CAUSALDIFFUSION: CAUSALLY RELATED TIME-SERIES GENERATION THROUGH DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

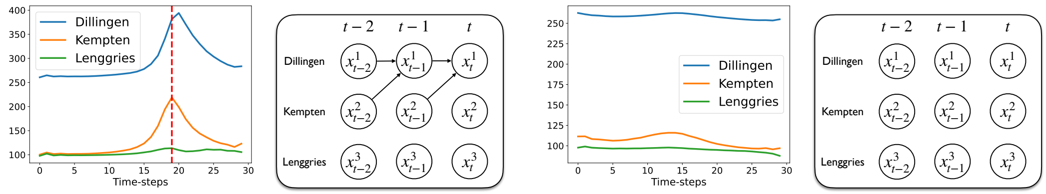
ABSTRACT

Understanding the intrinsic causal structure of time-series data is crucial for effective real-world interventions and decision-making. While several studies address the Time-Series Causal Discovery (TSCD) problem, the lack of high-quality datasets may limit the progress and evaluation of new methodologies. Many available datasets are derived from simplistic simulations, while real-world datasets are often limited in quantity, variety, and lack of ground-truth knowledge describing temporal causal relations. In this paper, we propose `CausalDiffusion`, the first diffusion model capable of generating multiple causally related time-series alongside a ground-truth causal graph, which abstracts their mutual temporal dependencies. `CausalDiffusion` employs a causal reconstruction of the output time-series, allowing it to be trained exclusively on time-series data. Our experiments demonstrate that `CausalDiffusion` outperforms state-of-the-art methods in generating realistic time-series, with causal graphs that closely resemble those of real-world phenomena. Finally, we provide a benchmark of widely used TSCD algorithms, highlighting the benefits of our synthetic data with respect to existing solutions.

1 INTRODUCTION

Many sequential temporal data (i.e., time-series) stemming from real-world phenomena have an inherent causal structure that describes the temporal and spatial interactions among the multiple system variables (Runge et al. (2023)). Understanding such causal relationships is a well-recognized and important challenge for decision-making and policy formulation, as it facilitates predicting the consequences of interventions on underlying systems and variables (Hasan et al. (2023)).

Over the years, several works have studied these underlying causal structures, starting from Granger causality (Granger (1969)). Unable to capture how time affects causal relationships between interdependent time-series, Granger causality has been complemented by efforts to formalize causal graphs (CG) that incorporate the temporal lag in which causality unfolds, as in the leading work of Pearl (2009). More recent studies have addressed deep learning frameworks for time-series causal discovery (TSCD), as explored by Cheng et al. (2023). Many approaches proposed for the TSCD problem (Hasan et al. (2023)) achieve satisfactory performance using statistical and machine learning techniques (Runge et al. (2019b); Pamfil et al. (2020); Sun et al. (2023)), with discovered causal graphs closely resembling the ground-truth counterparts. However, existing benchmark datasets for studying causal structures and evaluating TSCD algorithms are limited in both quantity and quality (Cheng et al. (2024)). The limited data available may hinder the development of new methodologies and studies, and raise concerns about how existing algorithms would perform in unseen real-world scenarios. Novel methodologies to generate realistic time-series with rigorously defined causal graphs are needed to support research and development of algorithms on time-series causal graphs. This challenge has been recently tackled by the works of Li et al. (2023) and Cheng et al. (2024), which marks an initial step in this direction, proposing two deep learning models to generate synthetic time-series data while extracting the corresponding causal graphs. The first approach focuses on the restricted case of Granger Causality (GC) and proposes a recurrent Variational Autoencoder (CR-VAE) framework that naturally encodes causality into the weight matrix connecting the input and hidden states. The second work introduces a comprehensive framework that supports prior causal graphs to generate realistic time-series data. However, when an input causal graph is not provided, the method extracts a hypothesized causal graph using explainability tools for feature



(a) (b)

Figure 1: An example of generated causal graphs and time-series representing three river discharges: sample (a) shows a graph in which Kempten (x^2) has an effect on Dillingen (x^1) with a lag of 1, as can be clearly observed in the corresponding time-series; sample (b) presents a graph with no edges, indicating the absence of causal relationships among the features — the time-series does not provide enough evidence of any underlying effect.

importance (e.g., DeepSHAP, Lundberg (2017)), which are inherently slow and only provide an imprecise approximation of the ground truth graph.

In this paper, we introduce a novel generative framework called `CausalDiffusion` that combines the advantages of previous approaches by naturally encoding a causal graph, along with the time-series, directly within a diffusion model architecture. Specifically, our model incorporates a τ-lag vector autoregressive (VAR(τ)) structure for multivariate time-series (Hamilton (2020)), enabling us to generate realistic time-series data and extract their corresponding ground-truth causal graphs from the VAR coefficients. `CausalDiffusion` can be trained directly on time-series data without requiring prior causal graphs, also eliminating the need for additional explainability tools.

We evaluated our framework on both real and synthetic datasets, benchmarking it against existing state-of-the-art methods. Our results indicate that our approach achieves superior performance in generating realistic time-series data and accurately recovers ground-truth causal graphs.

We can summarize the main contribution of our work as follows:

- We present `CausalDiffusion`, a novel pipeline that employs a diffusion model to generate realistic time-series along with their ground-truth causal graphs.
- We introduce new metrics to assess the accuracy of the generated causal graphs, providing more precise evaluation tools for this domain.
- With extensive experiments, we demonstrate that our method outperforms existing approaches, in terms of synthetic time-series quality and fidelity of causal graphs to real-world phenomena.
- We finally conduct an evaluation of existing causal discovery algorithms using our synthetically generated datasets, highlighting the practical benefits of our data.

We believe that our work may facilitate the research and development of efficient algorithms for uncovering cause-effect relationships in multivariate time-series across diverse fields. We emphasize that our approach specifically addresses the coherence between the synthetic sample and its corresponding causal graph. Figure 1 illustrates two generated data samples.

2 Related work

Synthetic time-series generation Several works have addressed the generation of synthetic time-series starting from real datasets (Yoon et al. (2019); Jarrett et al. (2021); Rasul et al. (2021)). Some approaches have focused on specific aspects, such as the correlation dynamics among variables (Seyfi et al. (2022); Masi et al. (2023)), user-specified constraints (Coletta et al. (2023)), or interpretable generation methods (Yuan & Qiao (2024); Fons et al. (2024)). However, only a few works delve into the generation of time-series along with their causal structure, (Li et al. (2023); Cheng et al. (2024)).

Li et al. (2023) proposed a VAE-based framework capable of learning Granger causal relationships from real multivariate time-series. This approach derives causal relationships from the weight matrix connecting the input and hidden states, allowing a unique causal graph to be learned from the data. All generated samples adhere to such a causal structure. A recent work of Cheng et al. (2024) proposed a pipeline to generate realistic time-series along with the full-time causal graph. However, their framework does not output an interpretable-by-design time-series but it performs the hypothetical causal graph inference through DeepSHAP (Sundararajan & Najmi (2020)) on the trained generative model, introducing a considerable time overhead.

Our goal is to further explore this area and address gaps in the current literature by extending the aforementioned works. Specifically, we aim to extend Granger causality by incorporating temporal lags, generate a unique causal graph for each synthetic sample to introduce greater variety in the data, and provide a naturally interpretable architecture that generates both the synthetic time-series and the causal graph explaining it.

Benchmarking Causal Discovery Algorithms Recent works have studied and tested causal discovery algorithms in several scenarios and domains. Hasan et al. (2023) provide a benchmark of 5 algorithms on both a synthetic and a real dataset (fMRI), evaluating them using several binary classification metrics. Lawrence et al. (2021) use their framework to generate numerical datasets and evaluate 5 causal discovery algorithms, with an in-depth performance analysis concerning their diverse assumptions and hyper-parameters selection. Finally, Cheng et al. (2024) employs the synthetic version of three real datasets to benchmark 13 representative state-of-the-art causal discovery algorithms. We also make use of our synthetic datasets to evaluate such algorithms in Section 6, while Appendix A.2 summarizes all the datasets commonly employed in both simulated and realistic scenarios.

3 Problem Formulation

3.1 Background Knowledge

Causal Discovery The Causal Discovery task aims to ferret out cause-effect relationships among the variables of a d-variate time-series $\boldsymbol{x} = (\boldsymbol{x}^0, \dots, \boldsymbol{x}^{d-1})$. We say that $\boldsymbol{x}^i$ has an effect on (or causes) $\boldsymbol{x}^j$ if the two variables are reflecting a real phenomenon in which a change of $\boldsymbol{x}^i$'s value affects $\boldsymbol{x}^j$. Trivially, the cause must precede the effect so it is important to consider also the lag τ that elapses between observing the cause event on $\boldsymbol{x}^i$ and the effect event on $\boldsymbol{x}^j$. Causal Discovery algorithms are employed to observe real data and point out the existence of causal relationships according to which $\boldsymbol{x}^i$ causes $\boldsymbol{x}^j$, after τ time-steps, returning $(\boldsymbol{x}^i, \boldsymbol{x}^j, \tau)$. We note that the existence of factors, called *confounders*, that influence both the independent variable (the cause) and the dependent variable (the effect) may lead to spurious associations making it harder to determine the true causal relationship. In the literature, it is common to assume the absence of latent confounders when constructing the working dataset.

Causal Graphs Causal relationships are often represented through the so-called Causal Graphs. Let $\tau_{max} \in \mathbb{N}^+$ be the maximum number of discrete time-steps we are interested in to model the cause-effect phenomena of $\boldsymbol{x}$. We define a Causal Graph $G = (V, E)$ where the vertices V represent the time-series variables for the various time-steps between 0 and τ_{max}, and the edges E represent their causal relationships. In particular, an edge $(x^i_{t_1}, x^j_{t_2}) \in E$ indicates that the variable x^i implies the variable x^j with a lag of $t_2 - t_1$ time-steps (i.e., $x^i_{t_1} \Rightarrow x^j_{t_2}$). Formally,

- $V = \{x^i_{t-l} \,|\, 0 \le i < d,\, 0 \le l \le \tau_{max}\}$
- $E = \{(x^i_{t_1}, x^j_{t_2}) \,|\, x^i \Rightarrow x^j$ with a lag of $t_2 - t_1 > 0\}$

Notice that G is a DAG since we are excluding instantaneous causal relationships. Figure 1 shows causal graphs illustrating the interdependencies of river levels.

Granger Causality If we are not interested in a specific lag τ of the cause-effect relation, we can simply resort to the evaluation of the Granger Causality. We say that $\boldsymbol{x}^i$ *Granger-causes* $\boldsymbol{x}^j$ if the past of values of $\boldsymbol{x}^i$ are useful to predict the present of $\boldsymbol{x}^j$ with statistical significance. This kind of relationship can be easily represented by a $d \times d$ matrix M, where $M[i, j] = 1$ means that $\boldsymbol{x}^i$ Granger-causes $\boldsymbol{x}^j$, $M[i, j] = 0$ otherwise.

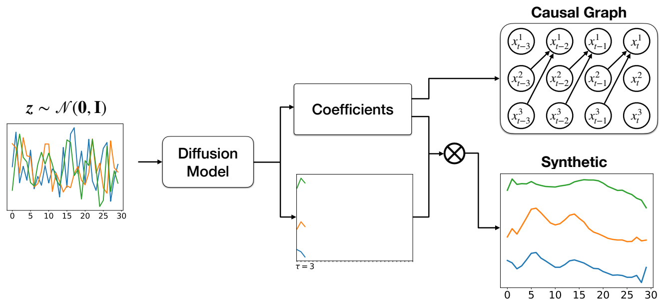


Figure 2: `CausalDiffusion` pipeline.

3.2 Task Definition

Let $\mathcal{D} = \{\boldsymbol{x} \mid \boldsymbol{x} \in \mathbb{R}^{L \times d}\}$ be a set of N d-dimensional input time-series of length L. Our goal is to use the training data $\mathcal{D}$ to learn a generative model that best approximates the distribution of the real time-series, while simultaneously learning the corresponding causal structures. In particular, we aim at generating a couple $\langle \hat{\boldsymbol{x}}, \hat{g} \rangle$ where $\hat{\boldsymbol{x}} \in \mathbb{R}^{L \times d}$ is a synthetic time-series similar to the ones in $\mathcal{D}$ and $\hat{g}$ is the associated causal graph that *explains* $\hat{\boldsymbol{x}}$ in terms of causal relationships.

4 Methodology

The methodology we propose hereby, illustrated in Figure 2, is based on a diffusion model (Ho et al. (2020)) able to map noisy Gaussian vectors $\boldsymbol{z} \in \mathbb{R}^{L \times d}$ to a synthetic sample $\langle \hat{\boldsymbol{x}}, \hat{g} \rangle$. Unless otherwise noted, we adopt common assumptions of the Causal Discovery literature (Cheng et al. (2024); Runge et al. (2019b); Pamfil et al. (2020); Sun et al. (2023): absence of instantaneous effects, Markovian conditions, faithfulness, and sufficiency, as amply discussed in Appendix A.1.

4.1 Diffusion Framework

A diffusion model is a type of latent variable model that operates through two key processes: the forward process and the reverse process. Given a sample $\boldsymbol{x}_0 \in \mathcal{D}$, the *forward process* gradually adds Gaussian noise to obtain a noisy sample $\boldsymbol{x}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Specifically, given the parameters $\beta_t \in (0, 1)$ to schedule the amount of noise added at diffusion step $t \in [1, T]$, the noisy sample is given by

$$\boldsymbol{x}_t = \sqrt{\hat{\alpha}_t} \cdot \boldsymbol{x}_0 + \sqrt{1 - \hat{\alpha}_t} \cdot \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\alpha_t = 1 - \beta_t$, and $\hat{\alpha}_t = \prod_{i=1}^{t} \alpha_i$.

The *reverse process* performs the actual generation of a new sample starting from Gaussian noise. Following the formulation of Yuan & Qiao (2024), we perform the denoising procedure of $\boldsymbol{x}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ as follows:

$$\boldsymbol{x}_{t-1} = \beta_t \cdot \frac{\sqrt{\hat{\alpha}_{t-1}}}{1 - \hat{\alpha}_t} \cdot \hat{\boldsymbol{x}}_0 + \frac{(1 - \hat{\alpha}_{t-1}) \cdot \sqrt{\alpha_t}}{1 - \hat{\alpha}_t} \cdot \boldsymbol{x}_t + \mathbb{1}_{\{t>0\}} \cdot \beta_t \cdot \frac{1 - \hat{\alpha}_{t-1}}{1 - \hat{\alpha}_t} \cdot \boldsymbol{\epsilon} \tag{2}$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\hat{\boldsymbol{x}}_0 = \textsc{Den}_\theta(\boldsymbol{x}_t, t)$ is the output of a neural network parametrized by θ trained with respect the following loss function:

$$\mathcal{L}_{Rec}(\boldsymbol{x}_0, \hat{\boldsymbol{x}}_0; \theta) = \|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_0\|_2^2, \tag{3}$$

where $\|\cdot\|_\rho$ indicates the ℓ_ρ-norm. In practice, $\textsc{Den}_\theta$ reconstructs the original sample taken from the dataset by filtering out the noise added during the forward process.

Other losses can be additionally computed to improve the performance of the reconstruction, as the Fourier-based term employed by Yuan & Qiao (2024):

$$\mathcal{L}_{Fourier}(\boldsymbol{x}_0, \hat{\boldsymbol{x}}_0; \theta) = \|\mathcal{FFT}(\boldsymbol{x}_0) - \mathcal{FFT}(\hat{\boldsymbol{x}}_0)\|_2^2, \tag{4}$$

where $\mathcal{FFT}(\cdot)$ indicates the Fast Fourier Transformation (Elliott & Rao (1982)), or the Dynamic Time Warping-based term $\mathcal{L}_{DTW}(\boldsymbol{x}_0, \hat{\boldsymbol{x}}_0; \theta)$ introduced by Cuturi & Blondel (2017).

Generally, the training objective can be formulated as:

$$\mathcal{L}(\boldsymbol{x}_0, \hat{\boldsymbol{x}}_0; \theta) = \mathop{\mathbb{E}}_{\substack{t\sim\mathcal{U}(1,T) \\ \boldsymbol{x}_0\sim\mathcal{D}}} [\lambda_1 \cdot \mathcal{L}_{Rec}(\boldsymbol{x}_0, \hat{\boldsymbol{x}}_0; \theta) + \lambda_2 \cdot \mathcal{L}_{Fourier}(\boldsymbol{x}_0, \hat{\boldsymbol{x}}_0; \theta) + \lambda_3 \cdot \mathcal{L}_{DTW}(\boldsymbol{x}_0, \hat{\boldsymbol{x}}_0; \theta)] \quad (5)$$

The architecture of DEN_θ consists of an initial convolutional layer followed by a series of RESNET and ATTENTION blocks (see Appendix A.3.3 for more details).

4.2 CAUSAL RECONSTRUCTION OF THE TIME-SERIES

This section details how the output process inherently embeds a causal structure, allowing for the generation of a coherent sample $\langle \hat{\boldsymbol{x}}_0, \hat{g} \rangle$.

Given $\boldsymbol{x}_0 \in \mathcal{D}$, we denote with $x_0^i(l)$ the value of the i-th feature of the time-series at time l, for $i \in [1, d]$ and $l \in [1, L]$[1]. The input and output shapes of DEN_θ must be identical since the network is designed to reconstruct the original sample from a noisy version.

Let $\tau_{max} \in \mathbb{N}^+$ be the maximum lag for which we model the causal relationships in the synthetic time-series[2]. Simultaneously for each feature i, DEN_θ outputs the first τ_{max} steps, i.e. $\hat{x}_0^i(l), \forall\, 1 \leq l \leq \tau_{max}$ and a set of coefficient vectors $\{\boldsymbol{c}^i(l) \mid \tau_{max} < l \leq L\}$, where $\boldsymbol{c}^i(l) = [c_1^1(l), \ldots, c_{\tau_{max}}^1(l), \ldots, c_1^d(l), \ldots, c_{\tau_{max}}^d(l)]$. The reconstruction of the whole time-series in a causal manner follows a Vector Autoregressive (VAR) model (Zivot & Wang (2006)). Proceeding one step at a time for all $\tau_{max} < l \leq L$, the coefficient vector $\boldsymbol{c}^i(l)$ of feature i is multiplied with the previously-defined window: $\hat{x}_0^i(l) = \boldsymbol{c}^i(l) \cdot \hat{\boldsymbol{x}}_0(l - \tau_{max} : l - 1)$.

We underline that, even though the reconstruction can be described by a VAR model, the generation framework is not autoregressive. This is because the model does not consider previously generated outputs as inputs. It instead generates the initial time-steps and the coefficients simultaneously.

Our approach is motivated by an acknowledged technique to identify causal relationships from the estimated VAR coefficients. For instance, the work of Hyvärinen et al. (2010) proves that if the time resolution of the measurements is higher than the time-scale of causal influences, one can estimate a classic autoregressive (AR) model with time-lagged variables and interpret the autoregressive coefficients as causal effects. In particular, they prove that causal effect matrices can be consistently, and computationally efficiently, estimated from the coefficients of the VAR model by means of least-squares methods. Therefore, in agreement with this result, we incorporated a VAR model in the reconstruction to provide guarantees about the identification of causal relationships after appropriate tuning of the sampling period and scaling of the intensity of the observed phenomena.

Finally, to encourage the model to learn sparse causal graphs, i.e. to focus on the most important causal relationships, we add a regularization term for the coefficients. While the ideal choice for such a function would be the ℓ_0-norm, this is difficult to optimize, therefore we consider both the ℓ_1-norm and the ℓ_2-norm, as in Sun et al. (2023); Li et al. (2023). Specifically, the regularization is defined as:

$$\mathcal{L}_{Spars}(\boldsymbol{x}_0; \theta) = \lambda_4 \cdot \|\boldsymbol{c}\|_1 + \lambda_5 \cdot \|\boldsymbol{c}\|_2, \quad (6)$$

where $\boldsymbol{c}$ is the vector of coefficients output by DEN_θ when reconstructing $\hat{\boldsymbol{x}}_0$, and λ_4 and λ_5 are the weights associated to such regularization terms.

4.3 CAUSAL GRAPH EXTRACTION

Given the coefficients vector, we can now extract the causal graph responsible for generating the time-series. The synthetic sample $\hat{\boldsymbol{x}}_0$ is reconstructed through the series of coefficients $\boldsymbol{c}$ of shape $[L - \tau_{max}, d, d \cdot \tau_{max}]$ meaning that for each time-step $\tau_{max} \leq l \leq L$, and for each feature $1 \leq d \leq d$, we have importance weights assigned to the previously generated time-steps, i.e. the window $\hat{\boldsymbol{x}}_0^{(l-\tau_{max}:l-1,:)}$. We also call these coefficients the *explanation* of the synthetic sample. To infer the causal graph $\hat{g}$, we summarize the causal relationships from the VAR coefficients respecting the following formal definition.

[1]To avoid confusion, note that in this section, l refers to the index of the temporal dimension of the time-series, and should not be confused with the diffusion step $t \in [1, T]$ as a subscript.

[2]The maximum lag should be set according to the time-series domain or based on expert domain knowledge.

Definition 4.1. *Let ρ be the percentage of causal relationships we want to keep in the synthetic dataset. For a synthetic sample $\hat{\boldsymbol{x}}$, we say that variable $\hat{\boldsymbol{x}}^i$ $\langle \rho, p \rangle$-causes variable $\hat{\boldsymbol{x}}^j$ with a lag of τ if the p-percentile of the coefficients of the VAR model at lag τ, i.e. giving the effect from $\hat{\boldsymbol{x}}^i(t-\tau)$ to $\hat{\boldsymbol{x}}^j(t)$, is among the ρ% highest values.* Notice that ρ and p refer to the dataset and the single sample, respectively.

This approach has been also employed in Cheng et al. (2024) but there are some other definitions of causality to be extracted from a VAR model, for instance in Hyvärinen et al. (2010). Our definition can be adopted by setting the related parameter values by using domain knowledge or through hyper-parameter tuning techniques. This means that there will be samples with more connections than others, while some may have no causal relationships at all. Notice that, unlike previous work, this allows us to not assume stationarity, as our causal graphs are strictly related to individual samples. This approach enables the generation of diverse samples, each associated with its own distinct causal graph, which may vary across the synthetic samples.

5 Experiments

In the experiments section, we show that the proposed pipeline is able to generate high-quality synthetic samples along with coherent and realistic causal graphs. In this regard, we conducted an experimental campaign involving three different datasets. We compared our models against two state-of-the-art approaches to highlight the advantages of our approach. We evaluate the generated samples both quantitatively and qualitatively, using well-established metrics for synthetic time-series as well as metrics specifically designed to assess the realism of the causal graphs.

5.1 Datasets

To evaluate the models' capability to generate time-series alongside their causal relationships, we utilize two real-world datasets and a synthetic dataset constructed using closed-form equations.

• **Hénon**: introduced by Li et al. (2023), it consists of six coupled Hénon chaotic maps (Kugiumtzis (2013)) described by the following equations:

$$\mathbf{x}^1_{t+1} = 1.4 - (\mathbf{x}^1_t)^2 + 0.3 \cdot \mathbf{x}^1_{t-1}$$

$$\mathbf{x}^p_{t+1} = 1.4 - (e \cdot \mathbf{x}^{p-1}_t + (1-e) \cdot \mathbf{x}^p_t)^2 + 0.3 \cdot \mathbf{x}^p_{t-1}$$

with $p = 2, \ldots, d$, where the number of dimensions $d = 6$ and $e = 0.3$. In this dataset, we have a maximum causal lag equal to 2. There is one positive ($x^p_{t-2} \Rightarrow x^p_t$) and two negative causal relationships ($x^p_{t-1} \Rightarrow x^p_t$ and $x^{p-1}_{t-1} \Rightarrow x^p_t$).

• **Rivers**: introduced by Ahmad et al. (2022), it consists of the average daily discharges of the Iller river at Kempten, the Danube river at Dillingen, and the Isar river at Lenggries between the year 2017 and 2019. The data are provided by the Bavarian Environmental Agency[3]. The Iller is a tributary of the Danube and we expect that an increase in the water level of the former will flow into the latter within a day, i.e., with a lag of 1 time-step. In this case, $d = 3$ and the only causal relationship is $x^{\text{Kempten}}_{t-1} \Rightarrow x^{\text{Dillingen}}_t$. For this dataset, there may be unobserved confounders, such as rainfall, allowing us to test the model's ability to distinguish spurious associations and real causal implications.

• **Air Quality Index** (AQI): introduced by Cheng et al. (2024), it consists of the PM2.5 pollution index monitored hourly over the course of one year by 36 stations spread across Chinese cities[4]. In this case, $d = 36$ and the available causal relationships are modeled through a Granger Causality matrix, which is based on the pairwise distances between sensors (see Appendix A.3.1 for more details).

5.2 Models

Benchmarks. We compare our model against the two most recent state-of-the-art approaches. The first one is CausalTime introduced by Cheng et al. (2024). It is an autoregressive model based on

[3] `https://www.gkd.bayern.de`

[4] `https://www.microsoft.com/en-us/research/project/urban-computing`

normalizing flows, able to observe some time-steps of the time-series and generate the subsequent step. Thanks to this architecture the authors can extract the importance of each feature in the input time-series using an explainability technique,i.e., DeepSHAP, provided by Sundararajan & Najmi (2020), and eventually extract a causal graph. The second one is CR-VAE introduced by Li et al. (2023). It is based on a recurrent VAE made up of a multi-head decoder, in which the p-th head is responsible for generating the p-th feature of the time-series. Encouraged by a sparsity penalty on the weights of the decoder, it learns a sparse causal matrix able to encode causal relationships among the variables. Since the causal matrix is part of the model's parameter, it will be the same for each synthetic sample generated by the model, in contrast to our approaches and CAUSALTIME. Moreover, a notable limitation of CR-VAE is that it is restricted to the notion of Granger Causality, implying that it does not consider the concept of lag in observing the causal relationships.

`CausalDiffusion` We trained our model using various loss functions and properly tuning the λ parameters in Equation (5) and Equation (6). In particular, we evaluate the following variants of our approach: OUR is trained by making use only of the standard reconstruction loss ($\lambda_1 = 10$); OUR W/L2 adds the ℓ_2-norm to sparsify the coefficients ($\lambda_5 = 1$); OUR W/L2 W/DTW considers also the DTW-based loss ($\lambda_3 = 0.01$). Additional loss functions, namely the ℓ_1-norm and the Fourier-based loss, are evaluated in a further ablation study presented in Appendix A.4.2.

All our models are trained with the hyper-parameter τ_{max} fixed to 2 for all three datasets (see Section 4.2). All the other hyper-parameters are shown in the Appendix in Table 5. The sequence length is fixed to 32 for Hénon and Rivers datasets, and to 24 for AQI dataset.

5.3 EVALUATION METRICS

To evaluate the quality of generated time-series and causal graphs, we selected a diverse set of metrics spanning various aspects of the synthetic samples.

Evaluation of time-series. We tested the quality of the synthetic time-series using well-known metrics for fidelity, usefulness, and diversity.

• DISCRIMINATIVE SCORE (Discr.) Yoon et al. (2019) measures the fidelity of synthetic time-series, evaluating to which extent they are indistinguishable from real ones. It consists in training an off-the-shelf 2-layer LSTM to distinguish real samples from synthetic ones. It is formally defined as $|0.5 - \text{AUROC}|$ where AUROC is the area under the ROC (Receiver-Operating Characteristic) curve of the trained discriminator.

• PREDICTIVE SCORE (Pred.) Yoon et al. (2019) measures the usefulness of synthetic time-series for a downstream prediction task. It involves training a post-hoc sequence-prediction model (2-layer LSTM) to predict the subsequent steps of a time-series by optimizing the ℓ_1 reconstruction loss. The predictor is trained on synthetic data and evaluated on real data in terms of the Mean Absolute Error (MAE) of the reconstructions.

• AUTHENTICITY (Auth.) Alaa et al. (2022) measures the portion of synthetic data that is *authentic*, i.e. the models should not simply memorize the training dataset by generating copies of real samples just observed but *invent* new samples.

• MAXIMUM MEAN DISCREPANCY (MMD) Gretton et al. (2006) measures the similarity of synthetic and real time-series distributions. Formally, it is defined as $\text{MMD}^2(P, Q) = \mathbb{E}_P[k(X, X)] - 2 \cdot \mathbb{E}_{P,Q}[k(X, Y)] + \mathbb{E}_Q[k(Y, Y)]$ where $k(\cdot, \cdot)$ is the Radial Basis Function (RBF) kernel.

• CROSS-CORRELATION (xCorr.) measures the extent to which synthetic time-series preserves the cross-correlation of real data. In detail, we evaluate the MAE between the correlation values of the real features and synthetic features.

• DIMENSIONALITY REDUCTION is used to evaluate the diversity of synthetic samples, i.e., they cover the full variability of real samples. We employed t-SNE (Van der Maaten & Hinton (2008)) and PCA (Bryant & Yarnold (1995)) on both real and synthetic data to easily visualize how similar the two distributions are in a 2-dimensional space.

Evaluation of Causal Graphs. To evaluate the corresponding causal graphs, we should first consider the following: despite the existence of a causal phenomenon relating the variables of the datasets, not all the samples extracted from the long time-series may exhibit such a phenomenon. For instance, concerning the Rivers dataset, even if the Iller is a tributary of the Danube, if there is no increase in the water level of the former, the phenomenon of causality cannot be observed. Indeed, the water level of the three rivers simply remains stable over substantial periods of time, as a result, many time-step windows extracted from the dataset will not provide evidence of the causal relationship. In this regard, we employ metrics that do not require that each sample show the causal relationships we expect. On the other hand, there are some causal relationships that we know for sure are not realistic and we focus on this kind of error the models make. The recent work of Ahmad et al. (2022) and Hasan et al. (2023) addressed such a problem by introducing metrics based on the false positive rate of causal relationships.

Accordingly, we introduce the GRANGER CAUSALITY FALSE POSITIVE RATE (GC-FPR) and the GRAPH FALSE POSITIVE RATE (Graph-FPR), which account for the fraction of connections in the graph that we know are incorrect. This evaluation metric is based on the idea that an implication must be considered true also when the hypothesis is not verified, as it does not penalize samples that do not exhibit the causal implication. At the same time, we are counting as errors other causal relationships that are not part of the real-world causal model. Notice that, since CR-VAE does not output a causal graph for each sample we compute the F1-SCORE to evaluate its causal relationships with respect to the ground-truth Granger Causality matrix.

We emphasize that these methods are not designed to be used as causal discovery algorithms, so we do not focus on the exact causal relationships expected to be extrapolated from the datasets. Rather than that, we focus on the realism of the causal graphs and their coherence with the corresponding generated synthetic time-series, ensured by the design of the generative pipeline.

Finally, we also evaluated the inference time (Inf. time) of the models to generate a synthetic sample and the corresponding graph.

5.4 RESULTS

In this section, we discuss the results of our experimental campaign. All the quantitative scores are shown in Table 1. We report the results for the two state-of-the-art approaches (namely, CAUSALTIME and CR-VAE) and three of our models (namely, OUR, OUR W/L2 and OUR W/L2 W/DTW) for comparisons. All the results report the mean and standard deviation across 10 different seeds.

Regarding the fidelity and the quality of the synthetic time-series, OUR W/L2 W/DTW outperforms the other approaches in terms of MMD on all three datasets, maintaining a satisfactory degree of AUTHENTICITY. It is also the best model concerning the DISCRIMINATIVE SCORE on two out of three datasets and in all the other cases it obtains scores very close to the benchmark. This validates our generated samples with respect to their originality, usefulness, and indistinguishability from real data.

Regarding the causal graphs OUR W/L2 W/DTW achieves both the best GC-FPR and the best Graph-FPR scores in all three datasets. This result is of critical importance given that it ensures the reliability of the graphs as a representation of the causal relationships exhibited by the time-series. For the AQI dataset, we report only the GC-FPR metric that evaluates the Granger Causality matrix, as no lag information is provided in the ground-truth causal phenomena. We recall that CR-VAE does not output a causal matrix for each sample, but it is learned and fixed in the trained model. The FPR metric does not fully capture the model's ability in this context. For this reason we reported the F1-score of the learned matrix with respect to the ground-truth GC matrix, highlighting room for improving performance.

Summarizing, we highlight that our model achieves the lowest DISCRIMINATIVE SCORE and PREDICTIVE SCORE along with the best Graph-FPR ensuring that the synthetic samples exhibit a high level of realness and the causal graphs are reliable.

Even though OUR W/L2 W/DTW turned out to be the best one, we also included the other models to point out the additional losses' impact on performance. As the results show, incorporating the ℓ_2-norm of the coefficients into the objective loss as an attempt to sparsify the causal graph reduces the number of wrong connections. Moreover, the DTW-based loss considerably aids in extracting synchronization signals among the temporal sequences, significantly improving overall

Table 1: Results of the models on the three datasets, where ↓ indicates *lower is better* and ↑ indicates *higher is better*. For each metric, the best result is highlighted in bold, and the second-best result is underlined.

Dataset	Metric	Model OUR	OUR w/L2	OUR w/L2 w/DTW	CAUSALTIME	CR-VAE
Hénon	Discr. ↓	0.09 ± 0.02	0.09 ± 0.01	$\mathbf{0.06 \pm 0.02}$	0.31 ± 0.14	0.24 ± 0.11
	Pred. ↓	$\mathbf{0.15 \pm 0.00}$	$\mathbf{0.15 \pm 0.00}$	$\mathbf{0.15 \pm 0.00}$	0.20 ± 0.01	0.24 ± 0.01
	Auth. ↑	0.59 ± 0.01	0.62 ± 0.01	0.65 ± 0.01	$\mathbf{0.72 \pm 0.03}$	0.65 ± 0.11
	MMD ↓	$\mathbf{0.001 \pm 0.000}$	$\mathbf{0.001 \pm 0.000}$	$\mathbf{0.001 \pm 0.000}$	0.002 ± 0.000	0.012 ± 0.009
	xCorr ↓	$\mathbf{0.03 \pm 0.00}$	0.04 ± 0.00	$\mathbf{0.03 \pm 0.00}$	0.06 ± 0.02	0.13 ± 0.03
	GC-FPR ↑	0.39 ± 0.00	$0.32 \pm 000.$	$\mathbf{0.31 \pm 0.00}$	0.48 ± 0.04	$0.52 \pm 0.07^{*}$
	Graph-FPR ↓	0.09 ± 0.00	0.08 ± 0.00	$\mathbf{0.04 \pm 0.00}$	0.23 ± 0.01	—
	Inf. time ↓	1548ms	1548ms	1548ms	8790ms	**194ms**
Rivers	Discr. ↓	0.08 ± 0.01	0.13 ± 0.01	$\mathbf{0.07 \pm 0.01}$	0.09 ± 0.05	0.11 ± 0.09
	Pred. ↓	0.035 ± 0.001	0.037 ± 0.001	0.033 ± 0.001	$\mathbf{0.026 \pm 0.001}$	0.036 ± 0.002
	Auth. ↑	0.58 ± 0.01	0.62 ± 0.01	0.63 ± 0.01	0.56 ± 0.03	$\mathbf{0.72 \pm 0.02}$
	MMD ↓	$\mathbf{0.001 \pm 0.000}$	$\mathbf{0.001 \pm 0.000}$	$\mathbf{0.001 \pm 0.000}$	0.009 ± 0.011	0.059 ± 0.029
	xCorr ↓	0.06 ± 0.00	0.06 ± 0.01	0.02 ± 0.01	$\mathbf{0.01 \pm 0.00}$	0.12 ± 0.02
	GC-FPR ↓	0.23 ± 0.00	$\mathbf{0.22 \pm 0.00}$	$\mathbf{0.22 \pm 0.00}$	0.57 ± 0.01	$0.37 \pm 0.14^{*}$
	Graph-FPR ↓	0.10 ± 0.00	$\mathbf{0.07 \pm 0.00}$	$\mathbf{0.07 \pm 0.00}$	0.22 ± 0.01	—
	Inf. time ↓	1492ms	1492ms	1492ms	4248ms	**148ms**
AQI	Discr. ↓	0.41 ± 0.02	0.43 ± 0.01	0.36 ± 0.05	0.46 ± 0.02	$\mathbf{0.25 \pm 0.04}$
	Pred. ↓	0.048 ± 0.001	0.048 ± 0.001	0.047 ± 0.001	0.054 ± 0.001	$\mathbf{0.043 \pm 0.001}$
	Auth. ↑	0.81 ± 0.02	0.81 ± 0.02	$\mathbf{0.82 \pm 0.01}$	0.77 ± 0.01	0.80 ± 0.10
	MMD ↓	$\mathbf{0.001 \pm 0.000}$	$\mathbf{0.001 \pm 0.000}$	$\mathbf{0.001 \pm 0.000}$	0.008 ± 0.001	0.017 ± 0.001
	xCorr ↓	0.09 ± 0.01	0.11 ± 0.01	0.10 ± 0.01	$\mathbf{0.03 \pm 0.01}$	0.12 ± 0.01
	GC-FPR ↑	0.48 ± 0.00	0.40 ± 0.00	$\mathbf{0.39 \pm 0.00}$	0.49 ± 0.00	$0.27 \pm 0.00^{*}$
	Graph-FPR ↓	—	—	—	—	—
	Inf. time ↓	1395ms	1395ms	1395ms	205s	**442ms**

performance. More ablation studies involving OUR W/L1 W/DTW, OUR W/L2 W/FOURIER can be found in Table 6 of the Appendix.

Finally, we evaluate the inference time of the models to obtain a sample, made up of the synthetic time-series and the corresponding causal graph. CR-VAE turned out to be the fastest, thanks to its VAE-based architecture. However, great time saving occurs because the causal graph is fixed for each sample since it is extracted from the parameters of the model. Our models achieve an inference time significantly lower than CAUSALTIME. Actually, CAUSALTIME is faster than OUR * in generating the synthetic time-series but the post-processing of the feature importance through DeepSHAP is very time-consuming. Instead, in our architecture the causal graph is generated simultaneously with the time-series, motivating a moderate overhead. Moreover, the sampling of diffusion models can be accelerated, using for example implicit diffusion models (DDIM, Song et al. (2021)).

Additional experiments and results can be found in the Appendix, including the evaluation of the time-series through dimensionality reduction techniques, namely t-SNE and PCA [Appendix A.4.3], and the evolution of the evaluation metrics during the training [Appendix A.4.5].

6 BENCHMARK OF CAUSAL DISCOVERY ALGORITHMS

To demonstrate the usefulness of our generative pipeline we employ our synthetic samples to benchmark several causal discovery algorithms. Given a generated couple $\langle \hat{\boldsymbol{x}}, \hat{g} \rangle$, we feed the algorithm with the generated time-series $\hat{\boldsymbol{x}}$ and we compare the predicted causal graph against the generated graph $\hat{g}$. We exclude the instantaneous relationships from the evaluation since our framework does not model them.

In our benchmark, we included:

• **Granger-Causality**-based approaches: Granger Causality (GC, Granger (1969)); Neural Granger Causality (NGC, Tank et al. (2021)); economy-SRU (eSRU, Khanna & Tan (2019)); Temporal Causal Discovery Framework (TCDF, Nauta et al. (2019)); CUTS (Cheng et al. (2022)); CUTS+ (Cheng et al. (2023));

• **Constraint**-based approaches: PCMCI+ (Runge et al. (2020)); NTS-NOTEARS (Sun et al. (2023)); DYNOTEARS (Pamfil et al. (2020)); Rhino (Gong et al. (2023));

• **CCM**-based approaches: Latent Convergent Cross Mapping (LCCM, De Brouwer et al. (2020));

• **Other** approach: Neural Graphical Model (NGM, Bellot et al. (2021)) employing neural ordinary differential equations.

The results of our benchmark are shown in Table 2, evaluated in terms of AUROC and AUPRC (Area Under Precision-Recall Curve). To always have a well-defined ground truth, for the benchmark we selected the strongest 15% causal connections for each sample. As additional experiments, we also executed the benchmark using the top 1% approach described in Section 4.3. These results are reported in Appendix A.6.2.

Table 2: Results of the benchmark of Causal Discovery Algorithms. Bold and underline are used to highlight the best and the second best result, respectively.

Method	**AUROC** Hénon	**AUROC** Rivers	**AUROC** AQI	**AUPRC** Hénon	**AUPRC** Rivers	**AUPRC** AQI
GC	0.52 ± 0.03	0.57 ± 0.07	0.50 ± 0.00	0.47 ± 0.13	0.46 ± 0.10	0.57 ± 0.09
DYNOTEARS	0.60 ± 0.04	0.51 ± 0.03	0.50 ± 0.00	0.52 ± 0.08	0.58 ± 0.03	0.65 ± 0.00
NTS-NOTEARS	0.57 ± 0.04	0.69 ± 0.10	0.50 ± 0.00	0.45 ± 0.07	0.54 ± 0.13	0.42 ± 0.10
PCMCI+	$\mathbf{0.74 \pm 0.02}$	$\mathbf{0.77 \pm 0.06}$	$\mathbf{0.74 \pm 0.00}$	0.68 ± 0.02	0.64 ± 0.05	0.73 ± 0.02
Rhino	0.51 ± 0.01	0.53 ± 0.06	0.50 ± 0.00	0.70 ± 0.07	0.66 ± 0.07	0.69 ± 0.04
CUTS	$\mathbf{0.75 \pm 0.02}$	$\underline{0.76 \pm 0.06}$	$\mathbf{0.74 \pm 0.00}$	0.68 ± 0.02	0.64 ± 0.05	0.73 ± 0.00
CUTS+	$\mathbf{0.75 \pm 0.02}$	$\underline{0.75 \pm 0.08}$	$\mathbf{0.74 \pm 0.00}$	0.68 ± 0.02	0.62 ± 0.08	0.73 ± 0.00
Neural-GC	0.72 ± 0.01	0.52 ± 0.05	0.50 ± 0.01	0.68 ± 0.01	0.55 ± 0.08	0.64 ± 0.05
NGM	0.61 ± 0.06	0.69 ± 0.12	0.50 ± 0.00	$\underline{0.77 \pm 0.05}$	0.73 ± 0.11	$\underline{0.80 \pm 0.05}$
LCCM	0.55 ± 0.00	0.50 ± 0.00	0.52 ± 0.00	0.67 ± 0.00	$\mathbf{0.78 \pm 0.00}$	0.57 ± 0.00
eSRU	0.50 ± 0.00	0.75 ± 0.11	0.50 ± 0.00	$\mathbf{0.78 \pm 0.02}$	$\underline{0.76 \pm 0.10}$	$\mathbf{0.81 \pm 0.00}$
TCDF	0.52 ± 0.03	0.50 ± 0.01	0.50 ± 0.00	0.35 ± 0.14	0.57 ± 0.03	0.64 ± 0.07

It turns out that three algorithms, namely PCMCI+, CUTS, and CUTS+, achieve the best tradeoff between AUROC and AUPRC on all datasets. Also, NGM obtains satisfying results on the Hénon and Rivers datasets, reaching AUPRC values among the highest. Instead, Neutral-GC performed well only on the synthetic dataset of our benchmark and eSRU only on the Rivers dataset. Among the constrained-based approaches only PCMCI+ achieved satisfying performances, while, in general, the Granger-Causality-based approaches proved to be the best ones. None of the methods got an AUROC lower than 0.5 meaning that there were no inverted classifications. The overall performance of tested algorithms is lower than what has been reported on simpler synthetic datasets, such as Lorenz-96 (Cheng et al. (2023); Tank et al. (2021)), where some methods achieved near-perfect scores. This performance gap may suggest that current algorithms are still inexact on some specific samples and datasets, and they could be further improved. In general, more challenging synthetic datasets should be used to rigorously test and potentially improve existing TSCD methods. The performance degradation observed in some algorithms when exposed to new data further underscores this need.

7 CONCLUSIONS & LIMITATIONS

We introduced `CausalDiffusion`, a novel pipeline to generate faithful time-series along with realistic and coherent causal graphs specifically suited for the TSCD task. To the best of our knowledge, this is the first work to incorporate diffusion models for causally related time-series generation dropping the stationarity assumption. We demonstrated that our model can effectively generate synthetic datasets to support the causal discovery community in enhancing their algorithms in various domains, learning directly from real-world observational data.

We acknowledge among the limitations of this work the assumption of causal sufficiency, i.e. no latent confounders in our datasets. Furthermore, only linear causal relationships are present in the synthetic samples. In future works, in addition to improving the approach to handle the above limitations, it can be extended in two directions. The first involves incorporating the modeling of instantaneous causal relationships. The second improvement is to add a loss-based guidance of the coefficients so that the generation can be conditioned on a prior-known causal graph. In fact, a key advantage of our approach is the realism and flexibility that diffusion models provide, which allows the implementation of sophisticated conditioning strategies on trained models.

Reproducibility Statement

To ensure reproducibility, we openly release the source code on GitHub (`https://anonymous.4open.science/r/causal-diffusion-AEB8`). All the datasets employed are easily accessible and described in Section 5.1. All the hyper-parameters are listed in Table 5.

Acknowledgments

References

Wasim Ahmad, Maha Shadaydeh, and Joachim Denzler. Causal discovery using model invariance through knockoff interventions. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022.

Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pp. 290–306. PMLR, 2022.

Alexis Bellot, Kim Branson, and Mihaela van der Schaar. Neural graphical modelling in continuous-time: consistency guarantees and algorithms. In *International Conference on Learning Representations*, 2021.

Fred B Bryant and Paul R Yarnold. Principal-components analysis and exploratory and confirmatory factor analysis. *L. G. Grimm & P. R. Yarnold (Eds.), Reading and understanding multivariate statistics (pp. 99–136). American Psychological Association.*, 1995.

Xuefei Cao, Björn Sandstede, and Xi Luo. A functional data method for causal dynamic network modeling of task-related fmri. *Frontiers in neuroscience*, 13:127, 2019.

Yuxiao Cheng, Runzhao Yang, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. Cuts: Neural causal discovery from irregular time-series data. In *The Eleventh International Conference on Learning Representations*, 2022.

Yuxiao Cheng, Runzhao Yang, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. Cuts: Neural causal discovery from irregular time-series data. In *The Eleventh International Conference on Learning Representations*, 2023.

Yuxiao Cheng, Ziqian Wang, Tingxiong Xiao, Qin Zhong, Jinli Suo, and Kunlun He. Causaltime: Realistically generated time-series for benchmarking of causal discovery. In *The Twelfth International Conference on Learning Representations*, 2024.

Andrea Coletta, Sriram Gopalakrishnan, Daniel Borrajo, and Svitlana Vyetrenko. On the constrained time-series generation problem. *Advances in Neural Information Processing Systems*, 36, 2023.

Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In *International conference on machine learning*, pp. 894–903. PMLR, 2017.

Edward De Brouwer, Adam Arany, Jaak Simm, and Yves Moreau. Latent convergent cross mapping. In *International Conference on Learning Representations*, 2020.

DF Elliott and KR Rao. Fast fourier transform and convolution algorithms, 1982.

Elizabeth Fons, Alejandro Sztrajman, Yousef El-Laham, Andrea Coletta, Alexandros Iosifidis, and Svitlana Vyetrenko. ihypertime: Interpretable time series generation with implicit neural representations. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=GSnGPgeoS5`.

Wenbo Gong, Joel Jennings, Cheng Zhang, and Nick Pawlowski. Rhino: Deep causal temporal relationship learning with history-dependent noise. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006.

James D Hamilton. *Time series analysis*. Princeton university press, 2020.

Uzma Hasan, Emam Hossain, and Md Osman Gani. A survey on causal discovery methods for iid and time series data. *Transactions on Machine Learning Research*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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

Daniel Jarrett, Ioana Bica, and Mihaela van der Schaar. Time-series generation by contrastive imitation. *Advances in Neural Information Processing Systems*, 34:28968–28982, 2021.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Alireza Karimi and Mark R Paul. Extensive chaos in the lorenz-96 model. *Chaos: An interdisciplinary journal of nonlinear science*, 20(4), 2010.

Saurabh Khanna and Vincent YF Tan. Economy statistical recurrent units for inferring nonlinear granger causality. In *International Conference on Learning Representations*, 2019.

Dimitris Kugiumtzis. Direct-coupling information measure from nonuniform embedding. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 87(6):062918, 2013.

Andrew R Lawrence, Marcus Kaiser, Rui Sampaio, and Maksim Sipos. Data generating process to evaluate causal discovery techniques for time series data. *stat*, 1050:16, 2021.

Hongming Li, Shujian Yu, and Jose Principe. Causal recurrent variational autoencoder for medical time series generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 8562–8570, 2023.

Scott Lundberg. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

Giuseppe Masi, Matteo Prata, Michele Conti, Novella Bartolini, and Svitlana Vyetrenko. On correlated stock market time series generation. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 524–532, 2023.

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

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*, pp. 1595–1605. Pmlr, 2020.

J Pearl. *Causality*. Cambridge University Press, 2009.

Robert J Prill, Daniel Marbach, Julio Saez-Rodriguez, Peter K Sorger, Leonidas G Alexopoulos, Xiaowei Xue, Neil D Clarke, Gregoire Altan-Bonnet, and Gustavo Stolovitzky. Towards a rigorous assessment of systems biology models: the dream3 challenges. *PloS one*, 5(2):e9202, 2010.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International Conference on Machine Learning*, pp. 8857–8868. PMLR, 2021.

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

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5 (11):eaau4996, 2019b.

Jakob Runge, Xavier-Andoni Tibau, Matthias Bruhns, Jordi Muñoz-Marí, and Gustau Camps-Valls. The causality for climate competition. In *NeurIPS 2019 Competition and Demonstration Track*, pp. 110–120. Pmlr, 2020.

Jakob Runge, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 4(7):487–505, 2023.

Ali Seyfi, Jean-Francois Rajotte, and Raymond Ng. Generating multivariate time series with common source coordinated gan (cosci-gan). *Advances in Neural Information Processing Systems*, 35:32777–32788, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

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*, pp. 1942–1964. PMLR, 2023.

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *International conference on machine learning*, pp. 9269–9278. PMLR, 2020.

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

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. Time-series generative adversarial networks. *Advances in neural information processing systems*, 32, 2019.

Xinyu Yuan and Yan Qiao. Diffusion-ts: Interpretable diffusion for general time series generation. In *The Twelfth International Conference on Learning Representations*, 2024.

Eric Zivot and Jiahui Wang. Vector autoregressive models for multivariate time series. *Modeling financial time series with S-PLUS®*, pp. 385–429, 2006.

A Appendix

A.1 Theory

Our work makes the following assumptions, aligned with several TSCD algorithms:

- **Markovian Condition:** The joint distribution of the multi-variate time-series can be factorized into $P(\boldsymbol{x}) = \prod_i P(x_i|\mathcal{P}(x_i))$, i.e., every variable is dependent only on its parents.
- **Causal Faithfulness:** It assumes that the relationships between variables in the data faithfully reflect the true causal connections between them.
- **Causal Sufficiency:** Also known as *no latent confounder*, it states that all common causes of all variables are observed. While this assumption may appear quite strong — since it is impossible to observe "all causes in the world" given the potentially infinite number of variables — it is a common and practical simplification in existing literature when constructing causal datasets.
- **No Instantaneous Effect:** It is very intuitive since it states that the cause must occur before its effect in the time-series. To satisfy this assumption it is sufficient to sample the data with a frequency higher than the causal effects.

On the other hand, thanks to the fact that our approach generates the time-series and its strictly associated causal graph we can drop the stationarity assumption, described below.

- **Causal Stationarity:** It states that all the causal links do not change over time.

A.2 Datasets

Most of the available datasets for studying causal structures in time-series can be classified into three categories.

Numerical datasets. This kind of datasets show causal dependencies that are manually designed through closed-form equations. Vector-autoregression (VAR), Lorenz-96 (Karimi & Paul (2010)), and the framework proposed by Lawrence et al. (2021) belongs to this category. In particular, the last one allows researchers to generate diverse data with several degrees of flexibility, but it lacks a connection to the dynamics of a real-world scenario Runge et al. (2020).

Quasi-real datasets. Similar to the previous category, the causal dynamics are manually designed but calibrated with real data. Several medical datasets rely on Functional Magnetic Resonance Imaging (fMRI, Cao et al. (2019)), a technique to investigate dynamic brain networks. For instance, Nauta et al. (2019) exploited the changes in blood flow to obtain an emulated blood oxygen level-dependent fMRI dataset, resembling the neural activity of different brain regions. Prill et al. (2010) introduced DREAM3, a dataset simulating gene expression, while Nauta et al. (2019) also employed a simulated dataset in the financial domain, based on the Fama-French Three-Factor Model.

Real datasets. Finally, we discuss the available real time-series dataset. Ahmad et al. (2022) introduced the rivers dataset made up of three rivers, in which one is a tributary of another, while Cheng et al. (2024) introduced three datasets, namely Air Quality Index, Traffic, and MIMIC-4 (details in Section 5.1). The first dataset consists of the PM2.5 pollution index monitored hourly; the second one collects traffic information from sensors in the San Francisco Bay Area[5]; while the last one consists of critical care data over a large number of patients in intensive care units (Johnson et al. (2023)). Other datasets that may include causal relationships are MoCap Tank et al. (2021), collecting human motion data, and S&P100 stock data Pamfil et al. (2020). However, most real datasets lack a ground truth causal graph, and even when such graphs are available, they are typically limited in both quantity and diversity.

CauseMe[6] is a platform released by Runge et al. (2019a) to collect many datasets mainly regarding climate scenarios, both synthetic and real.

[5] `https://pems.dot.ca.gov/`

[6] `https://causeme.uv.es`

A.3 IMPLEMENTATION DETAILS

A.3.1 DATASETS

Table 3 reports the most important statistics of our datasets. We also include additional details not discussed in Section 5.1.

Table 3: Statistics of our datasets.

Dataset	Number of Training Samples	Sequence Length	Number of Variables	Number of Causal Relations
Hénon	11295	32	6	11
Rivers	9969	32	3	1
AQI	7246	24	36	354

Hénon: The initial values are sampled from a standard Gaussian distribution, and then the time-series are computed according to the equations in Section 5.1. In this dataset, the causal graph consists of one positive relationship with a lag of 2 between a variable and itself ($x^p_{t-2} \Rightarrow x^p_t$) and two negative relationships. The first one is between the variable and itself with a lag of 1 ($x^p_{t-1} \Rightarrow x^p_t$); the second one is between two consecutive variables again with a lag of 1 ($x^{p-1}_{t-1} \Rightarrow x^p_t$).

Air Quality Index: We recall that, as Cheng et al. (2024) state, the causal relations in the AQI dataset are highly dependent on geometry distances. The graph contained in the dataset they released has been extracted considering Gaussian kernel and a threshold with respect to the geographic distances of the sensors. In particular,

$$w_{ij} = \begin{cases} 1, & \texttt{dist}\,(i,j) \leq \sigma \\ 0, & \text{otherwise} \end{cases}$$

where `dist` measures the distance between two sensors and σ is set to ≈ 40 km. See the work of Cheng et al. (2024) for more details.

A.3.2 BENCHMARK

We compare our approach against two state-of-the-art approaches, implemented from the respective repositories:

- CAUSALTIME Cheng et al. (2024): `https://github.com/jarrycyx/UNN`
- CR-VAE Li et al. (2023): `https://github.com/hongmingli1995/CR-VAE`

We tuned the hyper-parameters of both models on all the datasets and they are reported in Table 4.

Table 4: Hyper-parameters of CAUSALTIME and CR-VAE.

Model	Hyper-parameter	Dataset Hénon	Rivers	AQI
CAUSALTIME	Share type	Decoder	Decoder	Decoder
	N. Epochs Train Phase 1	40	20	10
	N. Epochs Train Phase 2	10	10	5
	Learning rate	0.001	0.0001	0.0001
	Batch size	32	32	32
	Hidden_size	128	128	128
	N. Layers	2	2	2
	N. Heads	4	4	4
	Dropout p	0.1	0.1	0.1
	Flow length	4	4	4
CR-VAE	Hidden	64	64	64
	N. Iterations Train Phase 1	1000	1000	1000
	N. Iterations Train Phase 2	90000	9000	90000
	Learning rate	0.05	0.05	0.05
	Batch size	1024	1024	1024

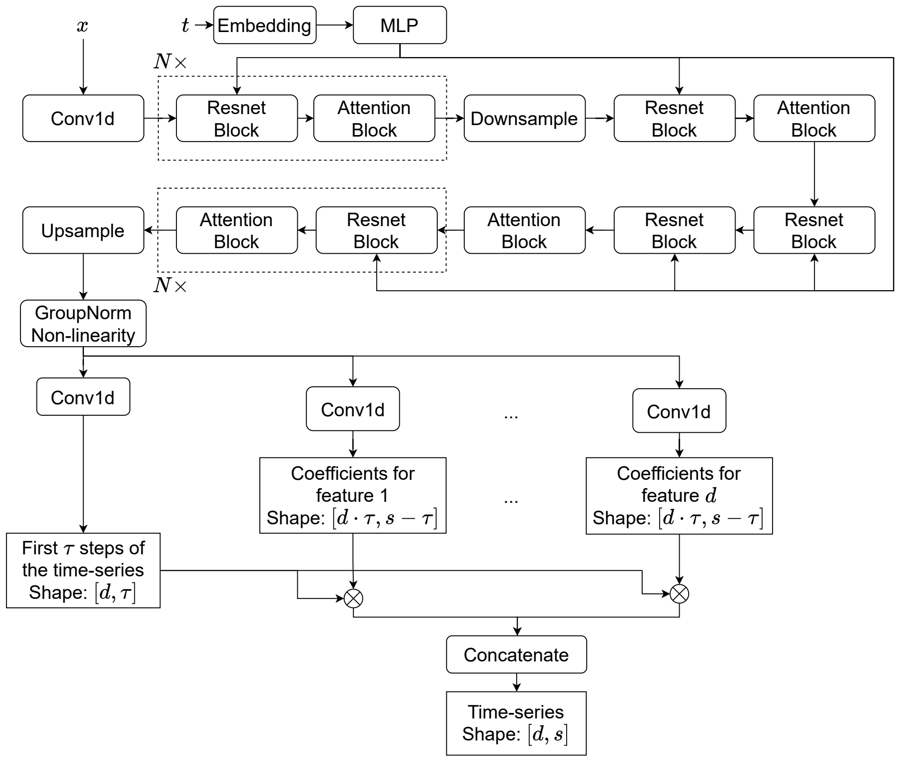


Figure 3: Architecture of DEN_θ.

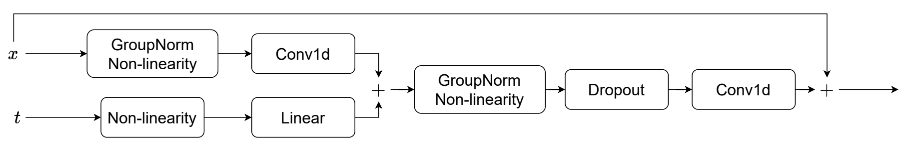


Figure 4: Architecture of the Resnet Block of DEN_θ.

A.3.3 ARCHITECTURE OF DENOISING NETWORK

Here we discuss our diffusion model network. We slightly modified the architecture of Song et al. (2021) released in their repository[7] by adding the convolution layer to output the coefficients. The overall architecture of DEN_θ is depicted in Figure 3. It consists of an initial convolution layer and a series of RESNET and ATTENTION blocks showed in Figure 4 and Figure 5, respectively.

To represent the denoising time-step t the model employs cosine embedding and an MLP block made up of 2 linear layers with the activation function $f(x) = x \cdot \sigma(x)$ in the middle, where σ represents the SIGMOID function $\sigma(x) = \frac{1}{1+e^{-x}}$. The time-step information is injected in all the RESNET blocks.

[7] `https://github.com/mirthAI/Fast-DDPM`

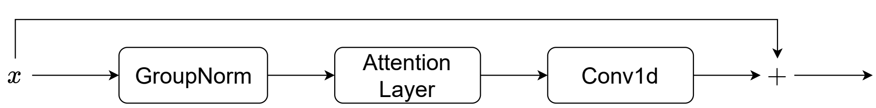


Figure 5: Architecture of the Attention Block of DEN_θ.

In the pictures, NON-LINEARITY and GROUPNORM refer to the function $f(x)$ and Group Normalization, respectively. The DOWNSAMPLE block is just a $1d$-convolution with a stride equal to 2. The UPSAMPLE block is made up of a Nearest Interpolation and a $1d$-convolution.

The end part of the architecture is made up of $d+1$ convolutional layers. The first one is responsible for outputting the first τ_{max} steps of the time-series. Each of the remaining d convolutional layers is responsible for the coefficients of one feature. Then, the coefficients are multiplied with the initial steps of the time-series and the final output is reconstructed following the formalization in Section 4.2.

A.3.4 HYPER-PARAMETERS

The most important hyper-parameters are reported in Table 5.

Table 5: Hyper-parameters of the generative model.

Training epochs	**Batch size**	**Learning rate**	**Diffusion Timesteps (T)**	β **schedule**	**Time-step embedding**
50	32	1e-4	100	Linear `start=0.0001`, `end=0.02`	Cosine `dim=128`

A.3.5 EVALUATION METRICS

• DISCRIMINATIVE SCORE: We trained a 2-layer LSTM for 30 epochs with a learning rate of $1e-4$, hidden size equals 8, and batch size set to 32. The loss function to be optimized is the BINARY CROSS ENTROPY where real samples are labeled as 1 and synthetic samples as 0. The score is formally defined as $|0.5 - \text{AUROC}|$, where AUROC is the area under the ROC curve of the trained discriminator.

• PREDICTIVE SCORE: Following the *train-on-synthetic* and *test-on-real* criterion, we tested the ability of the generated data to inherit the predictive characteristics of the original. We trained a 2-layer LSTM-based predictor to forecast the last $\frac{1}{10} \cdot \texttt{seq\_len}$ time-steps over each synthetic sample for 10 epochs, with a learning rate of $1e-3$, hidden size equals to 32, and batch size set to 32. The loss function to be optimized is the ℓ_1-loss. Then, the predictor is evaluated on real data and quantified through the Mean Absolute Error (MAE). Formally, given a real sequence $\boldsymbol{x}$ of length $\texttt{seq\_len}$ let $\boldsymbol{x}_{first}$ and $\boldsymbol{x}_{last}$ be the first $\frac{9}{10} \cdot \texttt{seq\_len}$ and the last $\frac{1}{10} \cdot \texttt{seq\_len}$ time-steps, respectively. The predictor observe $\boldsymbol{x}_{first}$ and predicts the subsequent $\frac{1}{10} \cdot \texttt{seq\_len}$ time-steps, denoted as $\tilde{\boldsymbol{x}}_{pred}$. The MAE-based performance consists of $\frac{1}{\frac{1}{10} \cdot \texttt{seq\_len}} \sum_{t=1}^{\frac{1}{10} \cdot \texttt{seq\_len}} |\boldsymbol{x}_{last}(t) - \tilde{\boldsymbol{x}}_{pred}(t)|$.

• AUTHENTICITY: We considered the original implementation provided by the work of Alaa et al. (2022). In detail, the authenticity $A \in [0,1]$ measures the portion of synthetic samples that are truly generated by the model, rather than just copied from the training data. The metric is evaluated through a hypothesis test for data copying, which employs a nearest-neighbor classifier. A synthetic sample is considered unauthentic if it is closest to a real training sample. A score close to 1 indicates that the model is generating novel, unseen data.

• MAXIMUM MEAN DISCREPANCY: We used the scikit-learn[8] implementation of the RBF kernel.

• CROSS-CORRELATION: We computed the Cross-Correlation distance for each lag up to 4. Formally, let $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ be a real and a synthetic sample respectively. Moreover, let $\boldsymbol{x}_i$ and $\hat{\boldsymbol{x}}_i$ be the i-th feature of the real and the synthetic sample ($\forall 1 \leq i \leq d$), respectively. The score is formally defined as $\sum_{\tau=0}^{4} \frac{1}{\binom{d}{2}} \cdot \sum_{\{i,j\} \in \binom{\{1,\ldots,d\}}{2}} |(\boldsymbol{x}_i \star \boldsymbol{x}_j)(\tau) - (\hat{\boldsymbol{x}}_i \star \hat{\boldsymbol{x}}_j)(\tau)|$, where $(\boldsymbol{x}_i \star \boldsymbol{x}_j)(\tau)$ denotes the cross-correlation between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ with respect to lag τ.

• DIMENSIONALITY REDUCTION: We used the scikit-learn[8] implementation for both PCA and t-SNE. For each sample, we flattened the dimension of the features by computing the mean.

[8] https://scikit-learn.org/

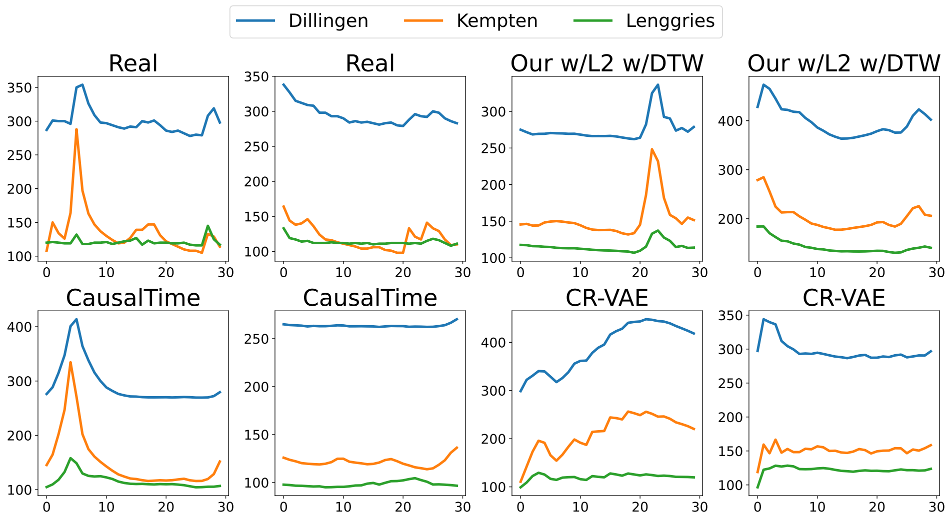


Figure 6: Examples from the Rivers dataset. We recall that the time-series sequence of CausalTime is obtained by feeding the model with a real sequence (seed) and it outputs the subsequent step since it is an autoregressive model. Instead, our method can truly generate new samples from random noise.

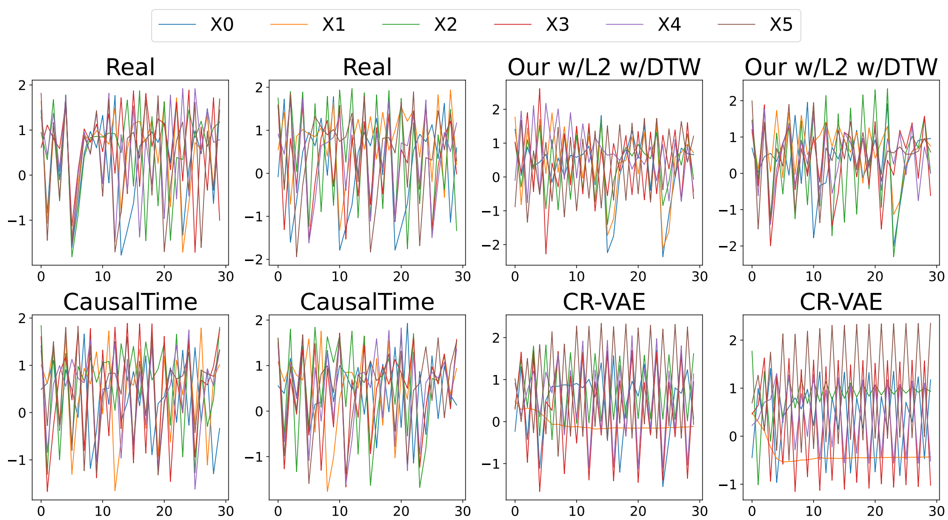


Figure 7: Examples from the Hénon dataset.

A.4 Additional Results

A.4.1 Samples

Figure 6 and Figure 7 show examples of real and generated samples for the Rivers and Hénon datasets, respectively.

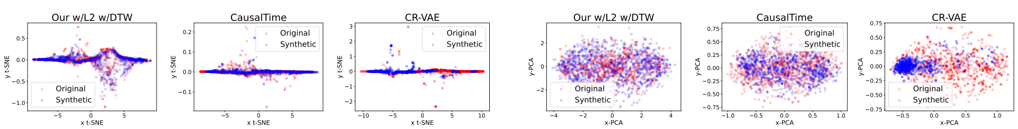
A.4.2 Additional ablation studies

Table 6 show the quantitative results for Our w/L1 w/DTW and Our w/L2 w/Fourier. In particular, the first model considers a DTW-based loss and a ℓ_1-norm where $\lambda_3 = 0.01$ and $\lambda_4 = 1$; while the latter considers a Fourier-based loss with $\lambda_2 = 100$, and ℓ_2-norm to sparsify the coefficients ($\lambda_5 = 1$).

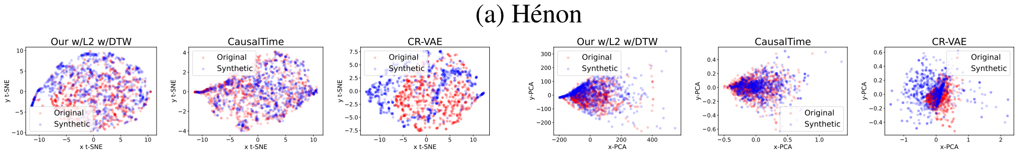
A.4.3 Dimensionality Reduction Plots

Table 6: Results of other models on the three datasets. ↓ indicates *lower is better* and ↑ indicates *higher is better*.

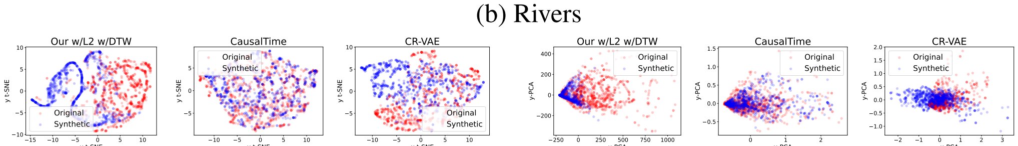
Dataset	**Metric**	**Model** OUR W/L1 W/DTW	OUR W/L2 W/FOURIER
Hénon	Discr. ↓	0.07 ± 0.02	0.09 ± 0.02
	Pred. ↓	0.16 ± 0.00	0.16 ± 0.00
	Auth. ↑	0.61 ± 0.01	0.63 ± 0.01
	MMD ↓	0.001 ± 0.000	0.001 ± 0.000
	xCorr. ↓	0.04 ± 0.00	0.04 ± 0.01
	GC-FPR. ↓	0.34 ± 0.00	0.33 ± 0.00
	Graph-FPR. ↓	0.04 ± 0.00	0.09 ± 0.00
Rivers	Discr. ↓	0.48 ± 0.01	0.10 ± 0.01
	Pred. ↓	0.043 ± 0.001	0.036 ± 0.001
	Auth. ↑	0.87 ± 0.02	0.61 ± 0.01
	MMD ↓	0.054 ± 0.003	0.001 ± 0.00
	xCorr. ↓	0.08 ± 0.01	0.07 ± 0.01
	GC-FPR. ↓	0.27 ± 0.00	0.23 ± 0.00
	Graph-FPR. ↓	0.16 ± 0.00	0.07 ± 0.00
AQI	Discr. ↓	0.44 ± 0.01	0.38 ± 0.03
	Pred. ↓	0.049 ± 0.001	0.050 ± 0.001
	Auth. ↑	0.82 ± 0.01	0.81 ± 0.01
	MMD ↓	$0.001 \pm 0.$	$0.001 \pm 0.$
	xCorr. ↓	0.13 ± 0.01	0.10 ± 0.01
	GC-FPR. ↓	$0.39 \pm 0.$	$0.40 \pm 0.$
	Graph-FPR. ↓	—	—



(a) Hénon



(b) Rivers



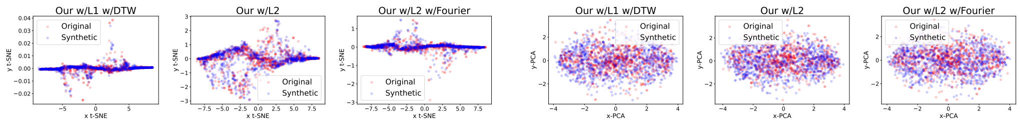
(c) AQI

Figure 8: Dimensionality reduction: t-SNE (left) and PCA (right).

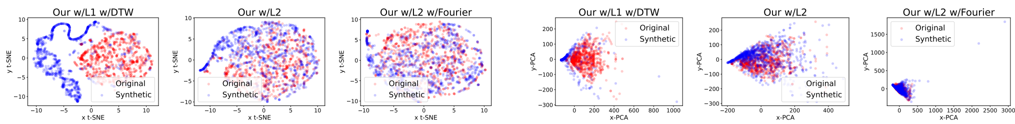
Figure 8 shows the t-SNE and PCA plots of our best model against the state-of-the-art approaches. It can be observed that the distribution of the synthetic samples closely resembles the real one in two of the three datasets (Hénon and Rivers). This visually ensures that the model is generating realistic time-series in a diverse set of fields. Figure 9 show the dimensionality reduction results for the other three variants of our model, namely (OUR W/L1 W/DTW, OUR W/L2, OUR W/L2 W/FOURIER), on all considered datasets.

A.4.4 INFERENCE TIME

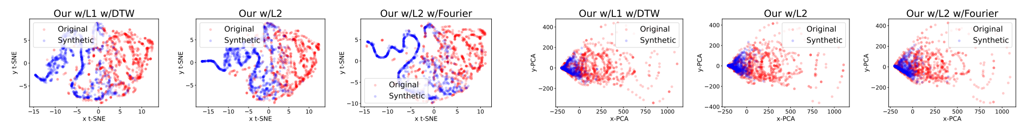
In more detail, Table 7 shows the inference time of the models isolating the generation of the time-series and the extraction of the graph. It turns out that even if CAUSALTIME is faster than OUR in generating the time-series, the graph extraction through DeepSHAP introduces an important overload making it the slowest model. We run this experiment on a machine equipped with Intel Core i9-10920X CPU @ 3.50GHz, NVIDIA GeForce RTX 2060 GPU, and 8×32 GB DDR4 RAM.



(a) Hénon



(b) Rivers



(c) AQI

Figure 9: Dimensionality reduction: t-SNE (left) and PCA (right).

Table 7: Inference time.

Dataset	Model				
	OUR	OUR with Graph	CAUSALTIME	CAUSALTIME with Graph	CR-VAE
Hénon	1481ms	1548ms	465ms	8790ms	194ms
Rivers	1425ms	1492ms	235ms	4248ms	148ms
AQI	1395ms	1535ms	1349ms	205s	442ms

A.4.5 EVALUATION METRICS DURING TRAINING

Figures 10 to 12 show the evolution of the evaluation metrics during training on the Hénon, Rivers, and AirQuality datasets, respectively.

A.5 ALGORITHMS

We show the algorithm to reconstruct the whole time-series from the output of DEN_θ (i.e. the initial time-steps $\boldsymbol{x}_{start}$ and the set of coefficients $\boldsymbol{c}$) in Algorithm 1.

The sampling procedure of a synthetic couple $\langle \hat{\boldsymbol{x}}, \hat{g} \rangle$ is described in Algorithm 2.

Algorithm 1 Reconstruction of $\hat{\boldsymbol{x}}$ from $\boldsymbol{x}_{start}$ and $\boldsymbol{c}$.

function RECONSTRUCT($\boldsymbol{x}_{start}, \boldsymbol{c}$)
 $\triangleright$ $\boldsymbol{x}_{start}$.shape = $[d, \ \tau_{max}]$
 $\triangleright$ $\boldsymbol{c}$.shape = $[d, \ d \cdot \tau_{max}, \ L - \tau_{max}]$
 $\hat{\boldsymbol{x}}_0 = \boldsymbol{x}_{start}$
 for all i from 0 to $L - \tau_{max}$ **do**
  sup $\leftarrow \hat{\boldsymbol{x}}_0$[: , $-\tau_{max}$:].flatten()
  $c \leftarrow \boldsymbol{c}$[: , : , i]
  $x \leftarrow$ torch.einsum('a,ba->b', sup, c)
  $\hat{\boldsymbol{x}}_0 \leftarrow$ torch.cat([$\hat{\boldsymbol{x}}_0$, x.unsqueeze(-1)], dim=-1)
 end for
 return $\hat{\boldsymbol{x}}_0$
end function

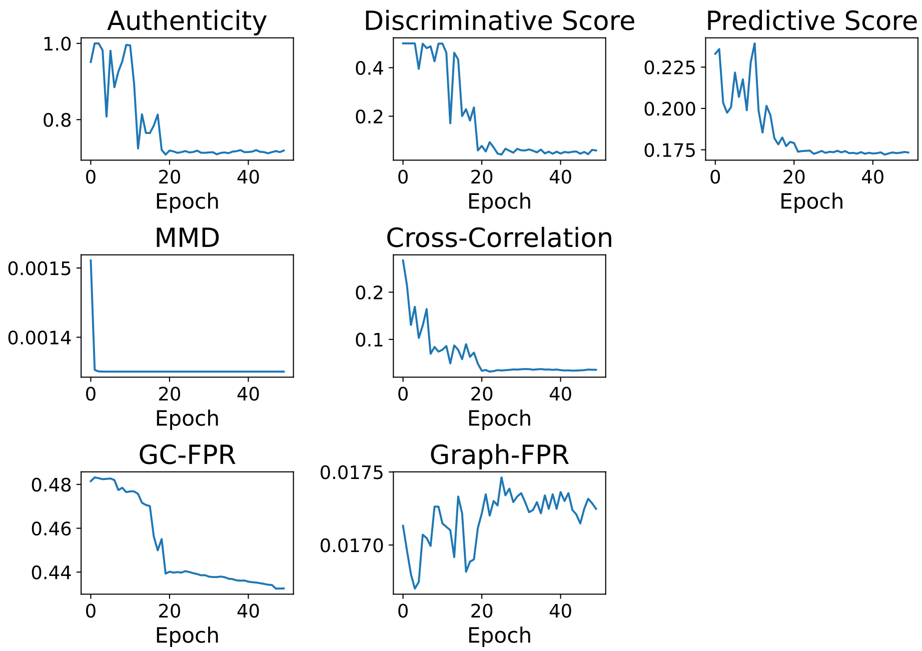


Figure 10: Evaluation metrics during training - Hénon dataset.

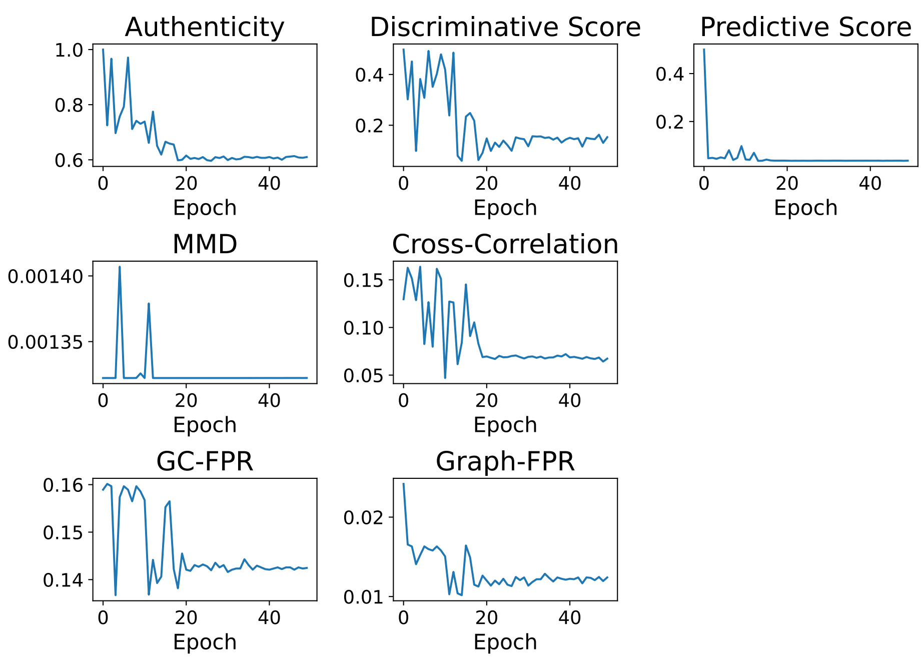


Figure 11: Evaluation metrics during training - Rivers dataset.

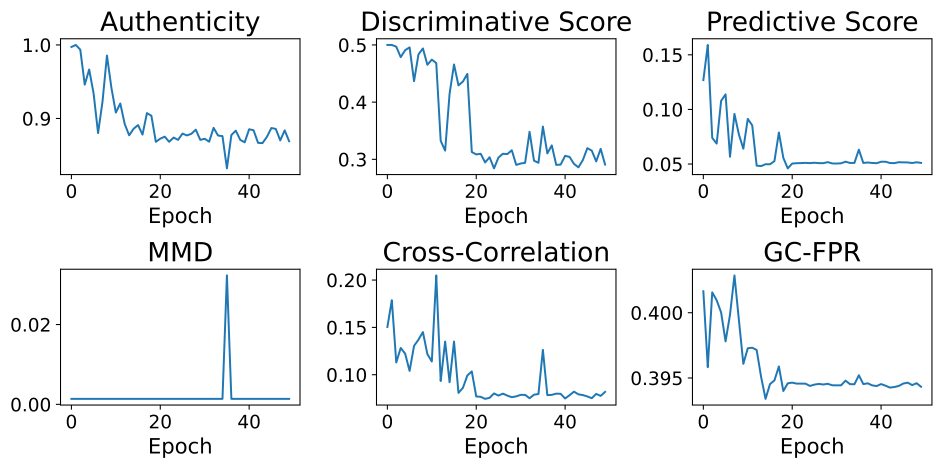


Figure 12: Evaluation metrics during training - AirQuality dataset.

Algorithm 2 Sampling of $\langle \hat{\boldsymbol{x}}, \hat{g} \rangle$.

Require: Trained denoising network DEN_θ
  $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  for all t from T to 0 **do**
    $(\boldsymbol{x}_{\text{start}}, \boldsymbol{c}) \leftarrow \text{DEN}_\theta\,(\boldsymbol{x}_t, t)$
    $\hat{\boldsymbol{x}}_0 \leftarrow \text{RECONSTRUCT}\,(\boldsymbol{x}_{\text{start}}, \boldsymbol{c})$
    $\boldsymbol{x}_{t-1} \leftarrow \beta_t \cdot \frac{\sqrt{\hat{\alpha}_{t-1}}}{1-\hat{\alpha}_t} \cdot \hat{\boldsymbol{x}}_0 + \frac{(1-\hat{\alpha}_{t-1}) \cdot \sqrt{\alpha_t}}{1-\hat{\alpha}_t} \cdot \boldsymbol{x}_t$
    if $t > 0$ **then**
      $\boldsymbol{x}_{t-1} \leftarrow \boldsymbol{x}_{t-1} + \beta_t \cdot \frac{1-\hat{\alpha}_{t-1}}{1-\hat{\alpha}_t} \cdot \boldsymbol{\epsilon}$
    end if
  end for
  return $\hat{\boldsymbol{x}}_0$

A.6 TSCD ALGORITHMS BENCHMARK

A.6.1 DETAILS ON THE ALGORITHMS

To evaluate the different TSCD algorithms we adapt/test them to our task using their source available code, whose repositories are listed below.

- GC: Granger Causality test implemented in the statsmodels[9] library.
- DYNOTEARS: `https://github.com/mckinsey/causalnex`
- NTS-NOTEARS: `https://github.com/xiangyu-sun-789/NTS-NOTEARS`
- PCMCI+: `https://github.com/jakobrunge/tigramite`
- Rhino: `https://github.com/microsoft/causica`
- CUTS / CUTS+: `https://github.com/jarrycyx/UNN`
- Neural-GC: `https://github.com/iancovert/Neural-GC`
- NGM: `https://github.com/alexisbellot/Graphical-modelling-continuous-time`
- LCCM: `https://github.com/edebrouwer/latentCCM`
- eSRU: `https://github.com/sakhanna/SRUforGCI`
- TCDF: `https://github.com/M-Nauta/TCDF`

The used hyper-parameters of the algorithms are reported in Table 8 (they are the same for all datasets).

Table 8: Hyper-parameters of the causal discovery algorithms.

Algorithm	Hyper-parameter	Value
GC	maxlag	2
DYNOTEARS	p max_iter	2 100
NTS-NOTEARS	lags w_threshold h_tol	2 0.3 $1e-60$
PCMCI+	τ_{max} PC_{α}	2 0.01
Rhino	Noise Distribution init_rho init_alpha	Gaussian 30 0.2
CUTS	Input step λ τ	2 0.1 $0.1 \to 1$
CUTS+	Input step λ τ	2 0.01 $0.1 \to 1$
Neural-GC	Learning rate λ_{ridge} λ	0.05 0.01 $0.002 \to 0.02$
NGM	Steps Horizon GL_reg	500 5 0.1
LCCM	hidden_size Learning rate	20 0.01
eSRU	μ_1 Learning rate Batch size Epochs	1 0.005 30 500
TCDF	τ Epochs Learning rate	10 1000 0.01

[9]https://www.statsmodels.org/stable/index.html

A.6.2 OTHER RESULTS

Table 9 shows the results of our benchmark on a synthetic dataset where the causal graphs are extracted globally, following the procedure in Section 4.3.

Table 9: Other results of the benchmark of Causal Discovery Algorithms. Bold and underline are used to highlight the best and the second best result, respectively.

Method	**AUROC**			**AUPRC**		
	Hénon	Rivers	AQI	Hénon	Rivers	AQI
GC	0.55 ± 0.10	0.73 ± 0.16	0.50 ± 0.00	0.45 ± 0.11	0.54 ± 0.09	0.48 ± 0.08
DYNOTEARS	0.45 ± 0.11	0.52 ± 0.08	0.50 ± 0.00	0.52 ± 0.15	0.56 ± 0.08	$\underline{0.51 \pm 0.02}$
NTS-NOTEARS	0.64 ± 0.14	0.73 ± 0.15	0.50 ± 0.00	0.40 ± 0.13	0.55 ± 0.14	0.30 ± 0.23
PCMCI+	$\mathbf{0.84 \pm 0.08}$	$\underline{0.82 \pm 0.08}$	$\mathbf{0.68 \pm 0.00}$	0.54 ± 0.09	0.64 ± 0.08	0.50 ± 0.03
Rhino	0.50 ± 0.02	0.57 ± 0.12	0.50 ± 0.00	0.52 ± 0.01	0.65 ± 0.10	0.51 ± 0.03
CUTS	0.81 ± 0.10	$\mathbf{0.86 \pm 0.09}$	$\underline{0.68 \pm 0.01}$	$\underline{0.54 \pm 0.07}$	0.55 ± 0.08	$\underline{0.51 \pm 0.02}$
CUTS+	0.81 ± 0.09	0.75 ± 0.09	0.67 ± 0.01	0.53 ± 0.07	0.58 ± 0.08	$\underline{0.51 \pm 0.02}$
Neural-GC	0.67 ± 0.00	0.52 ± 0.07	0.50 ± 0.01	0.52 ± 0.01	0.53 ± 0.05	0.48 ± 0.10
NGM	$\underline{0.84 \pm 0.13}$	0.80 ± 0.13	0.50 ± 0.01	$\mathbf{0.63 \pm 0.16}$	$\mathbf{0.81 \pm 0.12}$	0.47 ± 0.13
LCCM	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.51 ± 0.00	$\underline{0.78 \pm 0.00}$	0.21 ± 0.00
eSRU	0.50 ± 0.0	0.71 ± 0.10	0.50 ± 0.00	0.53 ± 0.01	0.76 ± 0.08	$\mathbf{0.53 \pm 0.01}$
TCDF	0.50 ± 0.0	0.50 ± 0.01	0.50 ± 0.00	0.50 ± 0.03	0.53 ± 0.09	0.45 ± 0.15