Causal Discovery for Irregularly Time Series with Consistency Guarantees

Weihong Li [1] **Baohong Li** [1] **Anpeng Wu** [1] **Zhihan Li** [2] **Ming Ma** [2] **Keting Yin** [1] **Kun Kuang** [1]

Abstract

This paper studies causal discovery in irregularly sampled time series—a key challenge in risk-sensitive domains like finance, healthcare, and climate science, where missing data and inconsistent sampling frequencies distort causal mechanisms. The main challenge comes from the interdependence between missing data imputation and causal structure recovery: errors in imputation and structure learning can reinforce each other, leading to an inaccurate causal graph. Existing methods either impute first and then discover, or jointly optimize both via neural representation learning, but lack explicit mechanisms to ensure mutual consistency of imputation and structure learning. We address this challenge with ReTimeCausal, an EM-based framework that alternates between imputation and structure learning, which encourages structural consistency throughout the optimization process. Our framework provides theoretical consistency guarantees for structure recovery and extends classical results to settings with irregular sampling and high missingness. ReTimeCausal combines kernel-based sparse regression and structural constraints in an alternating process that updates the completed data and the causal graph in turn. Experiments on synthetic and real-world datasets show that ReTimeCausal is more effective than existing methods under challenging irregular sampling and missing data.

1. Introduction

Understanding causal relationships in time series is crucial in domains such as finance, healthcare, and climate science. Unlike correlation-based analysis, causal discovery aims to identify cause-and-effect relationships and temporal mechanisms that govern complex systems (Gong et al., 2024; Assaad et al., 2022b). For instance, in finance, it helps predict market trends and manage risks (Sokolov et al., 2025; Sadeghi et al., 2023); in healthcare, it aids in developing effective medical interventions (Srikishan & Kleinberg, 2023; Smit et al., 2023; Sanchez et al., 2022); and in climate science, it enables better climate modeling and policy-making (Runge et al., 2019a). However, most existing causal discovery methods assume regularly sampled and fully observed data (Liu et al., 2023; Pamfil et al., 2020; Sun et al., 2023), which is rarely the case in real-world scenarios.

Such assumptions hinder practical causal discovery, especially in domains where time series frequently suffer from irregular sampling and missing data (Li & Marlin, 2020; Liu et al., 2025; Liu & Constantinou, 2023). For example, as shown in Figure 1(a), in the medical field, patient data may be collected at irregular intervals due to varying visit schedules or incomplete records. A patient's vital signs might be measured hourly during a hospital stay but only weekly during follow-up visits, and some data points may be missing altogether (Chauhan et al., 2024; Patharkar et al., 2024). This inconsistency in data collection frequency and completeness makes it difficult to apply existing causal discovery methods that assume regularly sampled and fully observed data. Coarsely aligning such data to a uniform time grid or ignoring missing values can distort temporal order and causal structure. Traditional causal discovery algorithms, which rely on the assumption of fully observed and regularly sampled sequences, can lead to incorrect identification of causal relationships. In the medical context, this could result in providing inaccurate treatment recommendations, potentially delaying effective medical interventions and worsening a patient's condition.

Existing causal discovery strategies are often inadequate when dealing with time series data that exhibit irregular sampling and missing values. For example, as shown in Figure 1(b), traditional impute-then-discover pipelines, which first fill in missing values and then apply standard structure learning algorithms, typically overlook the causal context during the imputation phase. For example, consider a case in the medical field where a patient's vital signs are measured hourly during a hospital stay but only weekly during follow-up visits, and some data points may be missing altogether. Simple interpolation may introduce artificial correlations. For instance, if antihypertensive medication dosage (X_1) is

[1]Zhejiang University, Hangzhou, China [2]Kuaishou Technology, Beijing, China. Correspondence to: Keting Yin <yinkt@zju.edu.cn>, Kun Kuang <kunkuang@zju.edu.cn>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

only recorded monthly, while blood pressure measurements (X_2) are taken weekly, simple interpolation may falsely create a direct link from medication to stroke risk score (X_3), bypassing the mediating influence of blood pressure. Similarly, if blood pressure (X_2) itself is sparsely observed, naive interpolation could lead to a spurious edge from stroke risk (X_3) to an unrelated variable such as LDL cholesterol (X_4), thereby distorting the true underlying causal pathway. Alignment-based approaches introduce additional pitfalls. As shown in Figure 1(c), these methods align all variables to a common temporal grid, typically at a coarse resolution like daily or monthly, without accounting for their native sampling frequencies. This coarse alignment can obscure the true temporal ordering of events and distort causal directionality. For example, suppose X_2 is recorded on January 10 and X_3 on January 20. If both are mapped to the same monthly time bin, they may be treated as simultaneous, which can mislead the discovery algorithm to infer a reversed causal relationship, with $X_3 \to X_2$ instead of the true $X_2 \to X_3$. Such distortions are especially problematic in irregularly sampled data, where variables often evolve at different temporal scales and even small timing discrepancies may carry significant causal information. While recent iterative frameworks like CUTS and CUTS+ attempt to jointly impute and discover causal structures under irregular sampling (Cheng et al., 2023; 2024a), they rely on recurrent neural networks. Although flexible, these neural models may provide limited interpretability.

To address the challenges of causal discovery in irregular and incomplete time series, we propose **ReTimeCausal** (**Re**covery for Irregular **Time**-series **Causal** Discovery), a framework that jointly performs missing data imputation and causal structure learning. Motivated by the tight coupling between these tasks, ReTimeCausal employs an iterative refinement process in which imputations and causal graphs are updated in tandem to ensure mutual consistency.

Specifically, ReTimeCausal formulates the learning process as an Expectation-Maximization (EM) algorithm, alternating between structure-aware data imputation (E-step) and sparse causal graph estimation (M-step). The framework accommodates both linear and nonlinear dynamics, captures lag-specific interactions, and ensures interpretability through sparsity-inducing regularization. We further provide a theoretical guarantee that the proposed joint method consistently recovers the true causal graph under given assumptions. The contributions of this work are summarized as follows:

- **A principled EM-based framework** for structure learning under severe missingness and temporal irregularity.
- **Theoretical guarantees** for causal discovery under irregular sampling and missing data, including consistency proofs in both idealized and realistic settings.
- **A structure-aware imputation strategy** that ensures consistency with the evolving causal graph through noise-aware completion.
- **Extensive empirical validation** on synthetic and real-world datasets shows our method outperforms baselines across the evaluated settings (e.g., achieving an F1 score of 0.463, exceeding the best baseline's 0.414 on CausalRivers).

2. Related work

Causal discovery with missing data. Several approaches aim to identify causal structures in the presence of missing values by modeling the dependencies between variables and their missingness. Graph-based methods such as m-graphs (Bhattacharya et al., 2020; Tu et al., 2019) provide recoverability conditions under MCAR, MAR, and some MNAR settings, but are limited to static contexts and rely heavily on conditional independence testing. Additive noise model (ANM)-based methods like MissDAG(Gao et al., 2022) uses a penalized EM framework to jointly learn the DAG structure and model parameters from incomplete observational data under ignorable missingness, but does not naturally extend to time-series settings with temporal dependence. While our approach is different in both modeling and estimation, it similarly adopts an EM-style alternating process to handle irregular time series.

Joint imputation and causal discovery in irregular time series. Recent methods like CUTS (Cheng et al., 2023) and CUTS+ (Cheng et al., 2024a) integrate causal discovery with imputation in irregularly sampled time series using neural architectures. While effective, these models provide limited interpretability and theoretical guarantees. Neural Graphical Modeling (Bellot et al., 2022) uses continuous-time Neural ODEs to avoid interpolation, but assumes infinitesimal-time dependencies. In contrast, our method targets a discretized setting rather than explicitly modeling continuous-time gaps or using elapsed time as an input. Irregular or asynchronous observations are represented on a fine time grid as partially observed values, enabling us to recover interpretable lag-specific graphs via sparse regression over discrete lagged variables, while providing theoretical guarantees under irregular sampling and high missingness. An extended discussion is provided in Appendix A.

3. Problem Setup

ReTimeCausal aims to uncover the underlying causal structure in multivariate time series data characterized by missing values and temporal irregularities. To handle such data

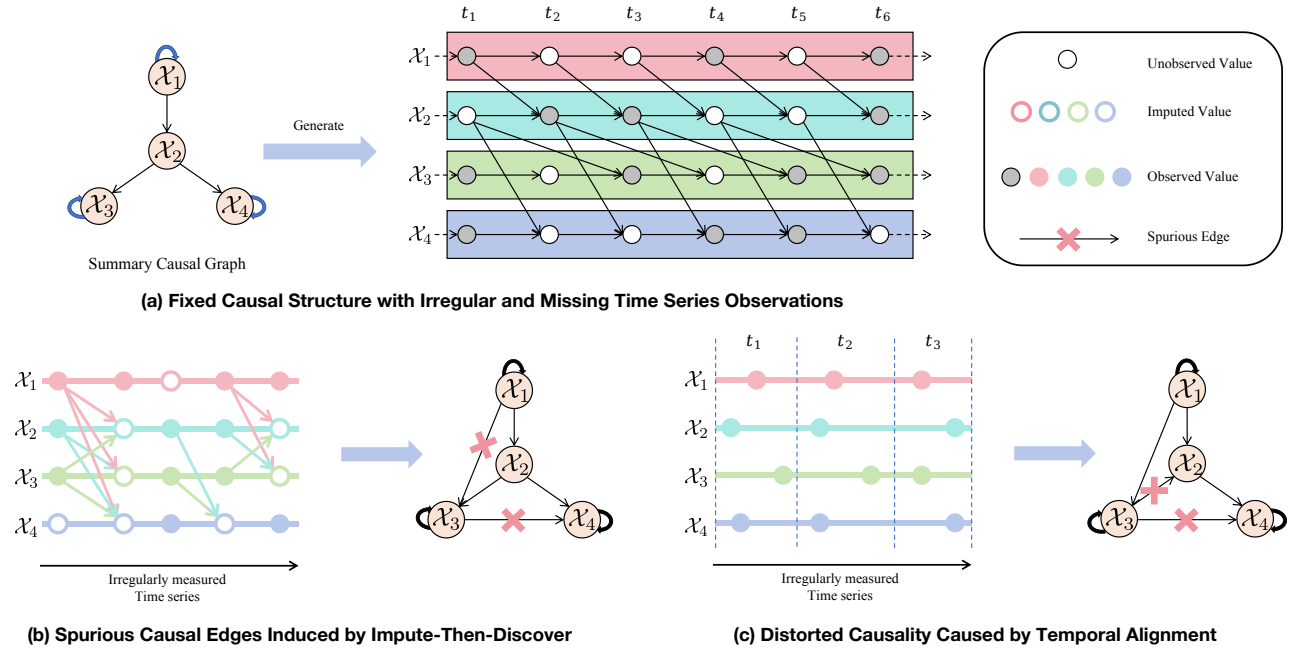

Figure 1. (a) A true causal graph generates irregularly sampled time series with missing data. (b) Impute-then-discover methods fill in missing values without causal context (e.g., by using all observed variables at the previous timestep), which may introduce spurious edges (e.g., $X_1 \to X_3$). (c) Temporal alignment maps variables to a coarse time grid, which distorts event order and potentially reversing causality ($X_3 \to X_2$).

within a unified framework, we adopt a discretized representation throughout this paper, where irregular or asynchronous observations are placed on a sufficiently fine discrete time grid and treated as partially observed values, rather than being modeled as native continuous-time trajectories. As formalized in Assumption 3.2, we assume an ignorable missingness mechanism (MCAR or MAR), consistent with prior work such as CUTS, CUTS+ and Miss-DAG (Cheng et al., 2023; 2024a; Gao et al., 2022). Under these assumptions, it is theoretically justified that unbiased causal inference can be achieved without explicitly modeling the missingness process. We further assume that the observed data is generated from a structural causal model with finite-order temporal dependencies, which forms the foundation for our method's design and analysis.

The Additive Noise Model (ANM) is a widely adopted structural causal framework in the causal discovery literature, where each variable is expressed as a function of its causal parents plus an independent noise term (Hoyer et al., 2008; Uemura et al., 2022; Gao et al., 2022; Montagna et al., 2023). At each time step t, the observed variables $\mathbf{X}^t = \{X_i^t\}_{i=1}^d$ are defined as:

$$X_i^t = f_i(\mathbf{Pa}_i^t) + \epsilon_i^t \tag{1}$$

where $\mathbf{Pa}_i^t$ denotes the set of parent variables of X_i^t, and ϵ_i^t is an exogenous noise term, assumed independent of Pa_i^t and of other contemporaneous noise terms. This formulation supports both linear and nonlinear mechanisms and enables identifiability under standard conditions.

In real-world scenarios, $\mathbf{X}_o$ denotes the observed values of the multivariate time series, which may correspond to vital signs collected at high frequency during a patient's hospitalization (e.g., hourly measurements of heart rate or blood pressure), while $\mathbf{X}_m$ denotes the unobserved values induced by sparse or missing records, such as those arising during post-discharge follow-ups or from incomplete documentation. We denote the incomplete multivariate time series by $\mathcal{D} = \{(\mathbf{X}^t, \mathbf{M}^t)\}_{t=1}^T$, where $\mathbf{M}^t \in \{0,1\}^d$ is the observation mask at time t; equivalently, $\mathbf{X}_o$ and $\mathbf{X}_m$ refer to the observed and missing subsets induced by this mask.

Definition 3.1 (Summary Causal Graph). The summary causal graph $\mathcal{G}_S$ is a directed graph over the variable set $\{\mathcal{X}_1, \ldots, \mathcal{X}_d\}$, where each node $\mathcal{X}_i$ represents the i-th variable aggregated across time. There is a directed edge from $\mathcal{X}_j$ to $\mathcal{X}_i$ whenever there exist time indices $t_1 < t_2$ such that the underlying time-indexed graph contains a direct causal arrow $X_j^{t_1} \to X_i^{t_2}$. Self-loops are allowed when a variable influences its own future values.

Assumption 3.2 (Ignorable Missingness). The missingness mechanism is said to be ignorable if it falls under either the Missing Completely at Random (MCAR) or Missing at Random (MAR) conditions. In such cases, the probability that an entry is missing depends only on the observed data,

and not on the unobserved (missing) values themselves.

Assumption 3.3 (Finite-Order Markov Property). The time series follows a finite-order Markov process of order L, in which the state at time $t+1$ depends only on the most recent L time steps:

$$P(\mathbf{X}^{t+1} \mid \mathbf{X}^t, \mathbf{X}^{t-1}, ..., \mathbf{X}^1) = P(\mathbf{X}^{t+1} \mid \mathbf{X}^t, ..., \mathbf{X}^{t-L+1}).$$

Assumption 3.4 (Consistency Throughout Time, Definition 7 in Assaad et al. (2022b)). The summary causal graph $\mathcal{G}_S$ for a multivariate time series $\mathbf{X}$ is said to be consistent throughout time if the causal directions encoded in $\mathcal{G}_S$ remain invariant over time.

Assumption 3.5 (Causal Sufficiency). A set of variables is said to be causally sufficient if all common causes of the observed variables are observed:

$$\forall i \neq j, \nexists Z \notin \{X_1^t, ..., X_d^t\}_t \text{ s.t.} Z \to X_i^{t_1} \wedge Z \to X_j^{t_2}$$

Assumption 3.6 (Faithfulness). The data-generating process is faithful to the true summary causal graph $\mathcal{G}_S$. For any two variables $\mathcal{X}_i$ and $\mathcal{X}_j$, and any subset $S \subseteq \{\mathcal{X}_1, \ldots, \mathcal{X}_d\} \setminus \{\mathcal{X}_i, \mathcal{X}_j\}$, if $\mathcal{X}_i \not\perp\!\!\!\perp \mathcal{X}_j \mid S$, then $\mathcal{X}_i$ and $\mathcal{X}_j$ are connected in $\mathcal{G}_S$ given S.

The ANM is a standard and widely used framework for causal discovery with fully observed data, and it can represent both linear and nonlinear relationships. However, most existing theoretical results focus on settings with complete observations and regular sampling. A few recent studies consider causal discovery with missing data under restrictive assumptions in static scenarios (Gao et al., 2022; Qiao et al., 2024), but theoretical guarantees for irregularly sampled time series remain scarce. Under our modeling assumptions, we bridge this gap by establishing a formal correctness result in this more general setting: Proposition 4.1 shows that our EM-style joint optimization recovers the true causal structure in the asymptotic regime.

4. ReTimeCausal

Motivated by the limitations of existing methods in handling missing and irregular time series, we propose **ReTimeCausal**, a framework that integrates Additive Noise Models (ANMs) with an EM-style optimization to jointly perform imputation and causal discovery. The goal is to recover sparse, interpretable, lagged causal graphs via explicit structural constraints and alternating updates.

As shown in Figure 2, ReTimeCausal consists of three components: (1) an *imputation module* in which neural structural functions estimate missing values in the E-step; (2) a *regression module* in which kernelized sparse regression operates on the completed data to recover lag-specific causal coefficients in the M-step, followed by thresholding; and (3) a *pruning mechanism* that removes spurious edges. These steps iterate via an EM-style loop, progressively refining both data and structure to align with the underlying causal process.

4.1. EM-based Imputation

Unlike impute-then-discover pipelines, our EM-based design ensures structure-aware imputation. This joint treatment of imputation and structure learning is realized via an Expectation-Maximization method, where each iteration consists of two alternating steps:

First, in the **E-step**, given the current parameters $\theta^{(k)}$ and the estimated summary causal graph $\mathcal{G}_S^{(k)}$, we compute the expected complete-data log-likelihood:

$$Q(\theta, \mathcal{G}; \theta^{(k)}, \mathcal{G}_S^{(k)}) = \mathbb{E}_{q^{(k)}(\mathbf{X}_m)} \left[\log p(\mathbf{X}_o, \mathbf{X}_m \mid \theta, \mathcal{G}_S)\right]$$

where $q^{(k)}(\mathbf{X}_m) \triangleq p(\mathbf{X}_m \mid \mathbf{X}_o; \theta^{(k)}, \mathcal{G}_S^{(k)})$. Missing values X_i^t are imputed via their conditional expectations:

$$\hat{X}_i^t = \mathbb{E}[X_i^t \mid \mathbf{Pa}_i^t, X_{1:d}^{t-L:t-1}; \theta^{(k)}] \tag{2}$$

Under the ANM with zero-mean noise $\mathbb{E}[\epsilon_i] = 0$, the conditional expectation of each variable directly reduces to its corresponding structural function:

$$\hat{X}_i^t = f_i(\mathbf{Pa}_i^t) \tag{3}$$

Second, in the **M-step**, the parameters of the structural functions are optimized by maximizing the expected complete-data log-likelihood under the current graph estimate:

$$\theta^{(k+1)} = \arg\max_{\theta} Q(\theta, \theta^{(k)}, \mathcal{G}_S^{(k)})$$

The causal graph is subsequently estimated from the imputed data:

$$\mathcal{G}_S^{(k+1)} = \mathcal{S}\left(\tilde{\mathbf{X}}^{(k)}\right)$$

Here, $\tilde{\mathbf{X}}^{(k)}$ denotes the completed time series at iteration k, formed by keeping observed values and filling missing values with the current imputation estimates:

$$\tilde{\mathbf{X}}^{(k)} \triangleq \mathbf{M} \odot \mathbf{X}_o + (\mathbf{1} - \mathbf{M}) \odot \hat{\mathbf{X}}^{(k)},$$

where $\mathbf{M} \in \{0,1\}^{T \times d}$ is the observation mask and $\odot$ denotes the element-wise product. The operator $\mathcal{S}(\cdot)$ denotes the structure learning map, implemented via kernelized sparse regression and further refined through binarization and pruning under the assumed causal mechanism.

4.2. Kernelized Sparse Regression for Nonlinear Causal Discovery

To fully utilize the potential of EM in nonlinear causal settings, we integrate kernel-based sparse regression into the

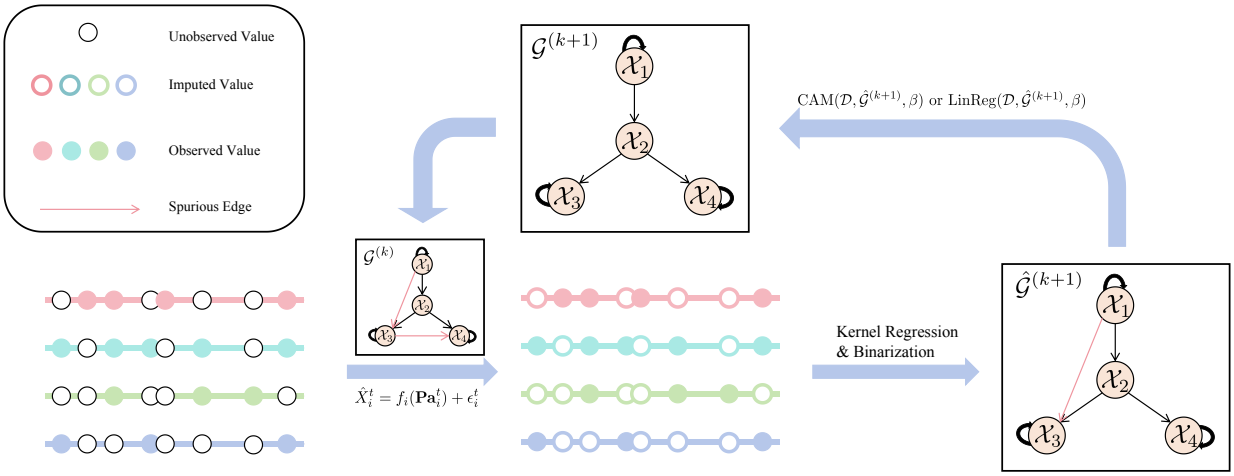

Figure 2. Overview of the ReTimeCausal Framework

M-step. This section presents the theoretical rationale for this integration and introduces kernel regression for modeling complex dependencies.

4.2.1. EM-BASED ANM

To capture nonlinear causal dependencies within the EM framework introduced in Section 4.1, we deliberately assign different roles to the two learning components. The E-step uses neural structural functions to estimate conditional expectations and impute missing values, while the M-step uses kernel-based regression to recover lag-specific coefficients for graph construction. We adopt kernel-based regression in the structure-learning stage because it enables universal function approximation (Schölkopf & Smola, 2002) while still providing an explicit coefficient representation that can be thresholded, projected, and pruned into an interpretable graph. This also preserves ANM identifiability under non-Gaussian noise (Hoyer et al., 2008).

Formally, each input vector $\tilde{\mathbf{X}}^t \in \mathbb{R}^d$ is mapped into a high-dimensional feature space via a kernel embedding $\phi : \mathbb{R}^d \to \mathcal{H}$, where $\mathcal{H}$ denotes an RKHS, and inner products $\langle \phi(x), \phi(x') \rangle$ are computed using a Mercer kernel $\kappa(x, x')$. For computational tractability, we approximate this mapping using a finite-dimensional representation $\phi(\tilde{\mathbf{X}}^t) \in \mathbb{R}^p$, where p is the number of kernel features determined by the approximation method. The temporal dynamics are modeled linearly in the feature space as:

$$\tilde{\mathbf{X}}^t = \sum_{\tau=1}^{L} \phi(\tilde{\mathbf{X}}^{t-\tau}) \mathbf{W}_{\text{high}}^{(\tau)} + \boldsymbol{\epsilon}^t \tag{4}$$

where each $\mathbf{W}_{\text{high}}^{(\tau)} \in \mathbb{R}^{p \times d}$ is a lag-specific transformation matrix. This formulation implies that the observed variable X_i^t depends nonlinearly on past multivariate lags $\mathbf{X}^{t-\tau}$, thereby supporting a broad class of nonlinear autoregressive models (NARs). The design is expressive enough to capture both short-term and long-range dependencies.

4.2.2. STRUCTURE RECOVERY AND SPARSITY

Since kernelized operations are not directly interpretable in the original input space, we introduce a projection mechanism to recover input-level causal structures. In our implementation, $\Phi \in \mathbb{R}^{d \times p}$ is a deterministic back-projection matrix induced by the Random Fourier Feature (RFF) map used to approximate the RBF kernel, rather than an additional learned parameter. Its entry $\Phi_{j,r}$ describes how the r-th feature-space coefficient contributes to the recovered influence associated with input dimension j. The reconstructed influence of variable $X_j^{t-\tau}$ on X_i^t at lag τ is therefore approximated by

$$\hat{W}_{j,i}^{(\tau)} = \sum_{r=1}^{p} \Phi_{j,r} \cdot W_{\text{high},r,i}^{(\tau)}. \tag{5}$$

The collection $\{\mathbf{W}^{(\tau)}\}_{\tau=1}^{L}$ defines an interpretable approximate summary of lagged structural causal relationships under the ANM. We emphasize that Eq. 5 provides an approximate feature-to-input back-projection induced by the chosen RFF map, rather than an exact variable-wise decomposition of the nonlinear kernel model. The resulting input-space coefficients support thresholding and pruning, thereby improving sparsity and interpretability.

We minimize a composite loss that balances prediction accuracy in RKHS with structural sparsity:

$$\min_{\mathbf{W}_{\text{high}}} \sum_{t=L}^{T} \left\| \tilde{\mathbf{X}}^t - \sum_{\tau=1}^{L} \phi(\tilde{\mathbf{X}}^{t-\tau}) \mathbf{W}_{\text{high}}^{(\tau)} \right\|_2^2 + \lambda \sum_{\tau=1}^{L} \|\mathbf{W}_{\text{high}}^{(\tau)}\|_1. \tag{6}$$

The ℓ_1 regularization acts on the feature-space coefficients $\mathbf{W}_{\text{high}}^{(\tau)}$, encouraging sparsity in the kernel representation.

However, this penalty alone does not guarantee variable-level sparsity after back-projection. In our pipeline, the final sparse causal graph is obtained only after projecting to input-space coefficients $\mathbf{W}^{(\tau)}$, followed by thresholding and pruning. This separation helps us distinguish weak or spurious dependencies from stable lag-specific causal effects even in the presence of nonlinear dynamics.

To construct the final causal graph, we apply a thresholding operation to the recovered weights. Specifically, for a given threshold $\gamma > 0$, we define the binary adjacency $\hat{\mathcal{G}}_S^{(k,\tau)} \in \{0,1\}^{d\times d}$ as:

$$\hat{\mathcal{G}}_S^{(k,\tau)} = \mathbb{I}\left(|\mathbf{W}^{(\tau)}| > \gamma\right) \tag{7}$$

This projection-based thresholding strategy removes weak or spurious dependencies, producing a sparse and interpretable lagged causal graph in the original input space. By recovering variable-level influence weights from the kernel feature space, it enables principled post-processing such as coefficient-based pruning and significance filtering, which in turn enhance structural clarity and robustness, particularly in high-dimensional or noisy settings.

Without loss of generality, when the kernel function degenerates to the identity function or when nonlinearities are mild, the entire model formulation reduces to a standard sparse linear regression over lagged variables. In such cases, the projected weight matrices $\mathbf{W}^{(\tau)}$ directly correspond to causal coefficients in the original input space, and structure learning becomes equivalent to sparse linear Granger modeling. To illustrate this special case and enhance interpretability, we provide a linear instantiation of our method in Appendix C.

4.3. Noise-Aware Imputation for Pruning-Consistent Recovery

Causal pruning techniques such as Causal Additive Models (CAM) (Bühlmann et al., 2014) rely on a fundamental assumption: the noise term must be statistically independent from its parent variables. However, this assumption can be violated when missing values are imputed deterministically using Eq. 3. In this case, the imputed values lie exactly on the regression manifold, effectively removing the noise term and introducing spurious dependence between $\hat{X}_i^t$ and its predictors. This undermines the post hoc pruning methods such as CAM, which may fail to identify false causal edges.

To prevent such artifacts, we propose a noise-aware imputation strategy for nonlinear models. Specifically, we modify the imputation step to explicitly inject noise during data completion:

$$\hat{X}_i^t = f_i(\mathbf{Pa}_i^t) + \epsilon_i^t \tag{8}$$

where ϵ_i^t is a noise term sampled from the empirical distribution of residuals $\hat{\epsilon}_i^t = X_i^t - \hat{f}_i(\mathbf{Pa}_i^t)$ to match the learned noise profile. This adjustment encourages the independence assumption between ϵ_i^t and $\mathbf{Pa}_i^t$, thereby making the downstream pruning process more statistically well-justified. Empirically, we find that this noise-injection mechanism improves pruning quality when applying CAM to kernel-based nonlinear models. It enables better separation between true and spurious causal links, particularly in high-dimensional settings with partial observability. The full pruning method, including CAM and linear-model variants, is detailed in Appendix D.

4.4. Iterative Optimization and Theoretical Analysis

This section provides implementation details of the underlying models, describes the iterative optimization method, and presents the theoretical guarantees that underpin ReTimeCausal.

4.4.1. Network and Model Implementation Details

To model complex causal dependencies, each structural function $f_i(\cdot)$ is parameterized as a three-layer feedforward neural network and optimized via gradient descent during the model update step. These neural networks are used to approximate the structural functions needed for the E-step imputation, rather than to directly parameterize graph edges. This design follows recent trends in causal discovery that integrate neural estimation into iterative learning methods (Wu et al., 2023; Lachapelle et al., 2020; Brouillard et al., 2020; Wang et al., 2020), enabling flexible modeling of both linear and nonlinear mechanisms. For nonlinear settings, we adopt the Gaussian Radial Basis Function (RBF) kernel, defined as:$\kappa(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$, where σ is a bandwidth hyperparameter. The RBF kernel is chosen for its universal approximation capabilities and smoothness properties, making it particularly well-suited for modeling nonlinear causal mechanisms under the ANM framework.

4.4.2. Iterative Optimization

ReTimeCausal operates as an alternating optimization method, where each iteration comprises two core phases: imputing missing values using current causal estimates, and updating model parameters and structure using the completed data. This cycle continues until convergence. To improve robustness across iterations, we apply exponential smoothing to the estimated weight matrices:

$$\mathbf{W}^{(k)} = \alpha\mathbf{W}^{(k-1)} + (1-\alpha)\hat{\mathbf{W}}^{(k)}, \tag{9}$$

where $\alpha \in [0,1]$ is a smoothing coefficient. We then apply thresholding to derive the binary graph $\mathcal{G}_S^{(k)}$, followed by model-specific pruning.

While ReTimeCausal adopts an imputation-update structure inspired by EM, it departs from the classical formulation by combining neural function estimation in the E-step with coefficient-based kernel structure recovery in the M-step. This modular separation provides both flexible nonlinear imputation and explicit graph extraction. A summary of the algorithmic method is provided in Appendix E.

4.4.3. Theoretical Convergence Guarantees

Proposition 4.1 (Structural Consistency)**.** *Let the variable sequence $\{\mathbf{X}^1, ..., \mathbf{X}^t\}$ be generated by a finite-order Markov process that satisfies Assumptions 3.2–3.6. If the regularization and thresholding parameters used in the algorithm decrease appropriately with the sample size, and the smoothing weight remains within a fixed range, then the estimated summary causal graph $\hat{\mathcal{G}}_S^{(k)}$ converges structurally to the true graph $\mathcal{G}_S^*$, satisfying:* $\lim_{k\to\infty} \lim_{n\to\infty} \Pr\left(\hat{\mathcal{G}}_S^{(k)} = \mathcal{G}_S^*\right) = 1.$

Proposition 4.1 should be interpreted as an asymptotic consistency result for the idealized EM-style alternating estimator underlying ReTimeCausal, rather than as a finite-sample or global convergence guarantee for every implementation detail of the practical pipeline. Specifically, it formalizes the principle that if the imputation step converges to the oracle conditional expectation and the structure estimation step is support consistent, then the recovered lagged graph converges to the true graph in the limit. The practical algorithm serves as a surrogate realization of this principle, with its empirical performance still affected by optimization and modeling choices. The proof builds on convergence results under missing data (Wu, 1983), consistency of universal function approximators (Schölkopf & Smola, 2002), and sparse structure recovery guarantees in high-dimensional regression (Wainwright, 2009). A formal proof is provided in Appendix G.1. Beyond this asymptotic result, we further discuss finite-sample and finite-iteration behavior in Appendix G.2.

5. Experiments

In this section, we evaluate the performance of ReTimeCausal on both synthetic and real-world datasets, under varying levels of missingness, graph sizes, and causal complexities. For clarity, we report only mean values in most result tables, as the standard deviations are below 0.001 and do not materially affect the comparative trends.

5.1. Baseline Algorithms

We compare ReTimeCausal against representative methods from different paradigms: PCMCI (Runge et al., 2019a) (constraint-based), DYNOTEARS (Pamfil et al., 2020) (score-based), CUTS+ (Cheng et al., 2024a) (iterative neural framework for irregular series), Rhino (Gong et al., 2023) (deep nonlinear temporal causal model), and AERCA (Han et al., 2025) (root-cause model with Granger-based structure learning). These baselines collectively cover classical causal discovery approaches as well as recent models tailored to irregular or high-dimensional time series. Since most methods require complete data, we additionally apply TimeMixer as a preprocessing step for fair comparison (Wang et al., 2024). Detailed descriptions of each baseline and evaluation metric are provided in Appendix H.

5.2. Experiments on Synthetic Data

We evaluate our proposed algorithm on synthetic time series data generated from Eq.1, incorporating both linear and nonlinear causal mechanisms. The data generation strictly follows the assumptions outlined in Section 3. The causal graph is generated using the Erdős–Rényi model, and weights are uniformly scaled to ensure numerical stability. Missing values are introduced randomly under the assumption 3.2, with missing rates (MR) ranging from 20% to 80%.

Following the design in Wu et al. (2024), we simulate causal mechanisms by applying element-wise transformations to parent variables before aggregation:

$$X_i^t = \sum_{\tau=1}^{L} \sum_{j=1}^{d} W_{ij}^{(\tau)} \cdot f_i(X_j^{t-\tau}) + \epsilon_i^t \tag{10}$$

We instantiate both linear and nonlinear settings within this general framework by varying the transformation function $f_i(\cdot)$ and the noise distribution. In both cases, additive noise is drawn from either Gaussian or Laplace distributions. In the linear setting, each variable is a weighted sum of its lagged parents. In the nonlinear setting, $f_i(\cdot)$ is selected from nonlinear functions such as $\sin$ or $\tanh$, and fix the function type for each experiment. This leads to more complex functional dependencies and stronger temporal entanglement. Similarly, we use [FunctionType]-[NoiseType]-[NumVars]-[NumEdges]-[MaxLag] to describe synthetic data configurations throughout the experiments.

As shown in Table 1, under linear-Gaussian dynamics with 10 and 20 variables, ReTimeCausal outperforms all competing approaches across all missing data rates. Specifically, in the 20-variable setting, it achieves an F1 score of 0.976 with fully observed data. When 60% of the values are missing, the F1 score remains 0.930, and it reaches 1.000 under 80% missingness. These results show that ReTimeCausal remains accurate even under high missingness. In contrast, the methods that rely on TimeMixer-imputed data show severe performance drop, with F1 scores falling below 0.2 in larger-scale settings with high missing rates. Although CUTS+ is designed for irregularly sampled data, it also fails

Table 1. Effect of Missing Data Rates on Model Performance in Synthetic Time Series

Methods	LR-gaussian-10-10-2			LR-gaussian-20-20-2		
	MR=0.0	MR=0.6	MR=0.8	MR=0.0	MR=0.6	MR=0.8
PCMCI+TimeMixer	0.889	0.400	0.294	0.857	0.152	0.141
PCMCI+EM	0.889	0.439	0.375	0.857	0.205	0.190
DYNOTEARS+TimeMixer	1.000	0.263	0.277	0.976	0.173	0.155
Rhino+TimeMixer	0.286	0.235	0.162	0.123	0.243	0.185
AERCA+TimeMixer	0.312	0.225	0.234	0.098	0.065	0.083
CUTS+	0.211	0.217	0.212	0.104	0.133	0.124
ReTimeCausal	**1.000**	**1.000**	**1.000**	**0.976**	**0.930**	**1.000**

Table 2. Scalability Evaluation on Nonlinear Synthetic Graphs of Increasing Size

Methods	TANH-laplace-10-20			TANH-laplace-20-40			TANH-laplace-50-100		
	SHD	SID	F1	SHD	SID	F1	SHD	SID	F1
PCMCI+TimeMixer	26	25	0.430	86	84	0.259	400	255	0.141
PCMCI+EM	15	27	0.545	43	110	0.318	124	216	0.311
DYNOTEARS*+TimeMixer	58	11	0.393	307	73	0.177	1717	84	0.092
Rhino+TimeMixer	43	15	0.308	263	55	0.165	253	258	0.087
AERCA+TimeMixer	52	22	0.312	194	55	0.192	1250	149	0.074
CUTS+	70	3	0.352	361	6	0.178	1145	235	0.085
ReTimeCausal	**4**	**2**	**0.909**	**6**	**4**	**0.923**	**39**	**39.5**	**0.816**

to sustain performance under such conditions, consistently yielding F1 scores below 0.13. This suggests the limitations of methods that rely on Granger-style assumptions and weak functional modeling, particularly when data quality is compromised by substantial missingness.

Table 2 further evaluates the scalability of ReTimeCausal under increasing graph sizes, with a fixed 60% missing rate and TANH-based nonlinear causal functions. For 10 variables, ReTimeCausal achieves an F1 score of 0.909, an SHD of 4, and a low SID of 2, which indicates strong alignment with both structural and interventional ground truth. As the graph expands to 20 and 50 variables, it maintains strong performance with F1 scores of 0.923 and 0.816, SHD values of 6 and 39, and SID values of 4 and 39.5. Meanwhile, all baseline methods exhibit significant performance degradation in both structural accuracy and interventional consistency. For instance, PCMCI+TimeMixer and PCMCI+EM report SID values exceeding 80 and 100 for the 20-variable setting, while CUTS+—despite reasonable SID scores in smaller graphs—reaches a SID of 235 for 50 variables, indicating instability in high-dimensional scenarios.

These results demonstrate that ReTimeCausal achieves accurate and scalable causal discovery in settings with high missingness, large variable dimensions, and nonlinear interactions. Its joint EM-style optimization and kernel-based modeling strategy allows it to avoid the structural inconsistencies and performance collapse observed in conventional impute-then-discover pipelines.

5.3. Experiments on Real-World Data

To evaluate our method on real-world data, we utilize a 10-node subgraph from the RiversEastGermany portion of the CausalRivers benchmark (Gideon et al., 2025), which includes a known ground truth causal graph. We simulate missing values by randomly removing 20% and 60% of the time-series data.

As shown in Table 3, although overall performance remains modest on this challenging real-world dataset, ReTimeCausal achieves the highest F1 scores at both missing rates (0.463 at 20% missingness and 0.414 at 60%), outperforming the strongest baseline by approximately 5 percentage points in each case. In addition, under 60% missingness, it attains the lowest SID (8.0), indicating comparatively strong interventional fidelity in the sparser regime. However, ReTimeCausal does not achieve the best SHD in these real-world experiments: PCMCI+TimeMixer attains the lowest SHD at both missing rates, and AERCA+TimeMixer attains the lowest SID at 20% missingness. We observe that PCMCI+TimeMixer performs competitively at 20% missingness (F1 = 0.414), but its advantage does not persist as the missing rate increases to 60%. These results suggest that ReTimeCausal is most effective in edge recovery as reflected by F1, while its advantage in SID is most evident under higher missingness.

Table 3. Empirical results (mean$_{\pm std}$) on CausalRivers Dataset

Methods	CausalRivers (MR = 0.2)			CausalRivers (MR = 0.6)		
	SHD	SID	F1	SHD	SID	F1
PCMCI+TimeMixer	**13.0**$_{\pm 0.00}$	54.0$_{\pm 0.00}$	0.414$_{\pm 0.01}$	**16.0**$_{\pm 0.00}$	59.0$_{\pm 0.00}$	0.273$_{\pm 0.00}$
PCMCI+EM	23.3$_{\pm 2.52}$	55.0$_{\pm 2.00}$	0.253$_{\pm 0.03}$	23.7$_{\pm 1.53}$	55.3$_{\pm 2.08}$	0.203$_{\pm 0.01}$
DYNOTEARS*+TimeMixer	59.0$_{\pm 1.00}$	43.0$_{\pm 0.00}$	0.392$_{\pm 0.00}$	47.3$_{\pm 0.58}$	13.0$_{\pm 0.00}$	0.355$_{\pm 0.00}$
Rhino+TimeMixer	45.2$_{\pm 12.1}$	34.8$_{\pm 8.41}$	0.260$_{\pm 0.03}$	51.4$_{\pm 3.21}$	30.0$_{\pm 9.00}$	0.265$_{\pm 0.06}$
AERCA+TimeMixer	47.4$_{\pm 1.67}$	**11.8**$_{\pm 3.35}$	0.400$_{\pm 0.02}$	51.0$_{\pm 2.00}$	18.0$_{\pm 4.00}$	0.354$_{\pm 0.03}$
CUTS+	31.0$_{\pm 1.00}$	16.0$_{\pm 0.00}$	0.367$_{\pm 0.00}$	38.0$_{\pm 0.00}$	35.0$_{\pm 0.58}$	0.319$_{\pm 0.00}$
ReTimeCausal	36.3$_{\pm 1.15}$	16.0$_{\pm 2.00}$	**0.463**$_{\pm 0.02}$	48.3$_{\pm 5.86}$	**8.0**$_{\pm 1.73}$	**0.414**$_{\pm 0.03}$

6. Conclusion

In this paper, we propose ReTimeCausal, an EM-augmented additive noise modeling framework for interpretable causal discovery in the presence of irregular time series data. ReTimeCausal integrates Expectation-Maximization with additive noise models, thereby allowing joint imputation and structure learning while preserving temporal resolution. Unlike impute-then-discover pipelines, our method learns lag-specific causal relationships that are noise-aware and interpretable. Theoretical analysis and empirical results demonstrate its robustness under linear and nonlinear settings with high missingness. Future work includes (1) extending ReTimeCausal to handle MNAR data via explicit modeling of missingness mechanisms, (2) improving computational efficiency through parallelizable or variational EM techniques, and (3) generalizing the framework to support latent confounders. We discuss limitations in Appendix L.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 62376243), the Key Research and Development Program of Zhejiang Province (2026C01021), and "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2025C02037). All opinions in this paper are those of the authors and donot necessarily reflect the views of the funding agencies.

Impact Statement

This work aims to advance causal discovery for irregularly sampled time series with missing values. It may benefit applications such as healthcare, finance, and climate analysis by providing more reliable and interpretable lag-specific causal graphs.

The main risk is misuse: if the method is applied outside its assumptions (e.g., unobserved confounding, non-random missingness, or distribution shift), the recovered graph may be misleading and could lead to inappropriate conclusions in downstream use. We therefore encourage practitioners to treat the output as supportive evidence, validate results with domain knowledge and robustness checks, and follow standard privacy practices when working with sensitive time series data.

References

Assaad, C. K., Devijver, E., and Gaussier, E. Entropy-based discovery of summary causal graphs in time series. *Entropy*, 24(8):1156, 2022a.

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

Balakrishnan, S., Wainwright, M. J., and Yu, B. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77 – 120, 2017. doi: 10.1214/16-AOS1435.

Basu, S. and Michailidis, G. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43:1535–1567, 08 2015. doi: 10.1214/15-AOS1315.

Bellot, A., Branson, K., and van der Schaar, M. Neural graphical modelling in continuous-time: consistency guarantees and algorithms. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=SsHBkfeRF9L.

Bhattacharya, R., Nabi, R., Shpitser, I., and Robins, J. M. Identification in missing data models represented by directed acyclic graphs. In Adams, R. P. and Gogate, V. (eds.), *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 1149–1158. PMLR, 22–25 Jul 2020. URL https://proceedings.mlr.press/v115/bhattacharya20b.html.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877, 2020.

Bühlmann, P. and Van De Geer, S. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Bühlmann, P., Peters, J., and Ernest, J. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526 – 2556, 10 2014. doi: 10.1214/14-AOS1260.

Chauhan, V. K., Thakur, A., O'Donoghue, O., Rohanian, O., Molaei, S., and Clifton, D. A. Continuous patient state attention model for addressing irregularity in electronic health records. *BMC Medical Informatics and Decision Making*, 24(1):117, 2024. doi: 10.1186/s12911-024-02514-2.

Cheng, Y., Yang, R., Xiao, T., Li, Z., Suo, J., He, K., and Dai, Q. CUTS: Neural causal discovery from irregular time-series data. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=UG8bQcD3Emv.

Cheng, Y., Li, L., Xiao, T., Li, Z., Suo, J., He, K., and Dai, Q. CUTS+: High-dimensional causal discovery from irregular time-series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11525–11533, 2024a.

Cheng, Y., Wang, Z., Xiao, T., Zhong, Q., Suo, J., and He, K. Causaltime: Realistically generated time-series for benchmarking of causal discovery. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=iad1yyyGme.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

Gao, E., Ng, I., Gong, M., Shen, L., Huang, W., Liu, T., Zhang, K., and Bondell, H. Missdag: causal discovery in the presence of missing data with continuous additive noise models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Gideon, S., Maha, S., Jan, B., Niklas, P., and Joachim, D. CausalRivers: Scaling up benchmarking of causal discovery for real-world time-series. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=wmV4cIbgl6.

Gong, C., Zhang, C., Yao, D., Bi, J., Li, W., and Xu, Y. Causal discovery from temporal data: An overview and new perspectives. *ACM Computing Surveys*, 57(4):1–38, 2024.

Gong, W., Jennings, J., Zhang, C., and Pawlowski, N. Rhino: Deep causal temporal relationship learning with history-dependent noise. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=i_1rbq8yFWC.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. *A distribution-free theory of nonparametric regression*. Springer, 2002.

Han, X., Absar, S., Zhang, L., and Yuan, S. Root cause analysis of anomalies in multivariate time series through granger causal discovery. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=JHC8jdmCrk.

Herman, R. J., Wahl, J., Ninad, U., and Runge, J. Unitless unrestricted markov-consistent scm generation: Better benchmark datasets for causal discovery. In Huang, B. and Drton, M. (eds.), *Proceedings of the Fourth Conference on Causal Learning and Reasoning*, volume 275 of *Proceedings of Machine Learning Research*, pp. 1506–1531. PMLR, 07–09 May 2025. URL https://proceedings.mlr.press/v275/herman25a.html.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

Lachapelle, S., Brouillard, P., Deleu, T., and Lacoste-Julien, S. Gradient-based neural dag learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rklbKA4YDS.

Li, S. C.-X. and Marlin, B. Learning from irregularly-sampled time series: A missing data perspective. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5937–5946. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/li20k.html.

Liu, M., Sun, X., Hu, L., and Wang, Y. Causal discovery from subsampled time series with proxy variables. In *Proceedings of the 37th International Conference on Neural*

Information Processing Systems, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Liu, M., Geißler, D., Bian, S., Zhou, B., and Lukowicz, P. Assessing the impact of sampling irregularity in time series data: Human activity recognition as a case study. In *2025 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pp. 311–316, 2025. doi: 10.1109/PerComWorkshops65533.2025.00083.

Liu, Y. and Constantinou, A. Improving the imputation of missing data with markov blanket discovery. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=GrpU6dxFmMN.

Lohse, C. and Wahl, J. Sortability of time series data. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=OGvmCpcHdV.

Merlevède, F., Peligrad, M., and Rio, E. A bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151, 03 2009. doi: 10.1007/s00440-010-0304-9.

Montagna, F., Noceti, N., Rosasco, L., Zhang, K., and Locatello, F. Causal discovery with score matching on additive models with arbitrary noise. In *Conference on Causal Learning and Reasoning*, pp. 726–751. PMLR, 2023.

Negahban, S., Ravikumar, P., Wainwright, M., and Yu, B. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 10 2012. doi: 10.1214/12-STS400.

Pamfil, R., Sriwattanaworachai, N., Desai, S., Pilgerstorfer, P., Georgatzis, K., Beaumont, P., and Aragam, B. DYNOTEARS: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pp. 1595–1605. PMLR, 2020.

Patharkar, A., Cai, F., Al-Hindawi, F., and Wu, T. Predictive modeling of biomedical temporal data in healthcare applications: review and future directions. *Frontiers in Physiology*, 15:1386760, 2024.

Qiao, J., Chen, Z., Yu, J., Cai, R., and Hao, Z. Identification of causal structure in the presence of missing data with additive noise model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18): 20516–20523, Mar. 2024. doi: 10.1609/aaai.v38i18.30036. URL https://ojs.aaai.org/index.php/AAAI/article/view/30036.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, NIPS'07, pp. 1177–1184, Red Hook, NY, USA, 2007. Curran Associates Inc. ISBN 9781605603520.

Reisach, A., Tami, M., Seiler, C., Chambaz, A., and Weichwald, S. A scale-invariant sorting criterion to find a causal order in additive noise models. *Advances in Neural Information Processing Systems*, 36:785–807, 2023.

Rosasco, L., Vito, E. D., Caponnetto, A., Piana, M., and Verri, A. Are loss functions all the same? *Neural Computation*, 16(5):1063–1076, 2004. doi: 10.1162/089976604773135104.

Rudi, A. and Rosasco, L. Generalization properties of learning with random features. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 3218–3228, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., van Nes, E. H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K., and Zscheischler, J. Inferring causation from time series in earth system sciences. *Nature Communications*, 10(1):2553, 2019a. doi: 10.1038/s41467-019-10105-3. URL https://doi.org/10.1038/s41467-019-10105-3.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11):eaau4996, 2019b. doi: 10.1126/sciadv.aau4996. URL https://www.science.org/doi/abs/10.1126/sciadv.aau4996.

Sadeghi, A., Gopal, A., and Fesanghary, M. Causal discovery in financial markets: A framework for nonstationary time-series data. *arXiv preprint arXiv:2312.17375*, 2023.

Sanchez, P., Voisey, J. P., Xia, T., Watson, H. I., O'Neil, A. Q., and Tsaftaris, S. A. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9(8):220638, 08 2022. ISSN 2054-5703. doi: 10.1098/rsos.220638. URL https://doi.org/10.1098/rsos.220638.

Schölkopf, B. and Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Smit, J., Krijthe, J. H., Kant, W., Labrecque, J., Komorowski, M., Gommers, D., van Bommel, J., Reinders, M. J., and van Genderen, M. E. Causal inference using

observational intensive care unit data: a scoping review and recommendations for future practice. *npj Digital Medicine*, 6(1):221, 2023.

Sokolov, A., Sabelli, F., Faraz, B. A., Li, W., and Seco, L. Toward automating causal discovery in financial markets and beyond. In *Transactions of ADIA Lab: Interdisciplinary Advances in Data and Computational Science*, pp. 359–392. World Scientific, 2025.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*. The MIT Press, 01 2001. ISBN 9780262284158. doi: 10.7551/mitpress/1754.001.0001. URL https://doi.org/10.7551/mitpress/1754.001.0001.

Srikishan, B. and Kleinberg, S. Causal discovery with stage variables for health time series. *arXiv preprint arXiv:2305.03662*, 2023.

Steinwart, I. and Christmann, A. *Support Vector Machines*. Information Science and Statistics. Springer New York, 2008. ISBN 9780387772417. URL https://books.google.co.jp/books?id=NQvYNAEACAAJ.

Sun, X., Schulte, O., Liu, G., and Poupart, P. NTS-NOTEARS: Learning nonparametric DBNs with prior knowledge. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 1942–1964. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/sun23c.html.

Tu, R., Zhang, C., Ackermann, P., Mohan, K., Kjellström, H., and Zhang, K. Causal discovery in the presence of missing data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1762–1770. PMLR, 2019. URL https://proceedings.mlr.press/v89/tu19a.html.

Uemura, K., Takagi, T., Takayuki, K., Yoshida, H., and Shimizu, S. A multivariate causal discovery based on post-nonlinear model. In *Conference on Causal Learning and Reasoning*, pp. 826–839. PMLR, 2022.

Wainwright, M. J. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

Wang, S., Wu, H., Shi, X., Hu, T., Luo, H., Ma, L., Zhang, J., and Zhou, J. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7oLshfEIC2.

Wang, Y., Menkovski, V., Wang, H., Du, X., and Pechenizkiy, M. Causal discovery from incomplete data: a deep learning approach. *arXiv preprint arXiv:2001.05343*, 2020.

Wu, A., Kuang, K., Xiong, R., Zhu, M., Liu, Y., Li, B., Liu, F., Wang, Z., and Wu, F. Learning instrumental variable from data fusion for treatment effect estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10324–10332, 2023.

Wu, A., Li, H., Kuang, K., Keli, Z., and Wu, F. Learning causal relations from subsampled time series with two time-slices. In *Forty-first International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53635–53655. PMLR, 2024. URL https://proceedings.mlr.press/v235/wu24p.html.

Wu, C. F. J. On the convergence properties of the em algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

Zhang, C.-H. and Zhang, S. S. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):217–242, 2014.

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

A. Related work

Graph-based identification methods aim to characterize the identifiability of causal structures under missing data by explicitly modeling the dependencies between variables and their missingness indicators. Approaches such as m-graphs (Bhattacharya et al., 2020; Tu et al., 2019) provide necessary and sufficient conditions for recoverability under MCAR, MAR, and some MNAR mechanisms. These methods enhance theoretical understanding by formalizing the graphical interplay between causal and missingness structures. However, they are mainly designed for static settings and rely on conditional independence testing, which becomes unreliable when data is sparse or temporally structured.

Additive noise model-based approaches extend structural causal models to handle missing values via generative modeling. For example, MissDAG (Gao et al., 2022) employs EM-style optimization to jointly learn causal graphs and impute missing data under MCAR/MAR assumptions, while Qiao et al. (2024) further extend this line by establishing identifiability under weak self-masking MNAR conditions. These methods support nonlinearity and partial observability, but are primarily designed for static data with complete conditional models. In contrast, our method extends this paradigm to irregular time series by explicitly modeling temporal dependencies through lagged causal effects and leveraging sparse kernel regression for interpretable nonlinear estimation. Moreover, we provide finite-sample and finite-iteration guarantees for structure recovery, complementing existing asymptotic identifiability results.

Joint imputation and causal discovery methods represent a recent advancement in handling irregular time series with missing data. Cheng et al. (2023) proposes an iterative framework that alternates between latent data prediction and sparse neural Granger graph estimation, enabling mutual refinement between imputed values and the learned causal structure. Cheng et al. (2024a) extends this paradigm to high-dimensional settings through a coarse-to-fine discovery strategy combined with message-passing GNNs, effectively addressing both feature redundancy and scalability. These models are among the few that jointly accommodate irregular sampling and high dimensionality, and outperform classical pipelines that separate imputation from structure learning.

Beyond these, several recent methods aim to tackle causal discovery under irregular sampling. The Neural Graphical Modeling framework (Bellot et al., 2022) introduces a continuous-time formulation using Neural ODEs with local independence constraints. While it provides consistency guarantees and avoids interpolation, it focuses on infinitesimal-time dependencies and assumes a continuous underlying process. In contrast, our method models lagged causal interactions directly at the discrete level, providing interpretable structures and finite-sample guarantees. The entropy-based method (Assaad et al., 2022a) infers summary causal graphs without interpolation, which provides robustness to sparse sampling. However, it lacks resolution at the lag level and does not model temporal causality explicitly. Our framework bridges this gap by supporting irregular timestamps while recovering lag-specific, nonlinear causal dependencies.

B. Notations Used Throughout This Paper

To improve clarity and readability, we summarize the key mathematical symbols used throughout the paper in Table 4.

C. Linear Causal Discovery as a Special Case

In domains where linear assumptions are valid or preferred for interpretability, ReTimeCausal can recover causal graphs using sparse linear regression. Each variable is regressed on its candidate parents:

$$\tilde{\mathbf{X}}^t = \sum_{\tau=1}^{L} \tilde{\mathbf{X}}^{t-\tau} \mathbf{W}^{(\tau)} + \boldsymbol{\epsilon}^t, \tag{11}$$

Here, $\mathbf{W}^{(\tau)} \in \mathbb{R}^{d\times d}$ denotes the weighted adjacency matrix that captures the causal influence from variables at lag τ, where the entry $\mathbf{W}_{ij}^{(\tau)}$ quantifies the causal effect of variable $X_j^{t-\tau}$ on X_i^t. The noise vector $\boldsymbol{\epsilon}^t \in \mathbb{R}^d$ consists of mutually independent components, each assumed to be independent of the predictors.

Unlike the nonlinear case, where kernel methods are required to model complex dependencies in a high-dimensional space, linear relationships are directly represented in the original feature space. This enables efficient recovery of the causal graph using sparse linear regression techniques.

A common approach is to formulate causal discovery as a sparse optimization problem by minimizing the reconstruction

Table 4. List of Notations used throughout the paper.

Symbol	Definition
$\mathbf{X}_o, \mathbf{X}_m$	Observed and missing subsets of the time series $\mathcal{D}$
$\mathbf{X}^t$	Multivariate observation at time step t
$\tilde{\mathbf{X}}^t$	Completed observation formed by observed values and imputed values
$\mathcal{X}_i$	The i-th variable (node) in the summary causal graph
$\mathcal{G}_S$	Summary causal graph over $\{\mathcal{X}_1, \ldots, \mathcal{X}_d\}$
L	Maximum lag order considered in temporal dependencies
T	Length of the time series
$f_i(\cdot)$	Structural function mapping from causal parents $\mathbf{Pa}_i$ to variable X_i
ϵ_t	Noise term
$\mathbf{W}^{(\tau)}$	Weight matrix capturing causal influence at lag τ
$\mathbf{W}^{(\tau)}_{\text{high}}$	High-dimensional kernel weight matrix in RKHS
Φ	RFF-induced back-projection matrix from kernel feature space to input space, with $\Phi \in \mathbb{R}^{d\times p}$
d	Number of variables in the causal graph
K	Number of EM iterations
k	Current EM iteration index
α	Exponential smoothing coefficient for weight averaging
γ, β	Thresholds for graph binarization and pruning significance
p	Number of random Fourier features (RFF dimension)
n	Number of independent multivariate time-series samples

error with an ℓ_1-penalty to enforce sparsity in the estimated graph:

$$\min_{\{\mathbf{W}^{(\tau)}\}_{\tau=1}^{L}} \sum_{t=1}^{T} \left\| \tilde{\mathbf{X}}^t - \sum_{\tau=1}^{L} \tilde{\mathbf{X}}^{t-\tau} \mathbf{W}^{(\tau)} \right\|_2^2 + \lambda \sum_{\tau=1}^{L} \|\mathbf{W}^{(\tau)}\|_1. \tag{12}$$

Where $\mathbf{W}^{(\tau)} \in \mathbb{R}^{d\times d}$ is the weighted adjacency matrix representing causal coefficients, and $\lambda > 0$ controls the trade-off between fit and sparsity.

This formulation is equivalent to solving d separate Lasso regression problems, one for each variable conditioned on the others, and is computationally efficient in high-dimensional settings. Moreover, under standard assumptions (e.g., causal sufficiency, acyclicity, and additive noise), such methods can identify the correct causal skeleton.

Nevertheless, when the true system exhibits nonlinear or interaction effects, purely linear models may yield biased or incomplete causal interpretations. Thus, linear models are best suited either as a baseline or in domains where linearity is a reasonable approximation.

D. Graph Pruning

The thresholding step in Section 4.2 produces an initial binary causal graph $\mathcal{G}_S$ based on regression weights. However, in high-dimensional or noisy settings, this graph may still contain spurious or unstable edges—particularly when the sample size is limited or the causal signal is weak. To address this, we introduce a model-specific pruning module tailored to the functional nature of the underlying data—linear or nonlinear. This post-processing step employs a unified pruning threshold β to filter out statistically insignificant or weakly contributing edges, thereby refining the causal graph and improving its interpretability, stability, and generalization.

Nonlinear setting. For nonlinear data modeled via kernel regression, we adopt a pruning approach based on Causal Additive Models (CAM) (Bühlmann et al., 2014). This method fits generalized additive models (GAMs) to each target

Algorithm 1 ReTimeCausal: Learning Causal Structures from Irregular Time Series

Input: Incomplete time series $\mathcal{D} = \{(\mathbf{X}^t, \mathbf{M}^t)\}_{t=1}^T$ with observation masks; initial regression weights $\mathbf{W}$ and neural estimators $\{f_i^{(0)}\}_{i=1}^d$
Output: Estimated causal graph $\mathcal{G}_S^{(K)}$ and functions $\{f_i^{(K)}\}$
for $k = 1$ **to** K **do**
  # *E-step: Impute missing values using current estimators* $\{f_i^{(k-1)}\}$
  $\tilde{\mathbf{X}}^{(k)} \leftarrow \texttt{Impute}(\mathcal{D}, \{f_i^{(k-1)}\})$
  # *M-step: (1)Update neural estimators* $\{f_i^{(k)}\}$
  $\arg\min \mathcal{L}_{\text{imp}} = \frac{1}{\|\mathbf{M}\|_1} \sum_{t,i} M_i^t \left\| X_i^t - f_i(\mathbf{Pa}_i^t) \right\|_2^2$
  # *M-step: (2)Update graph weights via kernelized sparse regression*
  $\arg\min \mathcal{L}_{\text{reg}} = \sum_t \left\| \tilde{\mathbf{X}}^t - \sum_\tau \phi(\tilde{\mathbf{X}}^{t-\tau}) \mathbf{W}_{\text{high}}^{(\tau)} \right\|^2 + \lambda \sum_\tau \|\mathbf{W}_{\text{high}}^{(\tau)}\|_1$
  for $\tau = 1$ **to** L **do**
    $\hat{\mathbf{W}}^{(k,\tau)} = \Phi \cdot \mathbf{W}_{\text{high}}^{(\tau)}$
  end for
  # *Weight smoothing*
  $\mathbf{W}^{(k)} \leftarrow \alpha\, \mathbf{W}^{(k-1)} + (1-\alpha)\, \hat{\mathbf{W}}^{(k)}$
  # *Graph extraction*
  $\hat{\mathcal{G}}_S^{(k)} \leftarrow \texttt{Binarization}(\mathbf{W}^{(k)}, \gamma)$
  # *Prune spurious edges*
  $\mathcal{G}_S^{(k)} \leftarrow \mathrm{CAM}(\mathcal{D}, \hat{\mathcal{G}}_S^{(k)}, \beta)$ **or** $\mathrm{LinReg}(\mathcal{D}, \hat{\mathcal{G}}_S^{(k)}, \beta)$
end for
return $\mathcal{G}_S^{(K)}, \{f_i^{(K)}\}$

variable and tests the significance of each input variable (including lagged parents), removing those with p-values exceeding a predefined threshold β. Although CAM is originally designed for static data, we apply it independently to each time-lagged parent set, leveraging its ability to capture smooth nonlinear influences while enforcing structural sparsity.

Linear setting. For linearly modeled systems, pruning is performed by applying a coefficient-based threshold to the absolute values of the estimated regression weights. Specifically, an edge $X_j^{t-\tau} \to X_i^t$ is retained only if $|W_{i,j}^{(\tau)}| > \beta$. This simple yet effective criterion eliminates weak dependencies that are likely to result from noise or overfitting, and is consistent with sparsity-inducing strategies used in prior DAG learning work (Zheng et al., 2018).

These pruning mechanisms act as the final refinement step in the ReTimeCausal pipeline, suppressing false positives and clarifying structural dependencies. The resulting pruned graph $\mathcal{G}_S$ is more parsimonious, statistically grounded, and better suited for interpretation and downstream applications. By separating pruning from initial graph construction, we enable modular design and allow for the incorporation of task- or domain-specific pruning strategies.

E. Motivation and Algorithmic Framework

The design of ReTimeCausal is motivated by the limitations of impute-then-discover pipelines under irregular and incomplete time series. Traditional methods either assume complete observations or perform imputation independently of the causal structure, which can introduce spurious dependencies or distort temporal dynamics. To address this, ReTimeCausal performs causal discovery and missing data imputation in a joint, iterative manner.

Specifically, ReTimeCausal integrates the Additive Noise Model (ANM) into an EM-style optimization framework. At each iteration, it imputes missing values based on current causal estimates (E-step) and updates the structure through sparse regression and pruning (M-step). This integration ensures that imputation respects the evolving causal graph, and that structure learning adapts to data refinement over time.

The algorithm supports both linear and nonlinear causal mechanisms, accommodates arbitrary sampling patterns, and yields interpretable, lag-specific graphs. To improve convergence robustness, we introduce exponential weight smoothing and apply post-hoc pruning based on causal independence principles. Algorithm 1 presents the pseudocode of ReTimeCausal.

Under mild conditions (bounded noise, identifiable mechanisms, compact domain), the learned functions f_i converge in ℓ_1 norm, and the estimated graph $\mathcal{G}_S^{(K)}$ converges in probability to the true summary graph $\mathcal{G}_S^*$.

F. Further Discussion on the Assumptions Used in This Paper

In this section, we provide further discussion on the assumptions made in our framework and clarify their roles in both modeling and theoretical analysis. While these assumptions are standard in the literature on time series causal discovery, understanding their necessity and scope of applicability is important for both the theoretical soundness and practical flexibility of our method.

F.1. Finite-Order Markov Property and Temporal Modeling

Assumption 3.3 (Finite-Order Markov Property) states that each variable at time t depends only on a finite number of past variables. This assumption is fundamental for representing temporal dependencies through sparse regression over a fixed lag window, which underpins the structure learning component of our method. Without such an assumption, modeling would require potentially infinite memory, making both estimation and interpretation infeasible. In practice, temporal dependence typically concentrates on recent history, making low-order Markov models a reasonable approximation. While our method is formulated using a fixed finite order, it is flexible enough to handle higher-order dependencies by increasing the lag window, which is straightforward to implement and evaluate in practice.

F.2. Temporal Consistency and the Validity of Summary Graphs

Assumption 3.4 (Consistency Throughout Time) requires that the causal directions summarized by $\mathcal{G}_S$ do not change across time. This assumption is essential for defining a well-posed summary causal graph: if causal directions were to drift over time, then aggregating them into a single static structure would be ill-defined or misleading. Importantly, this assumption concerns time-invariant causal directionality in the summary graph, not full distributional stationarity of the entire time series. In many applications, such as analyzing causal relations in physiological systems (e.g., EEG data) or economic indicators, this assumption holds reasonably well over moderate time spans. However, real-world systems may still exhibit nonstationarities or time-varying effects. In such cases, while our method can still be applied heuristically (e.g., in sliding windows), the theoretical guarantees derived under this assumption may no longer hold, and further investigation into time-varying extensions would be necessary.

F.3. Causal Sufficiency and the Impact of Latent Confounders

Assumption 3.5 (Causal Sufficiency) postulates that all common causes of observed variables are themselves observed. This is a standard assumption in many causal discovery frameworks, as it ensures that the observed statistical dependencies are not confounded by hidden variables. In our work, this assumption is particularly important for the consistency result in Proposition 1, which relies on the observed data faithfully representing the full causal structure. In practice, however, the presence of latent confounders is often unavoidable, especially in domains such as biology or social sciences. While our method does not explicitly account for hidden variables, it can be extended with techniques from the literature on causal discovery with latent confounding (e.g., FCI-based methods), which we leave as future work.

F.4. Faithfulness Assumption in Sparse Regression-Based Causal Discovery

Assumption 3.6 (Faithfulness) ensures that the statistical dependencies observed in the data correspond exactly to the causal structure, meaning that no causal relationships are masked by precise cancellation of effects. This assumption is necessary for identifying the true causal graph using statistical methods such as sparse regression, which relies on inferring structure from conditional independence patterns. While faithfulness is a strong condition, it is widely adopted in causal inference, and violations typically occur only under specific and rare parameterizations. In applied scenarios, approximate faithfulness is often sufficient to yield useful and interpretable causal graphs, especially when combined with regularization strategies that promote robustness to sampling noise.

F.5. Non-Gaussianity in Additive Noise Models and Identifiability

Finally, regarding the non-Gaussianity of noise, we emphasize that this is not a global requirement of our method. It is only invoked in the identifiability analysis of nonlinear additive noise models (ANMs), where classical results show that causal direction can be determined when the noise is non-Gaussian and independent of the cause. This assumption enables stronger theoretical guarantees in Section 4.2, particularly when learning causal direction from observational data. However, in practical applications, our method remains applicable even in Gaussian noise settings; in such cases, identifiability may be weaker, but the structure learning and estimation methods still produce meaningful and informative causal summaries. In real-world data—such as sensor networks, economic systems, or biological processes—non-Gaussian noise is often present due to heterogeneity, measurement error, or nonlinear dynamics, making this assumption not only theoretically useful but also practically reasonable in many settings.

G. Convergence Analysis

The core of ReTimeCausal lies in its iterative EM style optimization, where missing data is imputed based on current causal estimates, and structural functions are subsequently updated via sparse regression. Understanding the convergence behavior of this process is essential to ensure that the learned causal graph stabilizes and aligns with the underlying true structure as data grows.

This iterative learning process can be understood as a two-level optimization scheme: (i) an outer loop that alternates between imputing missing data and updating the causal graph (EM iterations), and (ii) an inner loop that solves kernel-based sparse regression problems to refine the structural functions. We analyze the convergence of both components under standard assumptions.

At each iteration, the imputed data $\hat{\mathbf{X}}_m^{(k)}$ is mapped into a kernel feature space, enabling the estimation of complex nonlinear structural functions via convex optimization. This forms a two-level learning process: (i)The outer loop (EM) alternates between imputing data and updating structure, and (ii)the inner loop (kernel regression) refines structural functions in a kernel-embedded space.

Under standard assumptions—such as bounded kernels, compact input domains, and noise terms that are independent of the parent variables—the learned structural functions converge to the true f_i in the L^2 norm. Specifically:

$$\forall \epsilon > 0,\ \exists\ f_i^{(\text{kernel})} \in \mathcal{H}_\kappa \quad \text{s.t.} \quad \|f_i - f_i^{(\text{kernel})}\|_{L^2} < \epsilon$$

Furthermore, sparse kernel regression ensures that the estimated graph $\hat{\mathcal{G}}_S$ converges in probability to the true summary graph $\mathcal{G}_S^*$ as $n \to \infty$, assuming faithfulness (Rosasco et al., 2004; Spirtes et al., 2001).

G.1. Proof of Proposition 4.1

To establish Proposition 4.1, we outline the proof of structural consistency by decomposing the argument into four key components. We assume throughout that Assumptions 3.2 (Ignorable Missingness), 3.3 (Finite-Order Markov Property), 3.4 (Consistency Throughout Time), 3.5 (Causal Sufficiency), and 3.6 (Faithfulness) hold.

Step 1 Under Assumption 3.2 and model identifiability, the maximum likelihood estimator based on the observed data is asymptotically consistent. In addition, provided that the EM procedure is initialized within sufficiently close to the global MLE and the likelihood satisfies standard regularity conditions, the classical EM iterations converge to this MLE (Dempster et al., 1977; Wu, 1983), we have

$$\hat{\theta}^{(k)} \xrightarrow{p} \theta^* \quad \text{and} \quad \hat{\mathbf{X}}_m^{(k)} \xrightarrow{p} \mathbf{X}_m^*,$$

where θ^* is the true parameter and $\mathbf{X}_m^* := \mathbb{E}[\mathbf{X}_m \mid \mathbf{X}_o; \theta^*]$ denotes the oracle E-step imputation target.

Step 2 For each structural equation $X_i^t = f_i(\mathbf{Pa}_i^t) + \epsilon_i^t$, assume f_i belongs to a function class $\mathcal{F}$ with universal approximation capacity, such as Reproducing Kernel Hilbert Spaces (RKHS) or sufficiently wide neural networks. Then, by standard learning theory (e.g., (Schölkopf & Smola, 2002; Györfi et al., 2002)), for any $\varepsilon > 0$, there exists an estimator $f_i^{(k)} \in \mathcal{F}$ such that

$$\|f_i^{(k)} - f_i^*\|_{L^2} < \varepsilon \quad \text{as } n \to \infty.$$

Hence the learned structural functions converge to the ground truth, provided the imputations are accurate (Step 1).

Step 3 Let $\mathcal{G}_S^*$ be the true causal graph and $\hat{\mathcal{G}}_S^{(k)}$ the estimate at iteration k. Assume (i) the structural functions admit identifiable ANMs, (ii) regularity of the design (e.g., the irrepresentability condition (Wainwright, 2009)), and (iii) appropriate regularization (e.g., $\lambda_n \to 0$ with $\sqrt{n}\lambda_n \to \infty$). Then sparse regression on $\hat{\mathbf{X}}^{(k)}$ achieves support recovery:

$$\Pr\Big(\mathrm{supp}(\hat{\mathbf{W}}^{(k)}) = \mathrm{supp}(\mathbf{W}^*)\Big) \to 1 \quad \text{as } n \to \infty.$$

Under Assumption 3.6 (faithfulness), the recovered supports correspond to the true causal edges.

Step 4 Building on the support recovery guarantee from Step 3, we now analyze the impact of exponential smoothing and entrywise thresholding on graph stability. In addition to support consistency, standard results from high-dimensional sparse estimation theory (Wainwright, 2009) imply that, under the same regularity assumptions, the estimation error in ℓ_1 norm satisfies

$$\|\hat{\mathbf{W}}^{(k)} - \mathbf{W}^*\|_1 = o(\gamma_n) \quad \text{as } n \to \infty,$$

where γ_n is the thresholding parameter. This result provides a foundation for analyzing the behavior of the smoothed weights $\mathbf{W}^{(k)}$, which are computed via

$$\mathbf{W}^{(k)} = \alpha\, \mathbf{W}^{(k-1)} + (1-\alpha)\, \hat{\mathbf{W}}^{(k)}.$$

subtract $\mathbf{W}^*$ from both sides and apply the triangle inequality, we have

$$\big\|\mathbf{W}^{(k)} - \mathbf{W}^*\big\|_1 \;\le\; \alpha \big\|\mathbf{W}^{(k-1)} - \mathbf{W}^*\big\|_1 \;+\; (1-\alpha)\big\|\hat{\mathbf{W}}^{(k)} - \mathbf{W}^*\big\|_1.$$

Unrolling the recursion yields, for any $k \ge 1$,

$$\big\|\mathbf{W}^{(k)} - \mathbf{W}^*\big\|_1 \;\le\; \alpha^k \big\|\mathbf{W}^{(0)} - \mathbf{W}^*\big\|_1 \;+\; (1-\alpha)\sum_{j=1}^{k} \alpha^{k-j} \big\|\hat{\mathbf{W}}^{(j)} - \mathbf{W}^*\big\|_1.$$

Since $\|\hat{\mathbf{W}}^{(j)} - \mathbf{W}^*\|_1 = o(\gamma_n)$ (uniformly over fixed $j \le k$) and $\alpha \in [0,1)$, it follows that $\|\mathbf{W}^{(k)} - \mathbf{W}^*\|_1 \le \alpha^k \big\|\mathbf{W}^{(0)} - \mathbf{W}^*\big\|_1 + O(\gamma_n)$ as $n \to \infty$ for each fixed k, and in fact $\mathbf{W}^{(k)} \to \mathbf{W}^*$ geometrically in k up to the vanishing estimation error.

To ensure correct support recovery under thresholding, we assume a positive margin parameter $\eta > 0$ such that the threshold satisfies $\gamma \in (\gamma_0, \gamma_{\min})$ and

$$\eta := \min\{\gamma_{\min} - \gamma,\ \gamma - \gamma_0\} > 0.$$

This margin separates true and false edges by at least η, enabling thresholding at level γ to recover the correct support with high probability.

With a symmetric margin $m > 0$ and a threshold $\gamma_n \in (\gamma_0, \gamma_{\min})$, this implies support stabilization for large k and n:

$$\lim_{k\to\infty} \mathrm{supp}(\mathbf{W}^{(k)}) = \mathrm{supp}(\mathbf{W}^*) \quad \text{and} \quad \lim_{k\to\infty}\lim_{n\to\infty} \Pr\Big(\hat{\mathcal{G}}_S^{(k)} = \mathcal{G}_S^*\Big) = 1.$$

G.2. Convergence in Realistic Settings

This section analyzes the error convergence of ReTimeCausal under realistic constraints—namely, finite samples and a finite number of EM iterations. In contrast to the idealized setting considered in Appendix G.1, we focus on three primary sources of error that arise in practice:(i) statistical estimation error due to limited sample size, (ii) kernel approximation error introduced by finite-dimensional feature mappings, and (iii) optimization residual caused by incomplete EM convergence.

Building on the assumptions established in the main text, we additionally introduce two mild conditions commonly used in high-dimensional EM analysis: (i) geometric absolute regularity to ensure concentration bounds for dependent observations (Merlevède et al., 2009), and (ii) gradient stability and restricted strong concavity (RSC) to guarantee local contractivity of the EM operator (Balakrishnan et al., 2017).

Under this framework, we formally characterize the evolution of the imputation error e_k at each EM iteration. By leveraging the sparsity structure induced by feature-to-input projection, we derive a quantitative bound on the Structural Hamming Distance (SHD), expressed in terms of sample size, feature dimension, and the number of EM iterations.

Theorem G.1. *Suppose the oracle EM operator satisfies local contractivity with rate $\rho < 1$ under mild curvature and stability assumptions. Let λ be a regularization level proportional to $\sqrt{\log(\widetilde{d})/n}$, and let τ_n denote the finite-sample deviation term. Then there exist constants $C_1 > 0$ such that*

$$\mathrm{SHD}\left(\hat{\mathcal{G}}_S^{(k)}, \mathcal{G}_S^*\right) \leq \frac{C_1}{\eta}\left(s\sqrt{\frac{\log(\widetilde{d})}{n}} + \delta_p + \rho^k e_0 + \frac{\tau_n}{1-\rho}\right).$$

Before proceeding with the proof, we recall the key assumptions and notation:

- Let $\hat{\mathbf{X}}^{(k)}$ denote the imputed data after the k-th EM iteration, and define the imputation error $e_k := \|\hat{\mathbf{X}}_m^{(k)} - \mathbf{X}_m^*\|_\infty$, where $\mathbf{X}_m^*$ denotes the oracle E-step fixed point.
- Let $\widetilde{d} := d^2 L$ denote the total number of pairwise temporal interactions across d variables with maximum lag L.
- The regularization level is set as:

  $$\lambda \asymp \sigma\sqrt{\frac{\log(\widetilde{d})}{n}}, \quad \text{with} \quad n \gtrsim s\log(\widetilde{d}),$$

  where s is the sparsity level of the true coefficient matrix $\mathbf{W}^*$ and σ^2 is the sub-Gaussian proxy variance (Bickel et al., 2009; Negahban et al., 2012).
- Feature map $\phi : \mathbb{R}^d \to \mathbb{R}^p$ is L_ϕ-Lipschitz and uniformly approximates a bounded kernel with error $\delta_p \to 0$ as $p \to \infty$ (Rahimi & Recht, 2007; Rudi & Rosasco, 2017; Steinwart & Christmann, 2008).
- The projection from feature space to input space is given by the back-projection matrix $\Phi \in \mathbb{R}^{d\times p}$, with $\|\Phi\|_{1\to 1} \leq C_\Phi$.
- The margin condition assumes that true edges have weight at least $\gamma_{\min}$ and false edges at most γ_0, with threshold $\gamma \in (\gamma_0, \gamma_{\min})$ and

  $$\eta := \min\{\gamma_{\min} - \gamma,\ \gamma - \gamma_0\} > 0.$$

  This condition is natural in the ground truth graph, where spurious edge weights are strictly smaller than the weakest true edge.

Step 1 We first establish an upper bound on the ℓ_1-norm error between the estimated coefficients $\hat{\mathbf{W}}^{(k)}$ and the true sparse tensor $\mathbf{W}^*$ after k EM iterations. By applying standard techniques for high-dimensional regression under absolute regularity dependence (e.g., primal–dual witness arguments and RE conditions) (Bickel et al., 2009; Basu & Michailidis, 2015), and using the stability of the RE constant under bounded imputation error e_k, we obtain:

$$\left\|\hat{\mathbf{W}}^{(k)} - \mathbf{W}^*\right\|_1 \leq \frac{4s\lambda}{\phi_0^2} + c_0 e_k + c_0\delta_p.$$

Here, c_0 is a constant depending only on smoothness and projection parameters (e.g., $L_f, L_\phi, \phi_0, C_\Phi$), and δ_p quantifies the kernel approximation error from the finite-dimensional feature map (Rahimi & Recht, 2007; Rudi & Rosasco, 2017).

Step 2 Next, we analyze the evolution of the imputation error e_k over EM iterations. Using the local contractivity of the oracle EM operator (under curvature and gradient stability assumptions) (Balakrishnan et al., 2017), combined with a high-probability uniform deviation bound between the sample and oracle operators (controlled via τ_n) (Merlevède et al., 2009), we obtain the recursion:

$$e_{k+1} \leq \rho\, e_k + \tau_n \quad \Rightarrow \quad e_k \leq \rho^k e_0 + \frac{\tau_n}{1-\rho},$$

where $\rho \in (\rho_{\mathrm{pop}}, 1)$ and $\tau_n = O\left(\sqrt{\frac{\log(\widetilde{d})}{n}}\right)$ captures the statistical fluctuation due to absolute regularity samples.

Step 3 Finally, we connect the coefficient estimation error to the Structural Hamming Distance (SHD) via the thresholding margin condition. Suppose all true edges have weights at least $\gamma_{\min}$, and all spurious edges have magnitudes at most γ_0. Then, thresholding at level $\gamma \in (\gamma_0, \gamma_{\min})$ correctly separates true and false edges, provided that the entrywise estimation error satisfies

$$\max_{i,j} |\hat{W}_{ij}^{(k)} - W_{ij}^*| < \eta.$$

This condition reflects a standard minimal signal strength assumption (also referred to as a margin condition), which is commonly used to ensure support recovery consistency under thresholding (Zhang & Zhang, 2014; Bühlmann & Van De Geer, 2011). Consequently, the Structural Hamming Distance (SHD) is upper bounded by the total ℓ_1 error divided by η, as each unit of SHD corresponds to at least η of deviation in the coefficient matrix:

$$\text{SHD}\left(\hat{\mathcal{G}}_S^{(k)}, \mathcal{G}_S^*\right) \leq \frac{1}{\eta} \left\| \hat{\mathbf{W}}^{(k)} - \mathbf{W}^* \right\|_1.$$

Step 4 Combining the results from Steps 1–3 and substituting the explicit form of λ and τ_n, we obtain the desired upper bound on the structural Hamming distance:

$$\text{SHD}\left(\hat{\mathcal{G}}_S^{(k)}, \mathcal{G}_S^*\right) \leq \frac{C_1}{\eta} \left(s\sqrt{\frac{\log(\widetilde{d})}{n}} + \delta_p + \rho^k e_0 + \frac{\tau_n}{1-\rho} \right),$$

which corresponds to the form presented in Theorem G.1.

G.2.1. Practical Limitations and Convergence Behavior

While the theoretical analysis establishes the asymptotic consistency of ReTimeCausal under standard assumptions, it is important to examine how these guarantees translate into practical settings with finite data, model approximations, and optimization heuristics. This section discusses several implementation-level factors that may influence convergence behavior in real-world applications.

First, although the EM-style method provides a principled framework for joint imputation and structure learning, its convergence rate can vary significantly depending on data characteristics such as the missingness rate (r), the degree of nonlinearity in causal functions $f_i(\cdot)$, and temporal sparsity. In our formulation, the E-step computes the imputed values based on current estimators.

In practice, we match this distribution to the residuals observed during function fitting (e.g., Laplace or empirical noise), ensuring compatibility with a wide range of noise models. High missingness rates ($r > 0.6$) or heavy-tailed noise can slow down EM convergence and may require additional iterations for structure stabilization.

Second, in the M-step, each causal function f_i is estimated using sparse regression in a kernel-induced space. To promote both convergence and interpretability, we apply ℓ_1-penalized regression over lagged features, as previously introduced in Eq. 6. This formulation encourages sparsity across temporal lags while enabling the recovery of nonlinear dynamics via the kernel mapping $\phi(\cdot)$. However, since the objective becomes non-convex when neural or RBF kernels are used, convergence to a global minimum is not guaranteed. We mitigate this through two stabilizing techniques:

- **Exponential smoothing**: We maintain a smoothed estimate of regression weights across iterations: $\mathbf{W}^{(k)} = \alpha \mathbf{W}^{(k-1)} + (1-\alpha)\hat{\mathbf{W}}^{(k)}$;
- **Threshold-based pruning**: Weak edges are removed using a lag-specific cutoff $\gamma > 0$, reducing variance in graph estimates and improving interpretability.

Third, although the method relies on alternating updates, it does not involve adversarial training or end-to-end neural inference. This enhances empirical robustness and enables direct control over convergence via interpretable hyperparameters such as λ, α, and γ.

Although the convergence speed and quality of ReTimeCausal may be affected by factors such as initialization and model complexity, we observe stable optimization behavior across a wide range of settings. Its modular architecture facilitates integration with more efficient optimization strategies or domain-specific priors, enhancing its adaptability to diverse real-world applications.

H. Discussion of Compared Baselines

To evaluate the performance of ReTimeCausal, we compare it against several widely used causal discovery algorithms that represent mainstream methodological paradigms. Specifically, we consider the following baselines:

PCMCI (Runge et al., 2019b) is a constraint-based method tailored for high-dimensional time series. It combines conditional independence testing with a temporal extension of the PC algorithm, providing statistical control under assumptions of causal sufficiency and time-lagged dependence.

DYNOTEARS (Pamfil et al., 2020) is a score-based approach that extends NOTEARS to dynamic settings. It optimizes a smooth objective with acyclicity constraints and sparse regularization, and estimates dynamic causal graphs with contemporaneous and lagged edges. Since the original DYNOTEARS is designed for linear dynamics, we include a kernelized variant denoted as DYNOTEARS* for comparison in nonlinear settings. DYNOTEARS* replaces the linear regression module with our kernelized sparse regression method, allowing it to model nonlinear dependencies and serve as a more appropriate baseline for evaluating ReTimeCausal in nonlinear regimes.

CUTS+ (Cheng et al., 2024a) is a recent iterative framework designed for irregular time series. It leverages uncertainty-aware structure learning, message-passing neural networks, and cycle-consistent data augmentation to enhance robustness and scalability.

Rhino (Gong et al., 2023) is a deep causal discovery method for time series. It models nonlinear relations, instantaneous effects, lagged effects, and history-dependent noise. It gives a flexible way to learn temporal causal graphs under complex real-world dynamics. Rhino serves as a strong baseline when the data have nonlinear or fast-changing patterns.

AERCA (Han et al., 2025) is a root-cause analysis method that also contains a built-in Granger causal discovery module. It models exogenous variables and their normal patterns. It then uses deviations in these variables to identify possible root causes. Because AERCA learns a Granger-style causal structure as part of its process, it provides an additional reference point for evaluating causal discovery in our setting.

Most of these methods do not handle missing data and require fully observed inputs. To ensure fair comparison, we apply **TimeMixer** (Wang et al., 2024) as a preprocessing step. TimeMixer decomposes a time series into multiple resolutions and uses learnable mixing operations. It captures short- and long-term patterns and gives high-quality imputations for downstream causal discovery.

Evaluation Metrics. To quantitatively compare performance, we report three standard metrics commonly used in causal discovery:

- **SHD (Structural Hamming Distance)**: Measures the number of edge additions, deletions, or reversals needed to transform the estimated causal graph into the ground truth. Lower values indicate better structural accuracy.
- **SID (Structural Intervention Distance)**: Captures the number of incorrect causal inferences under intervention. It reflects the mismatch between predicted and true interventional distributions, with lower values indicating better causal robustness.
- **F1 Score**: The F1 score represents the harmonic mean of precision and recall, evaluated on the predicted causal edges. It quantifies the balance between sensitivity and specificity, with higher values indicating more accurate and reliable structure recovery.

I. Additional Experiments and Ablation Studies

To further evaluate the robustness, effectiveness, and design choices of the ReTimeCausal framework, we conduct a series of additional experiments and ablation studies. While the main results in Section 5 demonstrate strong overall performance across various settings, it remains important to understand how individual components of the model contribute to its success, how it behaves under different nonlinear mechanisms, and how it scales with increasing complexity. This section presents targeted analyses that isolate specific factors such as noise injection, nonlinearity types, temporal lag orders, and real-world generalization, thereby providing a deeper empirical characterization of the framework's behavior.

I.1. Effect of Noise Injection

To assess the role of noise injection in ReTimeCausal, we compare the full model with a variant that disables this component. As shown in Table 5, removing noise injection results in a notable drop in F1 score (from 0.976 to 0.159) and a significant increase in SHD (from 1 to 212). These results show that noise injection is essential for accurate structure recovery.

Table 5. Robustness Comparison under Functionally Distinct Nonlinear Settings

Methods	SIN-laplace-20-20-1			TANH-laplace-20-20-1		
	SHD	**SID**	**F1**	**SHD**	**SID**	**F1**
PCMCI+TimeMixer	55	12	0.321	25	34	0.194
PCMCI+EM	22	35	0.154	21	33	0.276
DYNOTEARS*+TimeMixer	78	21	0.204	36	40	0.100
Rhino+TimeMixer	51	29	0.164	77	29	0.115
AERCA+TimeMixer	54	34	0.100	176	15	0.120
CUTS+	197	11	0.116	66	25	0.154
ReTimeCausal w/o Noise	212	**0**	0.159	185	**0**	0.178
ReTimeCausal	**1**	**0**	**0.976**	**1**	1	**0.974**

From a theoretical perspective, noise injection serves as a form of regularization in both optimization and statistical estimation. In the context of Expectation-Maximization, missing data can cause the algorithm to converge prematurely to suboptimal local minima, particularly when early-stage imputations are biased or overconfident. By perturbing the imputed residuals with small, zero-mean noise, we prevent the regression step from overfitting to inaccurate imputations, thereby promoting exploration and reducing the risk of early convergence to poor structures.

In kernel regression, this mechanism additionally mitigates overfitting in high-dimensional feature spaces, where exact interpolation of sparse or noisy points can lead to unstable estimates. Injecting noise smooths the learned functions in the RKHS, improving generalization.

Moreover, from a statistical testing viewpoint, our pruning method relies on testing independence between the regression residuals and potential parents. However, if residuals are deterministically derived from imperfect imputations, the independence assumption may be violated. Noise injection restores this independence in expectation, making the downstream conditional independence tests more statistically valid.

I.2. Adaptability to Nonlinear Complexity

To assess the adaptability of ReTimeCausal to increasing functional nonlinearity, we evaluate performance across synthetic datasets generated using distinct nonlinear structural functions, including $\sum \mathbf{Sin}(\mathbf{Pa}_i^t)$ and $\sum \mathbf{Tanh}(\mathbf{Pa}_i^t)$. All datasets share the same underlying causal graph and noise distribution. As shown in Table 5, ReTimeCausal maintains high accuracy across a wide spectrum of nonlinearities, demonstrating its expressive capacity and generalization.

Table 6. Causal Graph Recovery Performance Under Varying Maximum Lag Orders

Methods	$L = 2$			$L = 3$			$L = 4$		
	SHD	**SID**	**F1**	**SHD**	**SID**	**F1**	**SHD**	**SID**	**F1**
PCMCI+TimeMixer	19	6.0	0.230	25	23.0	0.144	32	16.0	0.130
PCMCI+EM	39	11.0	0.316	17	23.0	0.414	27	33.0	0.069
DYNOTEARS+TimeMixer	9	13.3	0.640	8	15.0	0.765	21	6.3	0.439
Rhino+TimeMixer	40	12.0	0.200	71	**3.0**	0.202	60	79	0.190
AERCA+TimeMixer	42	13.0	0.160	50	18.0	0.167	44	28.0	0.120
CUTS+	72	3.0	0.200	233	6.0	0.127	252	**5.0**	0.137
ReTimeCausal	**2**	**0.3**	**0.941**	**5**	10.3	**0.865**	**17**	6.3	**0.730**

I.3. Multi-order Lag Analysis

To evaluate the ability of ReTimeCausal to recover multi-order lagged causal structures, we conduct a targeted experiment on a synthetic linear time series dataset with 10 variables and 10 edges. The dataset includes 60% randomly missing values, simulating a challenging irregular setting.

We generate and test three variants of the dataset with different maximum lag orders: $L = 2$, $L = 3$, and $L = 4$. For each configuration, we run ReTimeCausal and evaluate the structure recovery performance using standard metrics such as SHD, SID, and F1 score. The results are summarized in Table 6. Additionally, Figure 3 visualizes the recovered second-order lagged causal graph, demonstrating that our regression-based structure learning accurately captures the underlying dependencies, while the pruning strategy effectively removes spurious edges, resulting in a clean and interpretable causal structure. All reported metrics are computed on the temporally unrolled graphs that merge all lag orders, which may lead to relatively large SHD values (e.g., exceeding 100) due to the expanded graph size. Since CUTS+ only outputs a single summary-level causal graph, we treat this graph as being shared across all lag orders during evaluation.

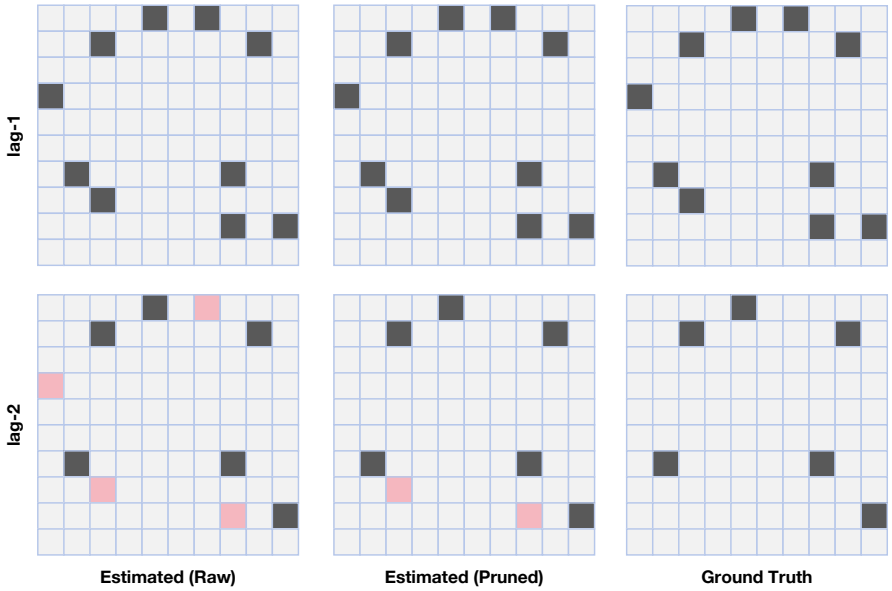

Figure 3. Visualization of recovered lagged causal graphs for the $L = 2$ setting. The top and bottom rows correspond to lag-1 and lag-2 structures, respectively. From left to right: estimated graph before pruning, pruned graph, and ground truth. Black squares denote correctly identified causal edges; red squares highlight spurious edges.

I.4. Evaluating Stability with Respect to R^2-Sortability

Recently, R^2-sortability has gained increasing attention as an artifact that can inadvertently boost the apparent performance of causal discovery methods (Reisach et al., 2023; Herman et al., 2025), thereby raising further concerns in time series causal discovery (Lohse & Wahl, 2025). The core idea is to use the coefficient of determination (R^2)—the proportion of variance in a variable explained by its direct causal parents—as a criterion for ordering variables along the causal direction. In systems where downstream variables tend to exhibit higher R^2 values (i.e., they can be better explained by their parents), the system is said to possess high R^2-sortability. Formally, R^2-sortability is defined as the average proportion of directed edges whose orientation is consistent with the ordering induced by R^2. Unlike varsortability, which is based on marginal variances, R^2-sortability emphasizes explanatory power, which provides a complementary perspective for characterizing data properties relevant to causal discovery.

In this section, we investigate how varying levels of R^2-sortability influence the stability and effectiveness of our method, and we demonstrate that its performance remains robust regardless of the degree of R^2-sortability.

We follow the synthetic data design principles proposed in the UUMC framework (Herman et al., 2025): (i) Unitless, (ii) Unrestricted, and (iii) Markov-Consistent. Based on the LR-gaussian-10-10-2 generation setting, we further adjust the distribution of noise standard deviations and the method for generating noise sortability to obtain a collection of synthetic

datasets whose R^2-sortability values lie within the range $[0.2, 0.8]$. The detailed settings and corresponding sortability values of the generated datasets are summarized as follows:

Table 7. Sortability under different noise settings in UUMC synthetic data

Level	Noise Std Range	Sorted Noise	Var-Sortability	R^2-Sortability
Low	[0.5, 2.0]	×	0.503	0.272
Mid	[1.0, 1.0]	×	0.003	0.503
High	[0.5, 2.0]	✓	0.012	0.794

Table 7 illustrates how varying noise levels and noise sorting strategies affect both var-sortability and R^2-sortability. To further investigate the impact of R^2-sortability on the performance of causal discovery methods, we evaluate standard metrics under different levels of R^2-sortability. The corresponding results are reported in Table 8.

Table 8. Evaluation metrics under different R^2-sortability levels

Level	SHD	SID	F1
Low	1	0	0.973
Mid	0	0	1.000
High	0	0	1.000

The evaluation results indicate that our method maintains robust performance across all three levels of sortability. While the performance under the low sortability condition is marginally reduced compared to the mid and high levels, it still achieves a high F1 score of 0.973. These results suggest that ReTimeCausal is relatively insensitive to changes in sortability, and its performance does not rely on favorable sortability structures.

I.5. Experiments on Additional Real-World Benchmarks

In addition to the CausalRivers evaluation, we further assess our method on two widely used benchmarks: **NetSim** and the recently released **CausalTime**. NetSim is a standard dataset for functional connectivity analysis in fMRI, providing biologically plausible synthetic time series with ground-truth networks, while accounting for hemodynamic variability and confounding effects. Following prior work, we introduce missing values through random masking with rates of 0.2 and 0.6 to emulate incomplete observations.

As shown in Table 9, ReTimeCausal achieves the best performance under both settings (F1 = 0.742 and 0.689), outperforming all competing methods. CUTS+ is competitive in the low-missing case but degrades under higher missingness, whereas AERCA, PCMCI-based, and DYNOTEARS-based baselines deteriorate more markedly. Rhino+TimeMixer shows the lowest F1 scores because Rhino depends on stable temporal patterns, while TimeMixer's coarse multiscale reconstruction introduces distortions that weaken Rhino's variational estimation.

Table 9. Performance comparison on the NetSim benchmark under different missing rates

Methods	Sim3 with 10 nodes 15 edges	
	MR=0.2	MR=0.6
PCMCI+TimeMixer	0.655	0.600
PCMCI+EM	0.405	0.408
DYNOTEARS*+TimeMixer	0.565	0.531
Rhino+TimeMixer	0.258	0.246
AERCA+TimeMixer	0.513	0.436
CUTS+	0.667	0.571
ReTimeCausal	**0.742**	**0.689**

Beyond NetSim, we extend our evaluation to the **CausalTime** benchmark (Cheng et al., 2024b), which contains multivariate

time series from ecological, financial, and medical domains. Compared to CausalRivers, CausalTime provides more diverse and realistic real-world conditions. In particular, we report results on datasets from the Medical and PM25 domains to examine performance in clinically relevant settings. As summarized in Table 10, ReTimeCausal achieves the best F1 scores in three of the four reported settings, namely Medical under MR = 0.6 and PM25 under both MR = 0.2 and 0.6, while CUTS+ achieves a higher F1 only for Medical under MR = 0.2. These results indicate that ReTimeCausal remains competitive across diverse real-world missing-data conditions, especially under higher missingness.

Table 10. Performance comparison on the CausalTime benchmark under different missing rates

Methods	**Medical**		**PM25**	
	MR=0.2	**MR=0.6**	**MR=0.2**	**MR=0.6**
PCMCI+TimeMixer	0.491	0.479	0.433	0.407
PCMCI+EM	0.160	0.207	0.508	0.264
DYNOTEARS*+TimeMixer	0.541	0.539	0.549	0.557
Rhino+TimeMixer	0.472	0.405	0.431	0.430
AERCA+TimeMixer	0.386	0.378	0.245	0.238
CUTS+	**0.667**	0.571	0.468	0.444
ReTimeCausal	0.658	**0.648**	**0.633**	**0.649**

I.6. Effect of Imputation–Discovery Interaction

Figure 4 illustrates the evolution of imputation quality and causal discovery accuracy across EM iterations in Linear setting (LR-gaussian-20-20-1). The two curves exhibit a clear mutual reinforcement pattern, which constitutes the core mechanism behind our joint EM updates.

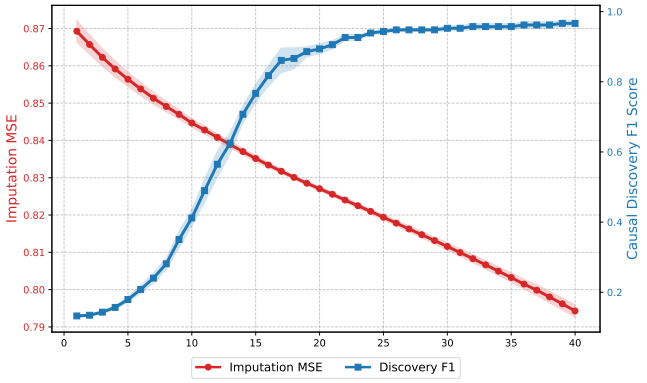

Figure 4. Co-evolution of imputation and causal discovery under the EM-style alternating process. Results are averaged over multiple runs on the same dataset. Shaded areas indicate the standard error.

This behavior can be explained as follows. Reducing the imputation error requires progressively updating the predicted values, whereas the estimated causal graph changes only when the edge set changes. After the graph stabilizes, subsequent iterations mainly improve predictive accuracy, which lowers the MSE while leaving the recovered structure unchanged.

I.7. Hyperparameter Sensitivity Analysis

We assess ReTimeCausal's resilience to its important hyperparameters, such as the smoothing coefficient α, pruning threshold β, and binarization threshold γ, in addition to the ablation studies mentioned above. We report performance on a representative nonlinear configuration (TANH–laplace–20–40–1) under 60% missingness, varying each hyperparameter's value over a wide range while holding all others constant.

ReTimeCausal is stable over a broad range of hyperparameter settings, as demonstrated by the results in Table 11. For values between 0.7 and 0.9, the smoothing coefficient α yields almost identical performance, suggesting that the exponential averaging mechanism is not sensitive to its precise selection. For the CAM pruning threshold β, a stricter setting such as

Table 11. Hyperparameter Sensitivity Analysis

Level	SHD	F1
$\alpha = 0.70$	4	0.952
$\alpha = 0.80$	5	0.938
$\alpha = 0.90$	3	0.964
$\beta = 0.01$	4	0.951
$\beta = 0.05$	5	0.938
$\beta = 0.10$	7	0.920
$\gamma = 0.05$	5	0.938
$\gamma = 0.15$	9	0.891
$\gamma = 0.25$	9	0.883

$\beta = 0.01$ provides the most reliable results; because CAM removes an edge only when its p-value exceeds β, smaller values enforce stronger significance requirements and lead to more effective pruning. In contrast, larger thresholds like $\beta = 0.10$ tend to retain statistically insignificant edges and introduce false positives. The binarization threshold γ also benefits from smaller values: thresholds around 0.05–0.15 preserve weak but genuine causal coefficients that CAM can further assess, whereas higher thresholds eliminate informative edges prematurely.

J. Complexity Analysis

We provide a consolidated analysis of the per-iteration computational complexity of the EM procedure. Let the batch size be B, sequence length T, number of variables d, lag order L, RFF feature dimension p, optimizer iterations I, and missing-rate ρ.

During the E-step, each missing entry, retrieving its lagged parent values costs $O(Ld)$, and evaluating the structural function $f_i(\cdot)$ costs $O(C_f)$. Since each f_i is a 3-layer MLP with constant hidden width, we have $C_f \approx O(Ld)$. With $M = \rho BTd$ missing values, the total E-step cost becomes $O(M(Ld + C_f)) = O(\rho BTLd^2) = O(BTLd^2)$.

During the M-step, updating all structural functions over the $V = (1-\rho)BTd$ observed samples requires $O(V(Ld+C_f)) = O(BTLd^2)$. Kernelized sparse regression, which includes constructing random Fourier features and performing sparse optimization, incurs a cost of $O(BTLdp) + O(ILBTdp) = O(ILBTdp)$. Nonlinear CAM pruning constructs, for each variable, a design matrix with $O(Ld)$ candidate parents and fits a spline-based GAM, yielding a per-variable cost $O(BT(Ld)^2)$; aggregated over all d variables, this gives a worst-case pruning complexity of $O(BTL^2d^3)$.

Combining all components, a single EM iteration incurs $O(BTLd^2) + O(ILBTdp) + O(BTL^2d^3)$. In practice, for small to moderate (d), the kernel regression term $O(ILBTdp)$ dominates; the theoretical worst-case term $O(BTL^2d^3)$ becomes significant only when the number of variables grows large.

K. Implementation Details

This section outlines the implementation details of the ReTimeCausal framework used in our experiments. We describe the model initialization, regression module configuration, kernel choices, imputation strategies, optimization routines, and hyperparameter settings. These settings are consistent across all experimental conditions unless otherwise noted. The goal is to facilitate reproducibility and clarify the technical setup behind the reported results.

For baseline algorithms, we primarily follow the hyperparameter settings recommended in the original papers or their official repositories (e.g., PCMCI[1], DYNOTEARS (Pamfil et al., 2020), Rhino[2], AERCA[3], TimeMixer[4], and CUTS+[5]). To ensure a fair comparison, we additionally perform hyperparameter tuning to identify the optimal settings for each baseline. The

[1]https://github.com/jakobrunge/tigramite
[2]https://github.com/microsoft/causica
[3]https://github.com/hanxiao0607/AERCA
[4]https://github.com/kwuking/TimeMixer
[5]https://github.com/jarrycyx/UNN

tuned hyperparameters are summarized in Table 12. All experiments were conducted on a single NVIDIA RTX 4090 GPU with 24GB memory.

Table 12. Hyperparameter Settings Used Across Main Datasets and Methods

Methods	Hyperparam	LR-10-10	TANH-10-10	TANH-20-20	SIN-20-20	CausalRivers	NetSim
PCMCI	PC_α	0.05	0.05	0.05	0.05	0.05	0.05
	CI Test	Parr	Parr	Parr	Parr	Parr	Parr
DYNOTEARS*	**W** lr	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}
	p	-	400	400	400	400	400
	σ^2	-	10	10	10	10	10
	ℓ_1	-	10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-4}
CUTS+	input_step	1	2	2	2	2	3
	n_groups	4	4	8	8	4	2
	model	multi_lstm	multi_lstm	multi_lstm	multi_lstm	multi_lstm	multi_lstm
	pred lr	$10^{-3} \to 10^{-4}$	$10^{-3} \to 10^{-4}$	$10^{-3} \to 10^{-4}$	$10^{-3} \to 10^{-4}$	$10^{-3} \to 10^{-4}$	$10^{-3} \to 10^{-4}$
	graph lr	$10^{-1} \to 10^{-2}$	$10^{-1} \to 10^{-2}$	$10^{-1} \to 10^{-2}$	$10^{-1} \to 10^{-2}$	$10^{-1} \to 10^{-2}$	$10^{-2} \to 10^{-3}$
	gru_layers	3	3	3	3	3	3
	lambda_s	$10^{-2} \to 10^{-3}$	$10^{-2} \to 10^{-3}$	$10^{-2} \to 10^{-3}$	$10^{-2} \to 10^{-3}$	$10^{-2} \to 10^{-3}$	$10^{-2} \to 10^{-3}$
	τ	1→0.1	1→0.1	1→0.1	1→0.1	1→0.1	1→0.1
Rhino	gumbel_temp	0.25	0.25	0.25	0.25	0.25	0.25
	lr	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}
AERCA	window_size	2	2	2	2	2	3
	β	0.5	0.5	0.5	0.5	0.5	0.5
	$\gamma_{encoder}$	0.1	0.15	0.15	0.15	0.15	0.15
	$\lambda_{encoder}$	0.01	0.2	0.2	0.2	0.2	0.2
	lr	10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-4}
	hidden layer	2	4	4	4	4	4
	layer size	50	100	100	100	100	100
ReTimeCausal	f_i lr	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}
	W lr	-	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}
	p	-	400	400	400	400	400
	σ^2	-	10	10	10	10	10
	α	0.7	0.7	0.7	0.7	0.7	0.7
	β	0.15	0.01	0.01	0.01	0.01	0.01
	γ	0.2	0.1	0.1	0.1	0.1	0.1
	λ	10^{-2}	10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-4}

L. Limitations

Despite its empirical performance and theoretical grounding, ReTimeCausal has several limitations that merit discussion. First, the method assumes that the missing data mechanism follows either the Missing Completely at Random (MCAR) or Missing At Random (MAR) assumptions. In practice, however, particularly in domains such as healthcare and finance, missingness is often Missing Not At Random (MNAR), where the probability of missing values depends on the unobserved data itself. ReTimeCausal does not explicitly model this setting, and applying it under MNAR may lead to biased estimates.

Second, the current framework only models lagged causal relationships and explicitly disallows instantaneous effects—i.e., causal dependencies occurring within the same time step. While this assumption is valid in many real-world applications where causal effects unfold over time, it limits applicability in high-frequency or coarsely sampled settings, where instantaneous interactions may exist.

From a computational perspective, while the M-step of our EM algorithm is efficient due to the use of Random Fourier Feature approximations, the E-step becomes increasingly expensive as the number of variables, time lags, and missing values grows. This bottleneck in the imputation step may hinder scalability to high-dimensional or long-horizon time series.

Another limitation lies in the assumption of time-invariant causal directions in the summary graph. ReTimeCausal assumes Consistency Throughout Time and fixed causal graphs, which precludes modeling dynamically evolving causal relationships. This should not be interpreted as requiring full distributional stationarity of the time series; rather, it restricts the method to settings where the summary-level causal directions remain stable over time. Extending the framework to genuinely time-varying causal graphs would be a valuable direction for future work.

Finally, the empirical evaluation is currently limited to synthetic datasets and a subset of real-world benchmarks with moderate size and complexity. While results are promising, further large-scale validation—especially in domains like climate modeling, electronic health records, or financial transaction systems—would strengthen the generalizability and practical impact of the proposed approach.