
    Replace $\tilde{\mathbf{z}}^{(m),j}_{t_0} \leftarrow \mathbf{z}^{(m),j}_{t_0}$ for all $j \in [d] \setminus (\mathcal{J} \cup \{i\})$
    $\tilde{\mathbf{x}}^{(m)}_{t_0} \leftarrow f_\theta(\tilde{\mathbf{z}}^{(m)}_{t_0} \mid \mathbf{h}^{(m)}_{\phi,t_0})$
    $\hat{\mathbf{x}}^{(m)}_{t_0} \leftarrow \tilde{\mathbf{x}}^{(m)}_{t_0}$
    $\mathcal{J} \leftarrow \mathcal{J} \cup \{i\}$
  Append $\tilde{\mathbf{x}}^{(m)}_{t_0}$ to sample set $\mathcal{S}$
return empirical distribution from $\mathcal{S}$

then the RMSE is computed as:

$$\mathrm{RMSE}(\hat{\mathbf{x}}, \mathbf{x}) = \sqrt{\frac{1}{Nkd}\sum_{i=1}^{N}\sum_{t=m+1}^{m+k}\sum_{j=1}^{d}\left(x^{(j)}_{t,i} - \hat{x}^{(j)}_{t,i}\right)^2}. \tag{18}$$

Maximum Mean Discrepancy (MMD; Gretton et al. 2012) Maximum Mean Discrepancy (MMD) is a non-parametric metric for quantifying the difference between two probability distributions based on samples. It measures the distance between the mean embeddings of the distributions. A larger MMD value indicates a greater difference between the distributions. Let $\mathbf{X} = \{\mathbf{x}_1,\dots,\mathbf{x}_{n_x}\}$ and $\mathbf{Y} = \{\mathbf{y}_1,\dots,\mathbf{y}_{n_y}\}$ be samples drawn from two distributions P and Q, respectively. Then the squared MMD between P and Q in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ associated with a kernel $k(\cdot,\cdot)$ is defined as:

$$\mathrm{MMD}^2(P,Q) = \|\mu_P - \mu_Q\|^2_{\mathcal{H}} \tag{19}$$

where $\mu_P = \mathbb{E}_{\mathbf{x}\sim P}\big[k(\mathbf{x},\cdot)\big]$ and $\mu_Q = \mathbb{E}_{\mathbf{y}\sim Q}\big[k(\mathbf{y},\cdot)\big]$ denote the mean embeddings of P and Q in $\mathcal{H}$. This quantity can be estimated empirically from finite samples. For computational simplicity, we use the following estimator:

$$\mathrm{MMD}^2(\mathbf{X},\mathbf{Y}) = \frac{1}{n_x^2}\sum_{i=1}^{n_x}\sum_{j=1}^{n_x} k\left(\mathbf{x}^{(i)},\mathbf{x}^{(j)}\right) + \frac{1}{n_y^2}\sum_{i=1}^{n_y}\sum_{j=1}^{n_y} k\left(\mathbf{y}^{(i)},\mathbf{y}^{(j)}\right) - \frac{2}{n_x n_y}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y} k\left(\mathbf{x}^{(i)},\mathbf{y}^{(j)}\right). \tag{20}$$

For the kernel choice, we adopt the Gaussian Radial Basis Function (RBF) kernel, which is defined as follows:

$$k\left(\mathbf{x}^{(i)},\mathbf{y}^{(j)}\right) = \exp\left(-\frac{\left\|\mathbf{x}^{(i)} - \mathbf{y}^{(j)}\right\|^2}{2\sigma^2}\right), \tag{21}$$

where σ is a bandwidth parameter that controls how sensitive the kernel is to differences between samples. We set the σ heuristically as the median distance between points in the aggregate samples following prior works [Gretton et al., 2012].

1-dimensional Wasserstein Distance [Ramdas et al., 2017] The 1-dimensional Wasserstein distance quantifies distributional shifts by measuring the minimum cost of transforming one distribution into another via mass transport. It captures

Algorithm 7 Monte Carlo Estimation of $P(\mathbf{X}_{t+k} \mid do(a_t), \ldots, do(a_{t+k-1}), \mathbf{h})$ (Implementation version)

Input: History $\mathbf{h} = \bar{\mathbf{x}}_{t-1}$, interventions $\{a_t, \ldots, a_{t+k-1}\}$ on treatment variable A (coordinate i), number of samples M
Output: Empirical distribution over $\mathbf{X}_{t+k}$

Initialize empty sample list $\mathcal{S} \leftarrow [\,]$
for $m = 1$ to M **do**
  Initialize history $\mathbf{h}_0^{(m)} \leftarrow \mathbf{h}$ {Start with observed history up to $t-1$}
  for $j = 0$ to $k-1$ **do**
    $\mathbf{h}_{\phi,t+j}^{(m)} \leftarrow E_\phi(\mathbf{M} \odot \mathbf{h}_j^{(m)})$ {Encode history using causal graph masking}
    $\mathbf{z}_{t+j}^{(m)} \sim P_z$ {Sample latent variable}
    $\hat{\mathbf{x}}_{t+j}^{(m)} \leftarrow f_\theta(\mathbf{z}_{t+j}^{(m)} \mid \mathbf{h}_{\phi,t+j}^{(m)})$ {Generate observational sample}
    $\hat{x}_{t+j}^{i,(m)} \leftarrow a_{t+j}$ {Apply intervention on coordinate i}
    $\tilde{\mathbf{z}}_{t+j}^{(m)} \leftarrow f_\theta^{-1}(\hat{\mathbf{x}}_{t+j}^{(m)} \mid \mathbf{h}_{\phi,t+j}^{(m)})$ {Invert flow for intervened value}
    Replace $\tilde{\mathbf{z}}_{t+j}^{(m)}$ with $\mathbf{z}_{t+j}^{(m)}$ except for dimension i {Update latent with intervention}
    $\tilde{\mathbf{x}}_{t+j}^{(m)} \leftarrow f_\theta(\tilde{\mathbf{z}}_{t+j}^{(m)} \mid \mathbf{h}_{\phi,t+j}^{(m)})$ {Generate intervened sample via conditional flow}
    $\mathbf{h}_{j+1}^{(m)} \leftarrow [\mathbf{h}_j^{(m)} \| \tilde{\mathbf{x}}_{t+j}^{(m)}]$ {Update history}
  $\mathbf{h}_{\phi,t+k}^{(m)} \leftarrow E_\phi(\mathbf{M} \odot \mathbf{h}_k^{(m)})$ {Context embedding for final step}
  $\mathbf{z}_{t+k}^{(m)} \sim P_z$ {Sample base noise}
  $\hat{\mathbf{x}}_{t+k}^{(m)} \leftarrow f_\theta(\mathbf{z}_{t+k}^{(m)} \mid \mathbf{h}_{\phi,t+k}^{(m)})$ {Generate final sample}
  Append $\hat{\mathbf{x}}_{t+k}^{(m)}$ to sample set $\mathcal{S}$
return empirical distribution from $\mathcal{S}$

differences in both location and shape between two probabilistic distributions. Given $X \in \mathbb{R}^{n_x}$ and $Y \in \mathbb{R}^{n_y}$, let $\hat{U}(t)$ and $\hat{V}(t)$ denote empirical cumulative distribution functions (CDFs) of X,Y respectively. The 1-dimensional Wasserstein Distance can be computed in two ways, depending on whether the sample sizes are equal. If the samples are of equal size $n_x = n_y$:

$$W_1(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \left| x^{(i)} - y^{(i)} \right|. \tag{22}$$

where $x^{(i)}$ and $y^{(i)}$ denote the i-th order statistics of X and Y, i.e., the sorted samples. Sorting aligns identical quantiles between the two empirical CDFs, which achieves the optimal transport cost in one dimension.

If the samples differ in size (i.e., $n_x \neq n_y$), the Wasserstein distance can be computed using the integral of the absolute difference between the empirical CDFs:

$$W_1(X, Y) = \int_{-\infty}^{+\infty} \left| \widehat{U}(t) - \widehat{V}(t) \right| dt. \tag{23}$$

In our evaluation, we primarily employ the empirical formulation for equal-sized samples, as it is computationally efficient and well-suited for comparing generated interventional samples and ground-truth samples. The Wasserstein distance provides an interpretable measure of distance even when the two distributions do not overlap perfectly and can capture differences in both central tendency and distributional shape.

Mean Distance [Wu et al., 2024] Mean distance is a simple metric that computes the absolute difference between the empirical means of two samples. Given two samples $X \in \mathbb{R}^{n_x}$ and $Y \in \mathbb{R}^{n_y}$, mean distance is defined as:

$$\text{Mean Dist.} = \left| \frac{1}{n_x} \sum_{i=1}^{n_x} x^{(i)} - \frac{1}{n_y} \sum_{i=1}^{n_y} y^{(i)} \right| \tag{24}$$

D.2 DATASET

Synthetic dataset for model ablation To evaluate the contributions of each architectural component while fully leveraging the model's ability to handle a wide range of causal queries (e.g., multiple treatments and multiple outcomes), we conduct

Algorithm 8 General Monte Carlo Estimation of $P\big(\mathbf{X}_{t_0+k} \mid \{do(X^i_{t_0+j} = a_{i,\,t_0+j})_{i\in\mathcal{I}_{t_0+j}}\}_{j=0}^{k},\ \mathbf{h}\big)$

Input: History $\mathbf{h} = \bar{\mathbf{x}}_{t_0-1}$, time-varying intervention sets $\big\{\mathcal{I}_{t_0+j} \subseteq [d],\ \{a_{i,t_0+j}\}_{i\in\mathcal{I}_{t_0+j}}\big\}_{j=0}^{k}$ with topological orders $\pi(\mathcal{I}_{t_0+j})$, prediction horizon $k \geq 0$ (with $\mathcal{I}_{t_0+k} = \varnothing$ unless specified), number of samples M
Output: Empirical distribution over $\mathbf{X}_{t_0+k} \setminus \{X^i_{t_0+k}\}_{i\in\mathcal{I}_{t_0+k}}$

Initialize empty sample list $\mathcal{S} \leftarrow [\,]$
for $m = 1$ to M **do**
  $\mathbf{h}_0^{(m)} \leftarrow \mathbf{h}$
  for $j = 0$ to k **do**
    $\mathbf{h}^{(m)}_{\phi,\,t_0+j} \leftarrow E_\phi\big(\mathbf{M} \odot \mathbf{h}_j^{(m)}\big)$ {Encode history}
    $\mathbf{z}^{(m)}_{t_0+j} \sim P_z$
    $\hat{\mathbf{x}}^{(m)}_{t_0+j} \leftarrow f_\theta\big(\mathbf{z}^{(m)}_{t_0+j} \mid \mathbf{h}^{(m)}_{\phi,\,t_0+j}\big)$ {Observational sample}
    $\tilde{\mathbf{x}}^{(m)}_{t_0+j} \leftarrow \hat{\mathbf{x}}^{(m)}_{t_0+j}$
    $\mathcal{J} \leftarrow \varnothing$
    for each $i \in \pi(\mathcal{I}_{t_0+j})$ **do**
      $\tilde{x}^{i,\,(m)}_{t_0+j} \leftarrow a_{i,\,t_0+j}$ {Apply intervention}
      $\tilde{\mathbf{z}}^{(m)}_{t_0+j} \leftarrow f_\theta^{-1}\big(\tilde{\mathbf{x}}^{(m)}_{t_0+j} \mid \mathbf{h}^{(m)}_{\phi,\,t_0+j}\big)$
      Replace $\tilde{z}^{\ell,\,(m)}_{t_0+j} \leftarrow z^{\ell,\,(m)}_{t_0+j}$ for all $\ell \in [d] \setminus (\mathcal{J} \cup \{i\})$
      $\tilde{\mathbf{x}}^{(m)}_{t_0+j} \leftarrow f_\theta\big(\tilde{\mathbf{z}}^{(m)}_{t_0+j} \mid \mathbf{h}^{(m)}_{\phi,\,t_0+j}\big)$
      $\mathcal{J} \leftarrow \mathcal{J} \cup \{i\}$
    if $j < k$ **then** $\mathbf{h}^{(m)}_{j+1} \leftarrow [\mathbf{h}^{(m)}_j \| \tilde{\mathbf{x}}^{(m)}_{t_0+j}]$ {Update history}
  Append $\tilde{\mathbf{x}}^{(m)}_{t_0+k}$ to $\mathcal{S}$
return empirical distribution from $\mathcal{S}$

ablation experiments on fully synthetic time-series data with the following DGP in Runge [2020]:

$$X^j_t = w_j X^j_{t-1} + \sum_{i=1}^{3} c_i f_i\left(X^{\pi(i)}_{t-\tau_i}\right) + \eta^j_t, \tag{25}$$

where w_j governs the coefficient for variable j, each c_i scales the function f_i over its parent $\pi(i)$ at lag τ_i, and $\eta^j_t \sim \mathcal{N}(0, \sigma^2)$ is Gaussian noise. We simulate a dataset with randomly generated window causal graph with six variables $\mathbf{X}_t = \big(X^1_t, X^2_t, X^3_t, X^4_t, X^5_t, X^6_t\big)$ and $\tau = 3$ as shown in Fig. 6. The treatment variables at time t are the second and fourth components, i.e., $\mathbf{A}_t = \big(X^2_t,\ X^4_t\big)$. We consider both single interventions (on either $do(X^2_t)$ or $do(X^4_t)$) and multiple interventions ($do(X^2_t, X^4_t)$) where both variables are intervened simultaneously.

Synthetic dataset for model comparison To evaluate the model's performance, we generate fully synthetic time-series data based on both linear and non-linear structural equation models, adapted from Robins et al. [1999] and Wu et al. [2024]. At each time step t, we observe a scalar covariate $X_t \in \mathbb{R}$, a scalar treatment $A_t \in \mathbb{R}$, and a scalar outcome $Y_t \in \mathbb{R}$. Let τ denote the size of maximal temporal dependency (i.e., the size of window causal graph). The data generating process (DGP) is defined as follows:

$$X_0 \sim \mathrm{Uniform}(0, 1)$$

$$f(x) = \begin{cases} x, & \text{(linear)} \\ x + 5x^2 e^{-x^2/20}, & \text{(non-linear)} \end{cases}$$

$$X_t = \gamma_0 + \sum_{k=t-\tau}^{t-1} \gamma_{t-k}\, f\big(A_k\big) + \sum_{k=t-\tau}^{t-1} \gamma_{\tau+t-k}\, f\big(X_k\big) + \varepsilon_{X,t}$$

$$A_t = \beta_0 + \sum_{k=t-\tau}^{t-1} \beta_{t-k}\, f\big(A_k\big) + \sum_{k=t-\tau}^{t} \beta_{\tau+t-k+1}\, f\big(X_k\big) + \varepsilon_{A,t}$$

$$Y_t = \alpha_0 + \sum_{k=t-\tau}^{t} \alpha_{t-k+1}\, f\big(A_k\big) + \sum_{k=t-\tau}^{t} \alpha_{\tau+t-k+2}\, f\big(X_k\big) + \varepsilon_{Y,t},$$

where $\varepsilon_{X,t}, \varepsilon_{A,t}, \varepsilon_{Y,t}$ are i.i.d. Gaussian noise terms. In our experiments we fix the same graph structure in Fig. 7a, but swap the linear/non-linear form of f, allowing us to evaluate model performance with varying complexity of functions. We employ both variants (linear and non-linear) for answering both marginal and history-adjusted queries.

Synthetic dataset for model comparison (sparse confounding structure) To further evaluate model performance with other baselines under different graph settings, we generate non-linear time-series data with sparser causal dependencies. We adapt the DGP proposed in Faruque et al. [2024]. The DGP comprises four endogenous time-series (X_t^1, X_t^2, A_t, Y_t) whose structural equations are

$$
\begin{aligned}
X_t^1 &= 2\left(\cos\frac{t}{10} + \log\big(|X_{t-2}^1 - X_{t-5}^1| + 1\big)\right) + 0.1\,\varepsilon_1 \\
X_t^2 &= 12\,e^{\frac{-(X_{t-1}^1)^2}{2}} - 4\,e^{\frac{-(X_t^1)^2}{2}} + \varepsilon_2 \\
A_t &= -10.5\,e^{-\frac{(X_{t-1}^1)^2}{2}} + 5.5\,e^{-\frac{(X_t^1)^2}{2}} - 3.5\,e^{-\frac{(X_t^2)^2}{2}} + \varepsilon_3 \\
Y_t &= -11.5\,e^{-\frac{(X_{t-1}^1)^2}{2}} + 13.5\,e^{-\frac{(A_{t-1})^2}{2}} + 5.5\,e^{-\frac{(X_t^1)^2}{2}} - 3.5\,e^{-\frac{(X_t^2)^2}{2}} - 5.0\,e^{-\frac{(A_t)^2}{2}} + \varepsilon_4
\end{aligned}
$$

where all noise terms are i.i.d. $\varepsilon_{it} \sim \mathcal{N}(0,1)$. Note that we retain the explicit time trend $\cos(t/10)$ in X_t^1; the resulting mechanism is mildly time-dependent, providing a stress test beyond exact causal stationarity. The window causal graph (Fig. 7b) mirrors that of the first non-linear dataset but with fewer parent-child links per node, making the true dependency structure more sparse. We use this dataset for comparative evaluation on the history-adjusted query.

Tumor Growth Dataset To evaluate the model's performance in a realistic setting, we employ a tumor growth simulation based on the Pharmacokinetic-Pharmacodynamic (PK-PD) framework proposed by Geng et al. [2017]. This model characterizes the daily evolution of tumor volume, denoted by V_t. The dynamics of V_t are governed by the following formula:

$$
V_t = \left(1 + \underbrace{\rho \log\left(\frac{K}{V_{t-1}}\right)}_{\text{Tumor Growth}} - \underbrace{\beta_c C_t}_{\text{Chemotherapy}} - \underbrace{\left(\alpha d_t + \beta d_t^2\right)}_{\text{Radiation}} + \underbrace{\epsilon_t}_{\text{Noise}}\right) V_{t-1} \tag{26}
$$

where the patient-specific parameters $\{\rho, K, \beta_c, \alpha, \beta\}$ are sampled from prior distributions as specified in Geng et al. [2017]. Here, d_t represents the radiation dose applied at time t, and C_t denotes the chemotherapy drug concentration, which is assumed to follow an exponential decay with a half-life of one day. The term ϵ_t accounts for environmental noise in the growth process. Following the approach of Lim et al. [2018], we introduce time-dependent confounding by making the treatment assignment dependent on the past tumor diameter. The probabilities p_t^c and p_t^d, representing the likelihood of assigning chemotherapy and radiotherapy respectively, are determined by the average tumor diameter $\bar{D}_t$ over the last 15 days:

$$
p_t^c = \sigma\left(\frac{\gamma_c}{D_{\max}}\left(\bar{D}_t - \theta_c\right)\right), p_t^d = \sigma\left(\frac{\gamma_d}{D_{\max}}\left(\bar{D}_t - \theta_d\right)\right) \tag{27}
$$

where $\sigma(\cdot)$ is the sigmoid function. We set the maximum tumor diameter $D_{\max} = 13cm$ and fix parameters $\theta_c = \theta_d = D_{\max}/2$ to ensure a 0.5 probability of treatment when the tumor reaches half its maximum size. The parameters γ_c and γ_d control the degree of time-dependent confounding. A higher value of γ assigns greater weight to the history of tumor diameter, thereby increasing the confounding bias in the treatment assignment. The window causal graphs are shown in Fig. 8, illustrating how the dependency structure varies with different values of γ.

D.3 BASELINES

We compare TSCNF with representative causal time-series methods that target marginal or history-adjusted interventional estimands under hypothetical treatment sequences in longitudinal observational settings. Beyond empirical performance, Table 5 summarizes the methodological scope of each baseline in terms of supported treatment types, multiple treatments and outcomes, density versus point estimation, flexibility in assigning treatment and target variables, query flexibility, graph input, and the handling of contemporaneous edges. The details of each baseline are provided below.

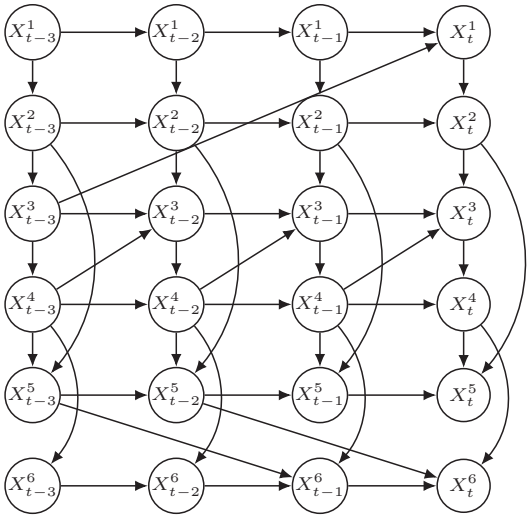


Figure 6: Window causal graph structuring the fully synthetic DGP (τ=3) used in ablation study experiments.

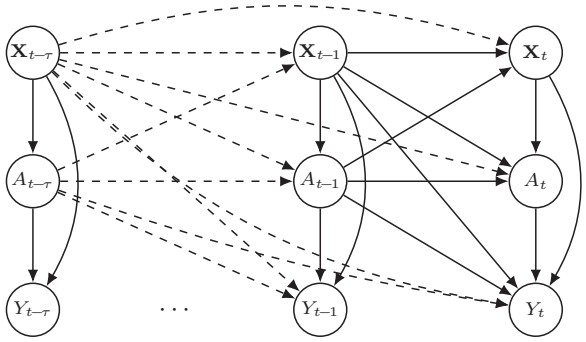
Model	Cont./Disc. Treat.	Multi. Treat.	Multi. Out.	Density/Point Est.	Treat./Target Var.	Mar./HA. Estimand	Graph Input	Contemp. Edge
RMSN	Both	✓	✓	Point	Fixed	HA.	Assumed	N/A
Causal CPC	Disc.	✓	×	Point	Fixed	HA.	Assumed	N/A
G-Net	Both	✓	×	Both	Fixed	HA.	Assumed	N/A
MSCVAE	Both	×	✓	Both	Fixed	Marginal	Assumed	N/A
DoFlow	Cont.	✓	✓	Both	Flexible	HA.	Provided	×
TSCNF	Both	✓	✓	Both	Flexible	Both	Provided	✓

Table 5: Comparison of causal time-series methods. Cont./Disc. Treat.: supported treatment types; Multi. Treat./Out.: support for multiple simultaneous treatments/outcomes; Density/Point Est.: type of estimation output; Treat./Target Var.: whether treatment and target variables are flexibly assignable; Mar./HA. Estimand: whether the method targets marginal or history-adjusted interventional estimands; Graph Input: whether an arbitrary graph is accepted as input; Contemp. Edge: whether contemporaneous edges are explicitly accounted for when the graph input is arbitrary (N/A when graph input is already assumed). Accepting an arbitrary graph as input means the method can be run with different user-provided graphs; it does not imply graph discovery or posterior graph uncertainty.

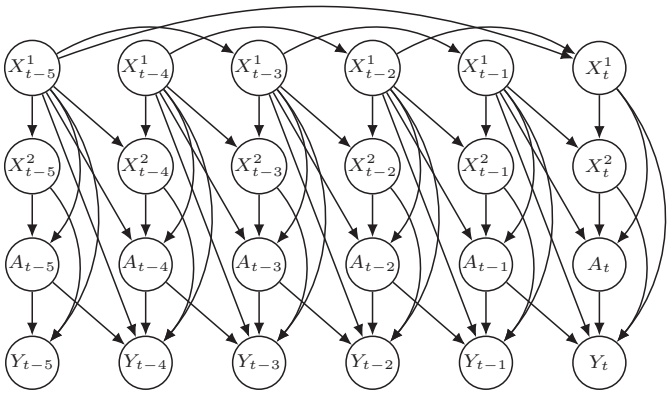
MSCVAE [Wu et al., 2024] MSCVAE is a conditional variational auto-encoder designed to generate samples for the counterfactual distribution under time-varying treatments. MSCVAE implicitly learns the counterfactual distribution by training a conditional generator g_θ. The model operates by approximating the true counterfactual distribution $f_{\bar{a}}(y)$ (i.e., $P(Y \mid do(A = \bar{a}))$ with a proxy conditional distribution $f_\theta(y \mid \bar{a})$, where $\bar{a}$ denotes a given treatment history of length d (e.g., $\bar{a} = (a_{t-d+1}, \ldots, a_t)$) and y is the outcome variable. The generator g_θ functions as the decoder of the variational autoencoder. It takes a random noise vector z and the treatment history $\bar{a}$ as input, defined as $g_\theta(z, \bar{a}) : \mathbb{R}^r \times \mathcal{A}^d \to \mathcal{Y}$, to produce samples of the outcome y. The key component of MSCVAE is the use of Inverse Probability Treatment Weighting (IPTW). Given observed data $(y, \bar{a}, \bar{x})$, IPTW $w_\phi(\bar{a}, \bar{x})$ is computed as the inverse of the product of propensity scores $f_\phi(a_\tau \mid \bar{a}_{\tau-1}, \bar{x}_\tau)$.

Since MSCVAE was originally introduced for discrete (binary) treatment sequences, its propensity score was computed with the joint Bernoulli likelihood of the observed treatment path. However, as our experiments involve continuous treatment, we adapt MSCVAE to use generalized propensity scores. For continuous treatments, we extend the original binary setting by modeling each propensity score as a conditional Gaussian distribution, where the mean and variance are estimated via ordinary least squares (OLS) regression on covariates and treatment history, following the formulation in Robins et al. [2000]. The resulting propensity score is computed as the probability density function value of the Gaussian evaluated at the observed treatment a_τ.

MSCVAE leverages IPTW by incorporating it into its learning objective to correct for treatment assignment bias stemming from observed confounders $\bar{x}$. The core learning objective for MSCVAE is to maximize the expected log-likelihood under



(a) Window causal graph structuring the fully synthetic DGP over a temporal window of size $\tau \in \{1, 3, 5, 30\}$.



(b) Window causal graph structuring sparse confounding with size $\tau = 5$.

Figure 7: Window causal graph used in comparison experiments with baselines. (a) Including time-varying covariates($\mathbf{X}_t$), treatments(A_t) and outcomes(Y_t), treatment assignments depend on all past covariates and treatments, reflecting time-varying confounding. (b) This graph includes same contemporaneous dependencies as in (a), but only a few recent covariates influence treatment.

the true counterfactual distribution, $E_A[E_{Y\sim f_a} \log f_\theta(y|a)]$. This can be approximated by an IPTW-weighted sum over the observed data:

$$\hat{\theta} = \arg\max_\theta \frac{1}{N} \sum_{(y,\bar{a},\bar{x})\in D} w_\phi(\bar{a}, \bar{x}) \log f_\theta(y|\bar{a}). \tag{28}$$

For its implementation as a conditional variational autoencoder (CVAE), $\log f_\theta(y|\bar{a})$ is approximated by its Evidence Lower Bound (ELBO):

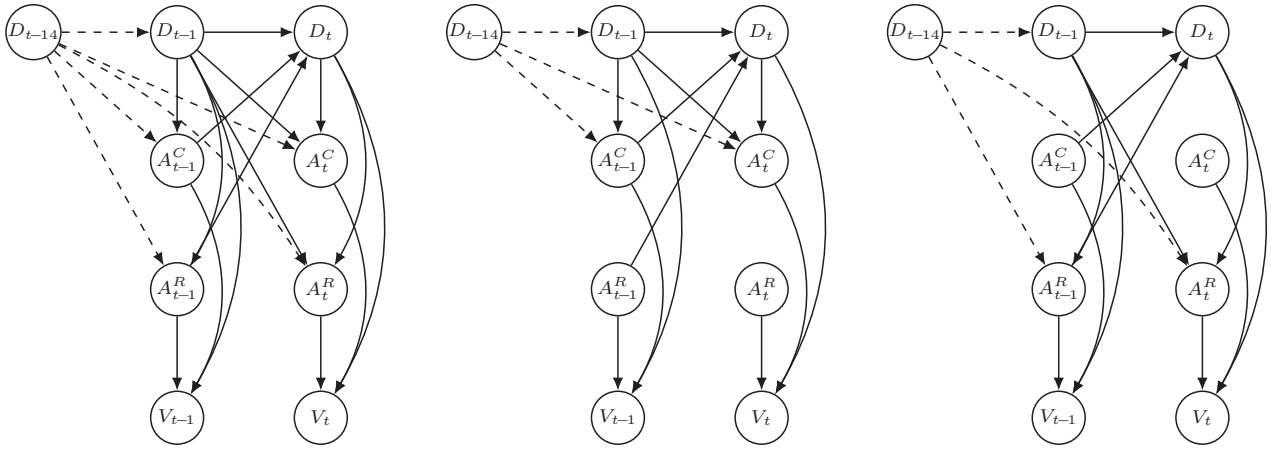
$$\log f_\theta(y|\bar{a}) \geq -D_{KL}(q(z|y,\bar{a}) \,||\, p_\theta(z|\bar{a})) + E_{q_\theta(z|y,\bar{a})}[\log p_\theta(y|z,\bar{a})], \tag{29}$$

where $q_\theta(z|y,\bar{a})$ is the encoder and $p_\theta(y|z,\bar{a})$ is the decoder (generator g_θ). By maximizing the IPTW-weighted sum of ELBOs, MSCVAE is trained to ensure that its generator g_θ is capable of producing samples that conform to the true counterfactual distribution $f_{\bar{a}}(y)$.

G-Net [Li et al., 2021] G-Net is a deep sequential learning framework for counterfactual prediction under dynamic treatment strategies, operating via the principles of g-computation. G-computation allows for the estimation of counterfactual distributions under a specified dynamic treatment strategy g. The desired distribution is defined by the g-formula [Robins, 1986] as:

$$\begin{aligned} p(Y_t(\bar{A}_{m-1}, \underline{g}_m) = y|H_m) = \int_{l_{m+1:t}} & p(Y_t = y|H_m, l_{m+1:t}, A_{m:t} = g(H_{m:t})) \\ & \times \prod_{j=m+1}^{t} p(L_j = l_j|H_m, l_{m+1:j-1}, A_{m,j-1} = g(H_m, l_{m+1:j-1})) \end{aligned} \tag{30}$$

Since the integral in this expression is generally intractable, it is common to approximate it through Monte Carlo simulation. Based on this formulation, G-Net framework operates in two main stages: **(i) Model training**: Using observational data, G-Net learns the conditional expectation of the next–step covariates L_{t+1} (including outcome Y_t) given the history H_t and treatment A_t. It employs recurrent neural networks (such as RNN/LSTM) to encode all prior steps into a latent state (history representation) R_t, enabling the network to capture complex temporal and non-linear dependencies. **(ii) Monte-Carlo simulation on Counterfactual Outcomes**: After training, counterfactual trajectories are generated recursively for each Monte Carlo sample. Starting from an observed history H_t, the simulation proceeds step-by-step into the future: 1. apply the dynamic treatment policy g_m to the current simulated history to obtain treatment A_t (e.g., $A_k = g(H_k)$); 2. sample L_{k+1} from the learned model based on H_k, A_k (e.g., $P_\theta(L_{k+1} \mid H_k, A_k)$). To reflect intrinsic stochasticity, noise sampled from the empirical residual distribution obtained on a hold-out set is added to the predicted L_{k+1}. Simulated L_{k+1} and



(a) $\gamma_c = 5, \gamma_d = 5$ (b) $\gamma_c = 5, \gamma_d = 0$ (c) $\gamma_c = 0, \gamma_d = 5$

Figure 8: Window-based causal graphs induced by the PK–PD tumor growth model under different (γ_c, γ_d) settings. Here, D_t denotes the tumor diameter deterministically computed from V_{t-1}, acting as a historical confounder.

determined A_k are added to history H_{k+1}, so that this history can be used as input for next step prediction. Repeating this recursive simulation process M times (e.g., $M = 100$) produces a collection of diverse counterfactual trajectories. This comprehensive set of simulated paths enables us to estimate the distribution of counterfactual outcomes under time-varying treatments.

RMSN [Lim et al., 2018] Recurrent Marginal Structural Networks (RMSN) is a deep learning framework designed to forecast a patient's expected response to a series of planned treatments while adjusting for time-dependent confounding. Drawing inspiration from marginal structural models (MSMs) in epidemiology, RMSN utilizes inverse probability of treatment weighting (IPTW) to correct for bias in observational data where treatment actions are contingent on time-varying biomarkers. The model architecture is subdivided into two primary components: **(i) Propensity Networks**: These networks consist of a set of Long-Short Term Memory (LSTM) units that compute the treatment and censoring probabilities required for stabilized IPTWs. By using LSTMs instead of the pooled logistic regression common in standard MSMs, RMSN can capture complex temporal dependencies and better specify the conditional probability of treatment assignments. The framework's flexibility allows it to handle various treatment forms, including binary interventions and continuous dosages. **(ii) Prediction Network**: The prediction network employs a sequence-to-sequence architecture to forecast multi-step treatment responses for a given set of planned interventions. It comprises an encoder that learns representations for the patient's current clinical state through one-step-ahead predictions, and a decoder that propagates these representations forward in time based solely on proposed treatments. To bridge the two, a memory adapter layer initializes the decoder's internal state using the encoder's final memory. This design enables RMSN to evaluate complex treatment scenarios over flexible horizons without the need to forecast intermediate input covariates.

Causal CPC [El Bouchattaoui et al., 2024] Causal Contrastive Predictive Coding (Causal CPC) is an approach for counterfactual regression over time that leverages the computational efficiency of RNNs combined with information-theoretic regularizations to capture complex temporal dependencies. Causal CPC employs a Gated Recurrent Unit (GRU) encoder to learn a context representation $\mathbf{C}_t$ from the history $\mathbf{H}_t$. To capture long-term dependencies, the encoder is pretrained by minimizing $\mathcal{L}^{CPC}$, which maximizes the mutual information (MI) between $\mathbf{C}_t$ and future local features $\mathbf{U}_{t+j}$ over multiple forecasting horizons $j = 1, \ldots, \tau$:

$$I\left(\mathbf{U}_{t+j}, \mathbf{C}_t\right) \geq \log(|\mathcal{B}|) - \mathcal{L}_j^{(InfoNCE)} \tag{31}$$

$$\mathcal{L}^{CPC}(\theta_1, \theta_2, \{\Gamma_j\}_{j=1}^{\tau}) := \frac{1}{\tau} \sum_{j=1}^{\tau} \mathcal{L}_j^{(InfoNCE)}(\theta_1, \theta_2, \Gamma_j). \tag{32}$$

To align with causal identification assumptions, the model incorporates the InfoMax principle to ensure the representation is reconstructable from the history, effectively retaining all necessary confounding information. This representation is further refined through adversarial learning designed to mitigate selection bias by minimizing the Contrastive Log-ratio Upper

Bound (CLUB) of the MI between the representation $\mathbf{\Phi}_t = \mathrm{SELU}\left(\mathrm{Linear}\left(\mathbf{C}_t\right)\right) = \Phi_{\theta_R}\left(\mathbf{H}_t\right)$ and the treatment assignment W_{t+1}, thereby achieving a balanced representation that is non-predictive of the treatment.

$$\begin{aligned} I_{\mathrm{CLUB}}\left(\Phi\left(\mathbf{H}_t\right), W_{t+1}; q_{\theta_W}\right) :=& \mathbb{E}_{\mathbb{P}_{\left(\Phi(\mathbf{H}_t), W_{t+1}\right)}}\left[\log q_{\theta_W}\left(W_{t+1} \mid \Phi\left(\mathbf{H}_{t+1}\right)\right)\right] \\ & - \mathbb{E}_{\mathbb{P}_{\Phi(\mathbf{H}_t)}} \mathbb{E}_{\mathbb{P}_{W_{t+1}}}\left[\log q_{\theta_W}\left(W_{t+1} \mid \Phi\left(\mathbf{H}_{t+1}\right)\right)\right]. \end{aligned} \tag{33}$$

The final counterfactual regression is performed by passing the balanced and invertible representation to a GRU-based decoder, which autoregressively predicts potential outcomes $Y_{t+1:t+\tau}$ under a given hypothetical treatment sequence $\omega_{t+1:t+\tau}$. By combining invertible history encoding with information-theoretic balancing, Causal CPC estimates individual treatment effects over long horizons while remaining computationally efficient.

DoFlow [Wu et al., 2026] DoFlow is a flow-based generative framework for multivariate time-series forecasting that explicitly embeds a causal DAG to support both interventional and counterfactual queries. The model captures structural dependencies by parameterizing, for each variable $x_{i,t}$, a continuous-time normalizing flow [Chen et al., 2018] whose dynamics are defined by a neural velocity field v_i. This design allows generation to follow the topological order of the causal graph during inference. To account for temporal dynamics, DoFlow employs an RNN-based encoder to summarize the past histories of each node into a recurrent state $h_{i,t} = \mathrm{RNN}\left(x_{i,t}, h_{i,t-1}\right)$. At time t, the conditioning context is constructed as:

$$H_{i,t-1} := \mathrm{concat}\left(h_{i,t-1}, h_{pa(i),t-1}\right), \tag{34}$$

which aggregates historical information from node i and its parents up to time $t-1$. This design ensures that generation is conditioned on both temporal and causal histories, while implicitly assuming that causal dependencies arise solely from past states without explicitly accounting for contemporaneous interactions. DoFlow is trained using conditional flow matching (CFM) [Lipman et al., 2023], which learns the velocity field v_i by minimizing the discrepancy between the modeled vector field and a target conditional flow. During interventional forecasting, intervened variables are fixed to given values and the remaining variables are generated via the reverse CNF dynamics, following the topological order of the causal DAG. For counterfactual inference, DoFlow performs the abduction-action-prediction procedure by inferring latent variables from factual observations and decoding the trajectory under an intervention.

D.4 OVERALL EXPERIMENTAL SETUP DETAILS

For all synthetic-data experiments, we train all models with 9,000 time-series training samples and validate on 1,000 validation samples. In both marginal and history-adjusted interventional sampling step, we draw 1,000 Monte Carlo samples to approximate the desired distribution. The same Monte Carlo sample count is used for TSCNF and the distributional baselines in the reported experiments. For marginal queries, we average results over 30 independent random seeds. For history-adjusted queries, we average results over 100 randomly selected observed histories. All experiments in this paper were executed on GPU-equipped environment. Detailed specifications are as follows:

- CPU: 2 × AMD EPYC 9124 (16 cores, 3.0 GHz)
- GPU: 4 × NVIDIA RTX A5000, 24 GB RAM memory per GPU
- Storage: 1.8 TB SSD + 11 TB HDD.

D.5 ABLATION EXPERIMENT DETAILS

To evaluate the contribution of each core component in our TSCNF, we perform a series of one-factor-at-a-time ablation experiments. We conduct each experiment under various types of causal queries, leveraging TSCNF's flexibility in handling a wide range of interventional queries.

Variants of TSCNF We compare three CNF-based variants to isolate the effects of feature-wise causal masking and feature-wise conditioning:

- **Standard conditional normalizing flow (without both)**: The variant equivalent to a standard conditional normalizing flow conditioning on the full history window of all variables. No feature-wise causal masking is applied (i.e. every variable sees the same full-window context), and no feature-wise conditioning is applied.

- **Feature-wise causal masked CNF (with feature-wise causal masking, without feature-wise conditioning)**: Feature-wise causal masking allows CNF to embed only causally relevant past values within the fixed window. Without feature-wise conditioning, a single context vector (which remains global) is shared across features.
- **TSCNF (both feature-wise causal masking and feature-wise conditioning)**: Our proposed model, which combines feature-wise causal masking and feature-wise conditioning. Each feature's own history embedding (masked to its causally relevant variables) is provided separately to the conditional flow.

Dataset We use a synthetic dataset (linear and non-linear) generated from the causal DAG with $\tau = 3$, as illustrated in Fig. 6. The dataset consists of $T = 100$ time steps, each containing six variables, among which two are designated as treatment variables.

Estimands We conduct multiple treatment and multiple outcome estimation. We intervene on two treatment variables X_t^2 and X_t^4, setting them at the median (50th percentile) of their empirical distributions, and estimate the distribution of all remaining variables. We evaluate three intervention settings: single intervention $P(\mathbf{X}_t \setminus \{X_t^2\} \mid do(X_t^2))$, $P(\mathbf{X}_t \setminus \{X_t^4\} \mid do(X_t^4))$ and multiple intervention $P(\mathbf{X}_t \setminus \{X_t^2, X_t^4\} \mid do(X_t^2, X_t^4))$. In this setting, we consider two types of interventional queries: marginal and history-adjusted.

Evaluation metrics For each variant and estimand, we report mean distance, maximum mean discrepancy (MMD), 1-dimensional Wasserstein distance. For each setting, we report the metric as the average computed over all non-intervened variables. In this setting, we evaluate how effectively each model component contributes to capturing inter-variable dependencies.

Implementation details In all ablation experiments, we use the following TSCNF configuration. For the causal history encoder, we use a masked 1D convolution network with kernel size = 3, hidden channels (H) = 32 and output dimension (o) = 5, and two encoder layers where the two encoder layers correspond to parallel masked convolutional blocks whose outputs are averaged before the final projection. For the causal conditional normalizing flow, we use 5 conditional masked autoregressive flow (MAF) layers, each with hidden units [32,32,32].

D.6 COMPARISON EXPERIMENTS DETAILS

We compare TSCNF against baselines that are tailored to either marginal or history-adjusted queries: **MSCVAE** and **G-Net**. We evaluate five static treatment strategies with quantile levels $q \in \{0.25, 0.37, 0.50, 0.63, 0.75\}$. For each strategy q, every treatment at each time step is set to the q-th empirical percentile of its training data distribution. This design probes performance across a broad range of hypothetical interventions.

Dataset We conduct both marginal and history-adjusted interventional experiments on fully synthetic datasets from linear and non-linear DGPs as described in Appendix D.2.

Marginal interventional query In marginal interventional experiments, we compare TSCNF to two baselines: MSCVAE and G-Net. MSCVAE is specifically designed to answer marginal queries, while G-Net is originally tailored for history-adjusted settings and thus requires conditioning on past covariate-treatment histories. To ensure a fair comparison, we provide historical input windows of size τ to both TSCNF and G-Net during inference, whereas MSCVAE—being incompatible with history conditioning—is evaluated in its default setup using only the treatment sequence $\mathbf{a}_{m+1:m+\tau}$ without access to $\bar{\mathbf{x}}_m$. We use training data with sequence length $T = 100$ and test using prior steps $m = \tau$, prediction horizons up to $k = \tau$, varying the causal window size as $\tau \in \{1, 3, 5\}$.

History-adjusted interventional query In history-adjusted interventional experiments, we compare TSCNF to G-Net. Note that the estimand in this experiment coincides with the main paper's target estimand, which also matches the setup used by G-Net, enabling a fair comparison (cf. the footnote in Sec. 4.3). G-Net is based on a fixed graphical model with a specific estimand, so we adapt our more flexible framework accordingly. In particular, G-Net targets the history-adjusted interventional distribution where the outcome is measured at the *final step of the intervention sequence*, rather than the step *after* the interventions as in Alg. 1. In our algorithm, this distinction is captured by whether or not the conditional observation sampling step (line 7) is performed. Specifically, while the default rollout of Alg. 1 predicts the outcome at the next step following the interventions—requiring line 7—G-Net's estimand, like our target estimand, corresponds to the outcome at the final intervention step itself. To match this within our framework, we simply skip line 7 in the sampling

procedure, which allows us to recover G-Net's estimand using our model. Below, we consider two evaluation regimes designed to probe different forms of time-varying confounding:

- DGP with $\tau = 30$: Both models are trained on sequences with $T = 30$ and causal window $\tau = 30$. In inference step, we evaluate with prior steps $m = 27$ and prediction horizons up to $k = 3$. Each model conditions on all 27 prior steps, and as each model autoregressively predicts up to $t = m + k$, the history is extended by appending the simulated covariates and applied treatments. Thus, the conditioning context grows autoregressively over the prediction horizon.
- DGP with $\tau = 5$: Both models are trained with $T = 100$ and $\tau = 5$. In inference step, we use prior steps $m = 5$ and prediction horizons up to $k = 5$. For each inference step $t = m + 1, \ldots, m + k$, G-Net and TSCNF condition only on a fixed window of the last $\tau = 5$ lags, ensuring both methods receive identically windowed inputs.

Evaluation metrics We report root mean square error (RMSE), maximum mean discrepancy (MMD), 1-dimensional Wasserstein distance, averaged over 100 individual histories and the five static treatment strategies. We report each result averaged over 30 repetitions.

Implementation details In all experiments, we provide hyper-parameters used for MSCVAE, G-Net and TSCNF in Table 10 and Table 11. The search ranges for all hyper-parameters are provided in Table 12. Parameter counts and inference-time comparisons are summarized in Table 13.

D.7 TUMOR GROWTH DATASET EXPERIMENT DETAILS

We compare TSCNF against baselines that estimate the expected outcomes under a hypothetical treatment sequence given the observed history, $\mathbb{E}\big[Y_t(\underline{a}_{m+1}) \mid \bar{\mathbf{x}}_m\big]$, including **RMSN**, **G-Net**, **Causal CPC** and **DoFlow**. We evaluate multi-step-ahead prediction performance up to horizon $k = 5$ under varying time-dependent confounding levels and static treatment strategies.

Dataset The dataset is simulated from the PK-PD framework proposed by Geng et al. [2017] as described in Appendix D.2. Our tumor growth dataset setting follows the CRN benchmark construction [Bica et al., 2020], enabling direct comparison with DoFlow [Wu et al., 2026], which adopts the same setting. In contrast, CausalCPC [El Bouchattaoui et al., 2024] uses a different dataset construction. Unlike CausalCPC, we vary (γ_c, γ_d) to cover multiple time-dependent confounding regimes. We construct three distinct datasets corresponding to each confounding level setting $(\gamma_c, \gamma_d) \in \{(5, 5), (5, 0), (0, 5)\}$. Each dataset contains 10,000 patient sequences for training. For evaluation, we generate four separate test sets within each setting by simulating 1,000 patient sequences for each static treatment strategie: no treatment, only chemotherapy, only radiotherapy, and both chemotherapy and radiotherapy. Following the baselines previously tested on this dataset, we utilize a mixed dataset of continuous and discrete variables, consisting of continuous tumor volume, binary chemotherapy and radiotherapy assignments and previous tumor volume as a historical confounder. Following Javaloy et al. [2023], we handle the binary treatment variables via dequantization, adding continuous noise to the discrete values so that the flow can model the mixed continuous–discrete data.

Estimand The target estimand is the expected outcome $\mathbb{E}\big[Y_t(\underline{a}_{m+1}) \mid \bar{\mathbf{x}}_m\big]$. For multi-step-ahead prediction, we fix the observed history length $m = 55$ and evaluate at each prediction step $k = 1, \ldots, 5$ (i.e., $t = m + 1, \ldots, m + 5$). Notably, TSCNF and G-Net model the conditional interventional distribution rather than directly estimating the conditional expectation. For these models, we approximate the point estimate by averaging 1,000 Monte Carlo samples per patient.

Evaluation Metrics RMSE is computed across all patient-treatment strategy pairs within each confounding level at each prediction step k. We report Normalized RMSE (nRMSE, %), defined as the RMSE expressed as a percentage of the maximum tumor volume (1,150 cm^3) as in [Lim et al., 2018, Li et al., 2021, El Bouchattaoui et al., 2024, Wu et al., 2026]. Results are averaged over 30 independent runs, and we report mean and standard deviation.

Implementation details For this experiment, we use a TSCNF architecture based on a masked LSTM causal history encoder and a conditional masked autoregressive flow (MAF). For each (γ_c, γ_d) confounding setting, we use the corresponding tuned configuration:

- $(\gamma_c, \gamma_d) = (5, 5)$: 15 MAF layers, output dimension $o = 4$, hidden dimension $H = 128$, and 2 encoder layers.
- $(\gamma_c, \gamma_d) = (5, 0)$: 5 MAF layers, output dimension $o = 16$, hidden dimension $H = 64$, and 1 encoder layer.
- $(\gamma_c, \gamma_d) = (0, 5)$: 1 MAF layer, output dimension $o = 8$, hidden dimension $H = 128$, and 4 encoder layers.

For all configurations, each MAF layer uses hidden units [32,32,32]. Parameter counts and inference-time comparisons are provided in Table 13.

D.8 ADDITIONAL EXPERIMENTAL RESULTS

We provide all additional experimental results for both the model ablation study and the comparative evaluation as follows: ablation results are presented in Table 6, Table 7 and Fig. 9, while comparisons with baselines are shown in Table 8, Table 9, Fig. 10, Fig. 11 and Fig. 12.

D.9 ON SENSITIVITY TO GRAPH MISSPECIFICATION

TSCNF is designed for causal inference and causal effect estimation given an input window causal graph, rather than for causal structure learning. In practice, this graph may be provided by prior domain knowledge or time-series causal discovery pipelines, and may therefore contain errors. To examine how such errors affect downstream causal estimation, we evaluate TSCNF under controlled perturbations of the input graph applied before training.

Experiment setup We assess the sensitivity of TSCNF to graph misspecification on the linear DGPs described in Appendix D.2. We use causal window size $\tau = 3$ and the same tuned architecture as in the main experiment. Before training, we perturb the true window causal graph in three ways: dropping existing lagged edges, dropping existing contemporaneous edges, and adding spurious lagged edges. For each perturbation mode, k edges are sampled without replacement from the corresponding eligible edge pool, with pool sizes 18, 3, and 9, respectively. Each perturbed setting is run with three perturbation seeds. We report held-out log-probability, observational MMD at the final time, and history-adjusted interventional MMD for $P(Y_{m+3}(\underline{a}_{m+1}) \mid \bar{\mathbf{x}}_m)$, averaged over fixed histories and intervention settings.

Results. Table 14 shows that removing true edges generally degrades performance, while adding spurious lagged edges has little effect. The impact is particularly clear for missing lagged edges, where held-out log-probability decreases and interventional MMD increases as the perturbation becomes stronger. In contrast, adding lagged edges keeps the metrics close to the true-graph baseline. These results suggest that, within the graph-given causal inference setting considered here, TSCNF is more robust to redundant lagged edges than to missing true causal dependencies.

D.10 ON CHOICE OF CAUSAL HISTORY ENCODER

The causal history encoder E_ϕ in TSCNF can be instantiated using various architectures, including standard 1D convolution, extended 1D convolution, LSTM, and attention mechanisms. The *extended 1D convolution* refers to a multi-layered, dilated convolution with multi-scale kernels, designed to expand the receptive field beyond that of a standard 1D convolution.

Experiment setup In this experiment, we assess the impact of different encoder choices under various conditions—specifically across both linear and non-linear DGPs, and under different causal window sizes $\tau \in \{1, 3, 5, 30\}$ with a fixed sequence length of $T = 31$. We focus on the history-adjusted interventional density estimation task, evaluating each encoder's ability to estimate $P(Y_{m+3}(\underline{a}_{m+1}) \mid \bar{\mathbf{x}}_m)$. Performance is measured using root mean square error (RMSE), maximum mean discrepancy (MMD), and 1-dimensional Wasserstein distance (Wass.).

Results Table 15 reports the performance of different encoders across varying causal window sizes. We summarize our key observations as follows.

- **Short τ:** When the causal window is short relative to the total sequence length, standard 1D convolution achieves the best average performance. Its localized receptive field and parameter sharing are well-suited for capturing short-range temporal dependencies. While the extended 1D convolution is competitive, its increased capacity is not strictly necessary for short windows.
- **Long τ:** As the causal window grows, attention-based encoders yield stronger performance by flexibly modeling long-range, cross-variable temporal dependencies. LSTM also performs comparably in these settings, benefitting from its recurrent architecture tailored for capturing long-range temporal dependencies.

Taken together, these findings suggest that no single encoder universally dominates across all settings. Instead, each architecture exhibits strengths under different temporal regimes—highlighting TSCNF's architectural flexibility. Depending

on the temporal scale and structural complexity of the underlying environment, the encoder can be chosen to best match the given causal structure.

D.11 EMPIRICAL EVIDENCE ABOUT PROP. 2

Javaloy et al. [2023] show that stacking multiple flow layers in CNFs can induce shortcut dependencies during the learning process, and they use an abductive single-layer construction or a causal-consistency regularizer as practical safeguards. Appendix B.4 gives a complementary estimand-level argument: when the true graph and topological order are known, redundant forward edges do not change the truncated-factorization causal effect computed from the same observational distribution. Here, we examine this empirically by comparing four CNF variants across different numbers of layers ($\{1, 2, 3, 5, 10\}$): (1) **Generative ($Z \to X$) model with regularization**, (2) **Generative ($Z \to X$) model without regularization**, (3) **Abductive ($X \to Z$) model with regularization**, and (4) **Abductive ($X \to Z$) model without regularization**. Our goal is to examine whether the estimation performance of these variants are significantly inferior to that of the baseline—namely, the single-layer abductive model without regularization under $\mathcal{G}$.

Dataset We construct static synthetic datasets based on both linear and non-linear DGPs, with each adhering to the causal structure illustrated in Fig. 13a. The linear DGP is defined as:

$$
\begin{aligned}
X^1 &= \varepsilon_1, \\
X^2 &= 10X^1 - \varepsilon_2 \\
X^3 &= 0.25X^2 + 2\varepsilon_3, \\
X^4 &= X^3 + \varepsilon_4 \\
X^5 &= -X^4 + \varepsilon_5
\end{aligned}
$$

and the non-linear DGP is defined as:

$$
\begin{aligned}
X^1 &= \varepsilon_1, \\
X^2 &= 0.8X^1 + 0.3\varepsilon_2 \\
X^3 &= \begin{cases} X^2 + 1.0 + 0.3\varepsilon_3, & \text{if } X^2 > 0 \\ -X^2 - 1.0 + 0.3\varepsilon_3, & \text{otherwise} \end{cases} \\
X^4 &= 0.6\sin(X^3) + 0.3\varepsilon_4 \\
X^5 &= 0.5X^4 + 0.4\sigma(X^3) + 0.3\varepsilon_5 \quad \text{where } \sigma(x) = \frac{1}{1+e^{-x}}
\end{aligned}
$$

where all noise terms $\varepsilon_i \sim \mathcal{N}(0, 1)$.

Estimand The target estimand is the interventional density of the outcome variable X^4 under the joint intervention $do(X^2, X^3)$. The goal is to examine whether redundant shortcut edges introduced during learning (see Fig. 13b) can distort valid causal inference in interventional settings. For this purpose, we define the estimand on a five-node causal graph, ordered as $X^1 \to X^2 \to X^3 \to X^4 \to X^5$.

Evaluation metrics We evaluate model performance by computing the Maximum Mean Discrepancy (MMD) and the 1-dimensional Wasserstein distance between the ground-truth and estimated samples from $\hat{P}(X^4 \mid do(X^2, X^3))$. Experiments on each configuration are repeated 30 times, and we report the mean and standard deviation of these metrics.

Results Fig. 14 presents the performance of different CNF variants in estimating $P(X^4 \mid do(X^2, X^3))$. Across all layer counts, the abductive model without regularization performs comparably to its regularized counterpart, indicating that the regularizer does not improve performance in this setting. Moreover, multi-layer configurations do not show significant degradation, suggesting that redundant same-order shortcut edges need not harm the truncated-factorization estimand when the graph order and observational distribution are fixed. This empirical result complements Proposition 2 and should be read as a setting-specific observation, while causal-consistency regularization remains a useful safeguard for learned multi-layer flows.

Metric	Method	$do(X_t^2)$	$do(X_t^4)$	$do(X_t^2, X_t^4)$
Mean Dist.	w/o both	0.82 (0.92)	1.00 (2.11)	1.27 (3.62)
	w/o feature-wise cond.	0.86 (0.58)	0.77 (0.40)	0.77 (0.45)
	TSCNF	**0.24 (0.12)**	**0.25 (0.11)**	**0.27 (0.14)**
MMD	w/o both	0.35 (0.16)	0.28 (0.13)	0.33 (0.18)
	w/o feature-wise cond.	0.33 (0.20)	0.30 (0.19)	0.32 (0.21)
	TSCNF	**0.03 (0.05)**	**0.04 (0.03)**	**0.06 (0.06)**
Wass.	w/o both	1.08 (0.90)	1.64 (3.75)	1.71 (4.98)
	w/o feature-wise cond.	1.83 (3.57)	0.95 (0.36)	0.97 (0.42)
	TSCNF	**1.06 (0.53)**	**0.32 (0.11)**	**0.36 (0.14)**

Table 6: Ablation study of TSCNF for history-adjusted interventional distributions under $do(X_t^2)$, $do(X_t^4)$ or $do(X_t^2, X_t^4)$ in the linear dataset ($\tau = 3$).

D.12 ON SENSITIVITY TO NORMALIZING-FLOW ARCHITECTURE AND REGULARIZATION

Following the static-CNF analysis in Appendix D.11, we examine the corresponding design choices in TSCNF under time-series interventional queries. Specifically, we vary the NF family, the number of flow layers, and the use of the causal-consistency regularizer.

Experiment setup. In this experiment, we evaluate TSCNF on the $\tau = 3$ linear and non-linear DGPs and the $\tau = 5$ sparse non-linear DGP described in Appendix D.2. We consider the history-adjusted interventional distribution of outcome variable at the final time point, $P(Y_{m+\tau}(\underline{a}_{m+1}) \mid \bar{\mathbf{x}}_m)$. We vary the normalizing-flow family, using either masked autoregressive flow (MAF) or neural spline flow (NSF, [Durkan et al., 2019]), and the number of stacked flow layers. For each architecture, we additionally compare training with and without the causal-consistency regularizer [Javaloy et al., 2023], which penalizes Jacobian dependencies inconsistent with the contemporaneous graph. All remaining model and optimization hyperparameters are fixed to the dataset-specific tuned values used in the main experiments. We report held-out log-probability, joint observational MMD at the final time point, outcome variable MMD for history-adjusted interventional estimand averaged over fixed histories and intervention settings, and training time. Results are reported as the mean and standard deviation over 10 runs.

Results Table 16 shows no systematic degradation in recovery of the target history-adjusted estimand under multiple flow layers or unregularized training, across the settings considered. In some cases, deeper flows or unregularized settings even achieve lower $\text{MMD}_{\text{HA}.}$. While the causal-consistency regularizer can serve as a practical safeguard against graph-inconsistent dependencies, the regularizer increases training time as discussed in Appendix D.1 of Javaloy et al. [2023]. Although the main experiments use MAF, the comparison with NSF shows that TSCNF is not tied to a single NF family. Overall, these results support the design flexibility of TSCNF: the NF family, flow depth, and causal-consistency regularization can be chosen according to dataset complexity and computational cost, rather than treated as fixed requirements for recovering the target history-adjusted estimand.

	Metric	Method	$do(X_t^2)$	$do(X_t^4)$	$do(X_t^2, X_t^4)$
Linear	Mean Dist.	w/o both	0.14 (0.06)	0.26 (0.13)	0.27 (0.13)
		w/o feature-wise cond.	0.12 (0.02)	0.09 (0.01)	0.13 (0.01)
		TSCNF (full)	**0.07 (0.01)**	**0.04 (0.01)**	**0.07 (0.02)**
	MMD ($\times 10^{-2}$)	w/o both	0.27 (0.02)	0.51 (0.06)	0.38 (0.04)
		w/o feature-wise cond.	1.09 (0.07)	0.83 (0.08)	0.60 (0.07)
		TSCNF (full)	**0.14 (0.04)**	**0.13 (0.04)**	**0.15 (0.07)**
	Wass.	w/o both	0.33 (0.09)	0.71 (0.12)	0.77 (0.10)
		w/o feature-wise cond.	0.34 (0.02)	0.23 (0.01)	0.24 (0.01)
		TSCNF (full)	**0.11 (0.01)**	**0.10 (0.01)**	**0.10 (0.01)**
Non-linear	Mean Dist.	w/o both	0.15 (0.02)	0.15 (0.02)	0.17 (0.03)
		w/o feature-wise cond.	0.13 (0.02)	0.15 (0.03)	0.17 (0.02)
		TSCNF	**0.11 (0.02)**	**0.11 (0.03)**	**0.11 (0.02)**
	MMD ($\times 10^{-2}$)	w/o both	0.11 (0.01)	0.09 (0.02)	0.12 (0.02)
		w/o feature-wise cond.	0.07 (0.01)	0.12 (0.02)	0.11 (0.01)
		TSCNF	**0.05 (0.01)**	**0.05 (0.02)**	**0.08 (0.01)**
	Wass.	w/o both	0.34 (0.01)	0.30 (0.01)	0.31 (0.01)
		w/o feature-wise cond.	0.26 (0.02)	0.34 (0.03)	0.29 (0.01)
		TSCNF	**0.21 (0.01)**	**0.18 (0.02)**	**0.18 (0.01)**

Table 7: Ablation study of TSCNF for marginal interventional distributions under $do(X_t^2)$, $do(X_t^4)$ or $do(X_t^2, X_t^4)$ in the linear (top) and the non-linear (bottom) dataset ($\tau = 3$).

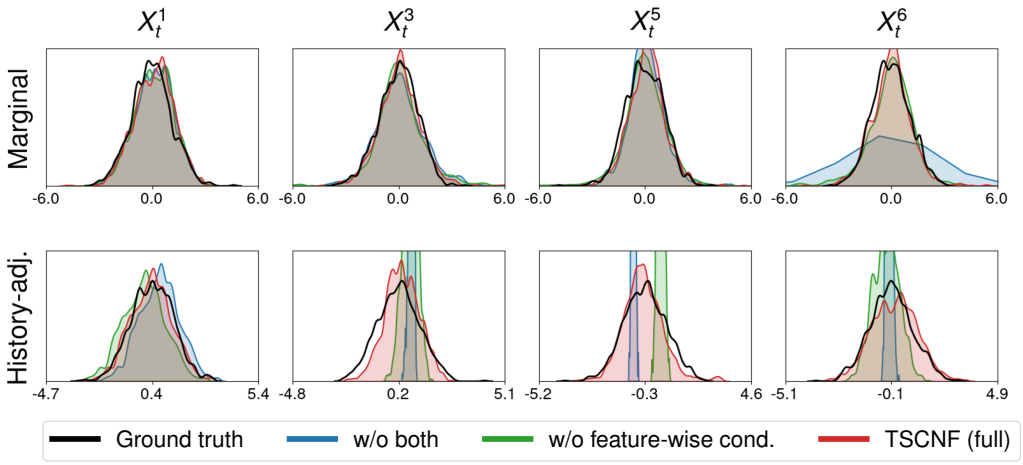


Figure 9: Marginal (top) and history-adjusted (bottom) interventional density plots of each variable under multiple treatment $do(X_t^2, X_t^4)$ for TSCNF and its ablations in the linear synthetic dataset ($\tau = 3$).

	Metric	Model	$\tau = 1$	$\tau = 3$	$\tau = 5$
Marginal	Mean Dist.	MSCVAE	0.08 (0.02)	1.18 (0.03)	0.14 (0.00)
		G-Net	**0.07 (0.04)**	**0.11 (0.06)**	**0.03 (0.02)**
		TSCNF	**0.07 (0.01)**	**0.11 (0.03)**	**0.03 (0.01)**
	MMD ($\times 10^{-2}$)	MSCVAE	0.72 (0.07)	6.98 (0.23)	30.41 (0.24)
		G-Net	**0.07 (0.06)**	0.12 (0.07)	0.09 (0.06)
		TSCNF	0.10 (0.03)	**0.10 (0.03)**	**0.05 (0.02)**
	Wass.	MSCVAE	0.26 (0.01)	1.77 (0.02)	0.65 (0.00)
		G-Net	**0.10 (0.03)**	0.29 (0.06)	0.06 (0.01)
		TSCNF	0.11 (0.01)	**0.27 (0.03)**	**0.05 (0.01)**

	Metric	Model	$k = 1$	$k = 2$	$k = 3$
History-adj.	RMSE	G-Net	0.13 (0.01)	0.12 (0.01)	0.11 (0.01)
		TSCNF	**0.05 (0.00)**	**0.05 (0.00)**	**0.05 (0.00)**
	MMD ($\times 10^{-2}$)	G-Net	0.36 (0.03)	0.29 (0.06)	0.24 (0.04)
		TSCNF	**0.11 (0.01)**	**0.10 (0.01)**	**0.13 (0.01)**
	Wass.	G-Net	0.10 (0.00)	0.10 (0.01)	0.09 (0.01)
		TSCNF	**0.06 (0.00)**	**0.06 (0.00)**	**0.06 (0.00)**

Table 8: (Top) Model performance for marginal interventional density estimation across window sizes $\tau \in \{1, 3, 5\}$ (Bottom) Model performance for history-adjusted interventional density estimation across prediction horizons from $k = 1$ to $k = 3$ in the linear dataset ($\tau = 30$).

Metric	Model	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
RMSE	G-Net	2.51 (0.20)	3.03 (0.31)	3.12 (0.28)	3.34 (0.46)	3.33 (0.44)
	TSCNF	**1.19 (0.19)**	**1.71 (0.03)**	**1.82 (0.02)**	**2.04 (0.05)**	**2.11 (0.05)**
MMD	G-Net	0.64 (0.04)	0.64 (0.07)	0.54 (0.06)	0.58 (0.13)	0.46 (0.05)
	TSCNF	**0.15 (0.03)**	**0.27 (0.02)**	**0.24 (0.02)**	**0.30 (0.03)**	**0.26 (0.01)**
Wass	G-Net	2.02 (0.17)	2.88 (0.32)	2.71 (0.26)	3.34 (0.79)	2.81 (0.29)
	TSCNF	**0.99 (0.11)**	**1.89 (0.05)**	**1.82 (0.05)**	**2.44 (0.08)**	**2.23 (0.07)**

Table 9: Model performances for history-adjusted interventional density estimation across prediction horizons from $k = 1$ to $k = 5$ with sparse temporal dependencies (non-linear setting, $\tau = 5$).

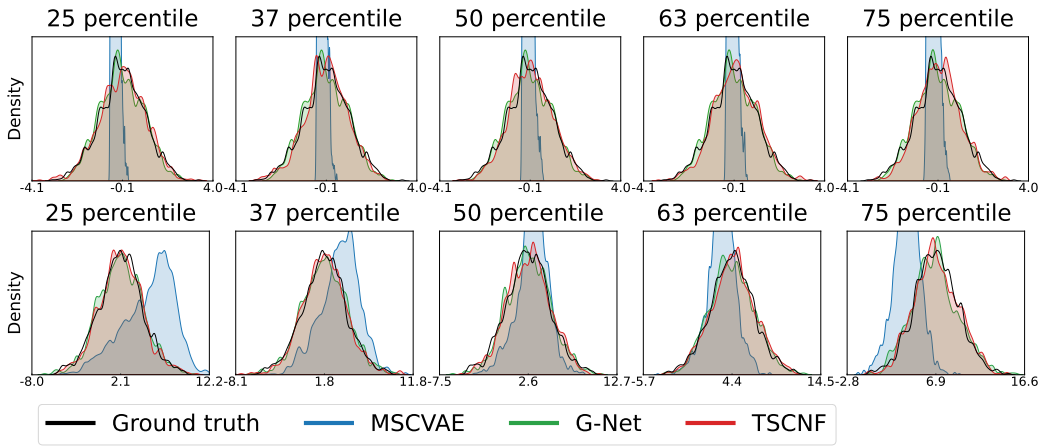


Figure 10: Density plots of marginal interventional distribution $P(Y_{m+5}(\underline{a}_{m+1}))$ under various time-varying treatment strategies in the linear (top) and the non-linear (bottom) dataset ($\tau = 5$).

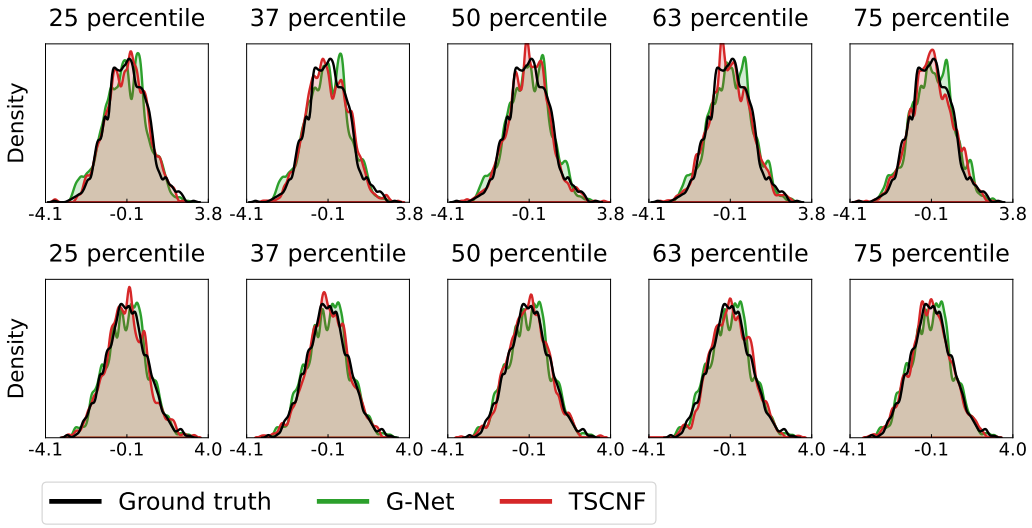


Figure 11: Density plots of the history-adjusted interventional distributions $P(Y_{m+3}(\underline{a}_{m+1}) \mid \bar{\mathbf{x}}_m)$ under various time-varying treatment strategies in the linear (top) and the non-linear (bottom) dataset ($\tau = 30$).

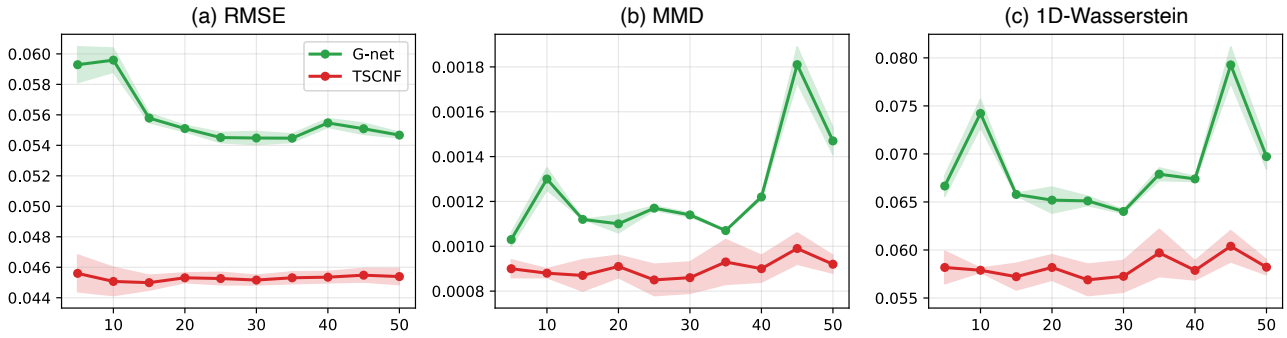


Figure 12: Model performance for history-adjusted interventional density estimation across prediction horizons from $k = 1$ to $k = 50$ in the linear dataset with $\tau = 5$. Results are averaged over 100 randomly sampled observed histories and 5 time-varying treatment strategies. Shaded areas indicate standard deviation.

(a) True causal graphs of the two DGPs

(b) Causal graph with redundant same-order shortcut edges (dashed)

Figure 13: (a) True causal structures of the linear and non-linear DGPs: both share the solid chain, and the non-linear DGP additionally contains $X^3 \to X^5$ (gray). (b) Redundant shortcut edges (dashed) potentially introduced by multi-layered CNF [Javaloy et al., 2023] without causal-consistency regularizer, illustrated relative to the chain graph.

Model	Hyper-parameters	Linear			Non-linear		
		$\tau = 1$	$\tau = 3$	$\tau = 5$	$\tau = 1$	$\tau = 3$	$\tau = 5$
MSCVAE	Latent dimension (r)	5	10	10	5	5	10
	Encoder, Decoder layers	4	4	1	4	2	4
	Layer width	500	2000	2000	1000	2000	2000
	Batch size	256	256	256	256	256	256
	Learning rate	1e-3	1e-3	1e-3	1e-3	1e-3	1e-3
	Epochs	300	300	300	300	300	300
G-Net	LSTM hidden units	8	64	8	16	8	8
	FC hidden units	32	128	32	32	64	32
	LSTM layers	4	1	4	1	1	1
	Batch size	256	256	256	256	256	256
	Learning rate	1e-3	1e-3	1e-3	1e-3	1e-3	1e-3
	Epochs	150	150	150	150	150	150
TSCNF	*Encoder type	1D Conv. net	1D Conv. net	1D Conv. net	1D Conv. net	1D Conv. net	1D Conv. net
	Kernel size	1	3	5	1	3	5
	Hidden channels (H)	32	32	32	16	32	64
	Output dim. (o)	3	3	3	16	3	3
	Encoder layers	8	8	8	2	8	8
	*MAF layers	7	5	10	5	10	5
	*MAF hidden units	[64,64,64]	[32,32,32]	[32,32,32]	[32,32,32]	[32,32,32]	[32,32,32]
	Batch size	100	100	100	100	100	100
	Learning rate	1e-3	1e-3	1e-3	1e-3	1e-3	1e-3
	Epochs	300	300	300	300	300	300

Table 10: Hyper-parameter configurations of MSCVAE, G-Net, and TSCNF across different causal window size $\tau \in \{1, 3, 5\}$ under linear and non-linear settings in *marginal* interventional density (i.e., $P(Y_t(\underline{a}_{m+1}))$) estimation experiments. *Encoder type: *1D Conv. net* denotes a causal history encoder implemented with masked 1D convolutional blocks; when the number of encoder layers is greater than one, the encoder uses multiple parallel blocks with the same kernel size. *MAF layers: the number of conditional masked autoregressive flow (MAF) layers used as the causal conditional normalizing flow. *MAF hidden units: the hidden dimension within a single MAF; for example, [a,b] represents two layers with a and b hidden units.

Model	Hyper-parameters	$\tau = 5$ (Linear)	$\tau = 30$ (Linear)	$\tau = 5$ (Non-linear, Sparse)	$\tau = 30$ (Non-linear)
G-Net	LSTM hidden units	8	32	64	16
	FC hidden units	32	128	64	64
	LSTM layers	4	1	1	4
	Batch size	256	256	256	256
	Learning rate	1e-3	1e-3	1e-3	1e-3
	Epochs	150	150	150	150
TSCNF	*Encoder type	1D Conv. net	1D Conv. extended	1D Conv. net	LSTM
	Kernel size	5	3,5,7,9	5	–
	Hidden channels (H)	32	32	64	64
	Output dim. (o)	3	1	4	16
	Encoder layers	8	4	8	4
	*MAF layers	10	5	7	10
	*MAF hidden units	[32,32,32]	[32,32,32]	[32,32,32]	[32,32,32]
	Batch size	100	100	150	100
	Learning rate	1e-3	1e-3	5e-3	1e-3
	Epochs	300	300	700	300

Table 11: Hyper-parameter settings for G-Net and TSCNF under causal window size ($\tau \in \{5, 30\}$) across linear and non-linear settings in *history-adjusted* interventional density (i.e., $P(Y_t(\underline{a}_{m+1}) \mid \bar{\mathbf{x}}_m)$) estimation experiments. *Encoder type: *1D Conv. net* denotes a causal history encoder implemented with masked 1D convolutional blocks; when the number of encoder layers is greater than one, the encoder uses multiple parallel blocks with the same kernel size. *1D Conv. extended* denotes a causal history encoder implemented with multiple dilated masked 1D convolution layers with multi-scaled kernels. *MAF layers: the number of conditional masked autoregressive flow (MAF) layers used as the causal conditional normalizing flow. *MAF hidden units: the hidden dimension within a single MAF; for example, [a,b] represents two layers with a and b hidden units.

Model	**Hyper-parameters**	$\tau \in \{1, 3, 5\}$	$\tau = 30$
MSCVAE	Latent dimension (r)	[2, 5, 10]	NA
	Encoder, Decoder layers	[1, 2, 4]	NA
	Layer width	[500, 1000, 2000]	NA
G-Net	LSTM hidden units	[8, 16, 32, 64]	[8, 16, 32, 64]
	FC hidden units	[32, 64, 128]	[32, 64, 128]
	LSTM layers	[1, 2, 4]	[1, 2, 4]
TSCNF	Encoder type	1D Conv. net	1D Conv. extended, attention, LSTM
	Kernel size	same as τ	NA
	Hidden channels (H)	[16, 32, 64, 128]	[16, 32, 64]
	Output dimension (o)	[1, 3, 4, 8, 12, 16]	[1, 3, 5, 16]
	Encoder layers	[1, 2, 8]	[2, 4, 8]
	*MAF layers	[1, 3, 5, 7, 10]	[5, 10]
	*MAF hidden units	([32,32,32], [64,64], [64,64,64], [128,128])	[32, 32, 32]

Table 12: Hyper-parameter search ranges for each model across different values of τ.

Table 2 exp.			Table 3 exp.		
Method	#Params	Inference (ms/batch)	Method	#Params	Inference (ms/batch)
MSCVAE	16088021	2.003 ± 0.984	RMSN	4966	7.055 ± 1.372
G-Net	38466	3.285 ± 1.782	G-Net	3050	3.466 ± 1.713
TSCNF	22465	5.580 ± 0.547	Causal CPC	2051	1.759 ± 0.936
			DoFlow	20717	11.787 ± 0.037
			TSCNF	1867788	7.355 ± 1.415

Table 13: Parameter counts and inference time of each model.

Mode	Pool	k	Fraction	log-prob.↑	$\text{MMD}_{\text{obs.}}$ ($\times 10^{-3}$)↓	$\text{MMD}_{\text{HA.}}$ ($\times 10^{-3}$)↓
baseline (true G)	—	0	0%	-4.26	1.48	1.04
drop (lagged)	18	1	5.6%	-4.49 (0.25)	2.47 (1.85)	29.3 (36.2)
		2	11%	-4.49 (0.25)	2.32 (1.77)	28.8 (35.8)
		3	17%	-4.33 (0.13)	1.29 (0.37)	6.08 (7.85)
		6	33%	-4.67 (0.24)	3.37 (1.54)	39.1 (26.2)
		9	50%	-4.62 (0.10)	1.65 (0.20)	32.2 (13.0)
		12	67%	-5.16 (0.35)	6.66 (2.58)	76.2 (35.9)
		15	83%	-5.48 (0.09)	7.25 (0.84)	33.5 (4.4)
drop (contemp.)	3	1	33%	-4.27 (0.01)	1.11 (0.03)	1.64 (0.39)
		2	67%	-5.07 (0.00)	1.43 (0.00)	1.27 (0.01)
add (lagged)	9	1	11%	-4.26 (0.00)	1.04 (0.00)	1.37 (0.02)
		2	22%	-4.26 (0.00)	1.04 (0.01)	1.41 (0.14)
		3	33%	-4.26 (0.00)	1.05 (0.00)	1.43 (0.07)

Table 14: Sensitivity analysis of TSCNF to graph misspecification on the fully synthetic τ=3 setup. We consider three perturbation modes applied to the input causal graph: *drop (lagged)* removes existing time-lagged edges, *drop (contemp.)* removes contemporaneous edges, and *add (lagged)* inserts spurious time-lagged edges. For each mode, k denotes the number of edges perturbed and fraction denotes the proportion of the eligible edge pool affected; the pool is the set of candidate edges eligible for perturbation under each mode. We report held-out log-probability together with joint MMD over all variables for observational inference ($\text{MMD}_{\text{obs.}}$) and MMD over outcome variable for history-adjusted interventional inference ($\text{MMD}_{\text{HA.}}$); the latter is averaged over fixed histories and intervention settings. All metrics are measured on the test set (over 3 perturbation seeds); the baseline (k=0) uses the unperturbed graph (single run).

Metric	Encoder E_ϕ	Linear				Non-linear			
		$\tau = 1$	$\tau = 3$	$\tau = 5$	$\tau = 30$	$\tau = 1$	$\tau = 3$	$\tau = 5$	$\tau = 30$
RMSE	1D Conv.	**0.09**	**0.18**	**0.05**	0.06	**0.08**	0.36	0.75	0.05
	1D Conv. extended	0.16	0.82	0.09	0.06	0.14	1.10	1.47	0.05
	LSTM	0.13	0.29	0.11	**0.05**	0.74	**0.24**	**0.54**	0.05
	Attention	1.14	3.00	0.10	0.05	0.67	5.62	7.03	**0.05**
MMD ($\times 10^{-2}$)	1D Conv.	**0.10**	**0.11**	**0.10**	0.16	**0.14**	0.99	5.33	0.16
	1D Conv. extended	0.28	1.13	0.19	0.14	0.28	8.55	17.31	0.11
	LSTM	0.22	0.18	0.29	**0.10**	3.96	**0.63**	**2.68**	0.12
	Attention	5.91	6.23	0.25	0.10	2.86	30.73	48.52	**0.10**
Wass.	1D Conv.	**0.12**	**0.25**	**0.06**	0.07	**0.09**	0.27	0.52	0.07
	1D Conv. extended	0.19	0.78	0.08	0.07	0.13	0.89	1.24	0.06
	LSTM	0.16	0.32	0.10	**0.06**	0.59	**0.21**	**0.35**	0.07
	Attention	0.87	2.27	0.09	0.06	0.49	4.61	6.14	**0.06**

Table 15: Performance comparison for history-adjusted interventional density estimation $P(Y_{m+3}(\underline{a}_{m+1}) \mid \bar{\mathbf{x}}_m)$ across different causal history encoder E_ϕ under varying $\tau \in \{1, 3, 5, 30\}$ with fixed sequence length $T = 31$. History encoders are evaluated under linear and non-linear DGP following Fig. 7a.

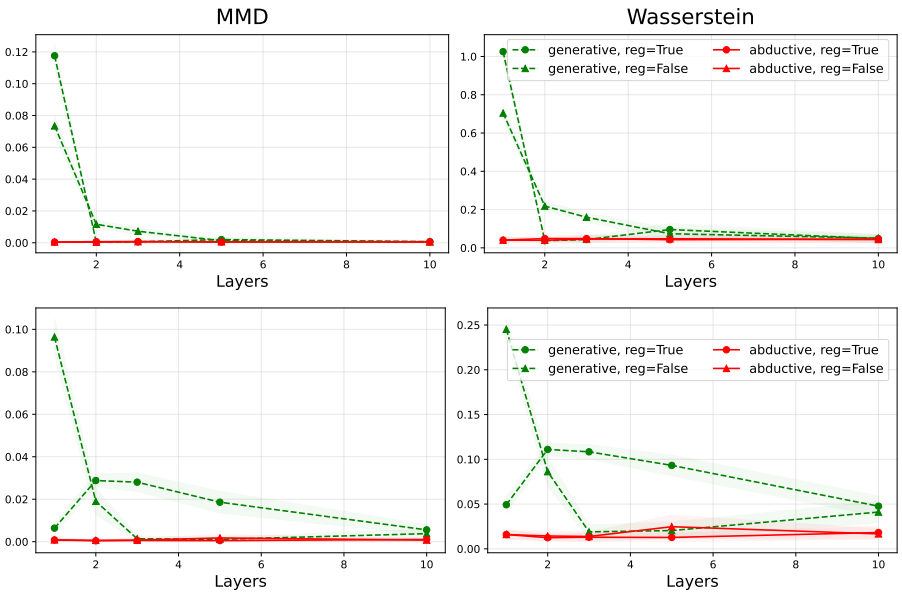


Figure 14: Performance for interventional density estimation $P(X^4 \mid do(X^2, X^3))$ across CNF [Javaloy et al., 2023] variants with and without causal-consistency regularization. Left and right columns show MMD and 1-d Wasserstein Distance, respectively. Green curves represent generative models ($Z \to X$) and red curves denote abductive models ($X \to Z$). Circles are with the causal-consistency regularizer, while triangles indicate models without it. Experiments are repeated 30 times with shaded areas indicating standard deviation, on both the static linear (top) and the static non-linear (bottom) dataset following the corresponding graph shown in Fig. 13a.

Flow	Layers	With regularizer				Without regularizer			
		log-prob. ↑	$\text{MMD}_{\text{obs.}}$ ↓ ($\times 10^{-3}$)	$\text{MMD}_{\text{HA.}}$ ↓ ($\times 10^{-3}$)	Train (min.)	log-prob. ↑	$\text{MMD}_{\text{obs.}}$ ↓ ($\times 10^{-3}$)	$\text{MMD}_{\text{HA.}}$ ↓ ($\times 10^{-3}$)	Train (min.)
Linear synthetic ($\tau = 3$)									
MAF	1	-4.26 (0.00)	2.19 (0.54)	1.05 (0.19)	**12.4 (0.52)**	-4.26 (0.00)	**2.17 (0.53)**	1.05 (0.17)	**9.45 (0.47)**
	3	-4.26 (0.00)	2.19 (0.59)	**0.99 (0.11)**	21.1 (1.55)	-4.26 (0.00)	2.19 (0.58)	1.02 (0.15)	11.1 (0.18)
	5	-4.26 (0.00)	**2.17 (0.57)**	1.03 (0.17)	23.4 (1.17)	-4.26 (0.00)	2.21 (0.59)	**0.97 (0.10)**	12.3 (0.45)
	10	-4.26 (0.00)	2.20 (0.57)	1.01 (0.13)	25.9 (0.12)	-4.26 (0.00)	2.19 (0.57)	1.00 (0.14)	15.7 (1.18)
NSF	1	-4.26 (0.00)	2.20 (0.58)	1.11 (0.18)	20.9 (1.45)	-4.26 (0.00)	2.21 (0.58)	1.12 (0.18)	11.0 (0.43)
	3	-4.26 (0.00)	2.18 (0.57)	1.09 (0.18)	23.0 (0.20)	-4.26 (0.00)	2.19 (0.58)	1.09 (0.15)	15.3 (0.62)
	5	-4.26 (0.00)	2.20 (0.58)	1.07 (0.07)	34.4 (0.41)	-4.26 (0.00)	2.19 (0.58)	1.06 (0.06)	17.7 (0.35)
	10	-4.26 (0.00)	2.23 (0.59)	1.07 (0.15)	60.1 (0.64)	-4.26 (0.00)	2.22 (0.57)	1.02 (0.13)	25.8 (0.76)
Non-linear synthetic ($\tau = 3$)									
MAF	1	-4.35 (0.01)	1.79 (0.63)	**10.1 (3.74)**	**12.3 (0.41)**	-4.35 (0.01)	1.81 (0.65)	**12.4 (4.11)**	**9.28 (0.63)**
	3	-4.35 (0.01)	1.77 (0.68)	14.4 (4.34)	20.2 (2.67)	-4.35 (0.01)	1.78 (0.58)	12.5 (3.90)	11.0 (0.33)
	5	-4.35 (0.01)	1.80 (0.63)	12.6 (8.40)	22.4 (1.41)	-4.35 (0.01)	1.78 (0.62)	12.7 (7.82)	12.6 (0.31)
	10	-4.35 (0.01)	1.79 (0.60)	15.9 (8.00)	25.9 (0.20)	-4.35 (0.01)	1.82 (0.67)	14.6 (6.13)	15.8 (0.72)
NSF	1	-4.37 (0.01)	1.79 (0.54)	12.3 (3.20)	19.5 (1.86)	-4.37 (0.01)	1.78 (0.55)	**12.4 (4.23)**	10.8 (0.59)
	3	-4.37 (0.01)	1.85 (0.65)	15.6 (6.74)	23.1 (0.16)	-4.37 (0.01)	1.81 (0.67)	15.9 (5.79)	14.6 (1.11)
	5	-4.37 (0.01)	**1.75 (0.61)**	20.7 (12.3)	34.2 (0.35)	-4.37 (0.01)	1.80 (0.58)	21.1 (10.8)	17.4 (0.38)
	10	-4.38 (0.01)	1.81 (0.57)	15.3 (4.47)	60.1 (0.65)	-4.37 (0.01)	**1.71 (0.56)**	15.8 (4.81)	26.2 (0.88)
Sparse non-linear synthetic ($\tau = 5$)									
MAF	1	-4.68 (0.07)	3.23 (0.53)	268 (35.8)	**15.7 (0.50)**	-4.67 (0.07)	3.13 (0.59)	268 (29.7)	**12.3 (0.32)**
	3	-4.21 (0.06)	2.88 (0.48)	233 (18.0)	24.7 (1.77)	-4.21 (0.06)	2.95 (0.57)	229 (25.0)	13.6 (0.44)
	5	-3.81 (0.11)	**2.82 (0.50)**	245 (29.9)	26.6 (1.01)	-3.84 (0.11)	**2.83 (0.52)**	245 (38.6)	15.5 (0.35)
	10	-3.31 (0.18)	3.22 (0.58)	252 (35.6)	31.7 (0.32)	-3.33 (0.16)	2.94 (0.55)	253 (35.8)	18.6 (0.44)
NSF	1	-4.40 (0.06)	3.38 (0.62)	253 (39.3)	24.4 (1.23)	-4.39 (0.07)	3.38 (0.75)	245 (46.5)	13.8 (0.57)
	3	-3.62 (0.15)	3.03 (0.38)	215 (36.0)	29.1 (0.17)	-3.63 (0.10)	3.22 (0.64)	214 (38.0)	18.1 (0.75)
	5	-3.24 (0.13)	2.97 (0.53)	209 (26.2)	44.1 (0.87)	-3.25 (0.11)	3.18 (0.63)	**211 (16.1)**	20.2 (0.47)
	10	**-2.61 (0.19)**	3.23 (0.64)	**205 (20.9)**	76.0 (0.64)	**-2.61 (0.23)**	3.04 (0.62)	218 (31.6)	29.0 (0.71)

Table 16: Sensitivity of TSCNF to the choice of NF family, number of NF layers, and causal-consistency regularization, measured by the log-probability on the test set, MMD, and training time. We report joint MMD over all variables for observational inference ($\text{MMD}_{\text{obs.}}$) and MMD over outcome variable for history-adjusted interventional inference ($\text{MMD}_{\text{HA.}}$); the latter is averaged over fixed histories and intervention settings. Bold denotes the best value per column within each dataset; all values are mean (standard deviation) over 10 random seeds.