TIME SERIES CAUSAL DISCOVERY VIA CONTEXT-CONDITIONED AND CAUSALITY-AUGMENTED PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal discovery from time series is critical for many real-world applications, such as tracing the root causes of anomalies. Existing approaches typically rely on dataset-specific optimization, making it difficult to transfer their causal discovery capabilities to new time series governed by diverse causal mechanisms. In this paper, we propose **PTCD**, a novel **P**retraining framework for **T**ime-series **C**ausal **D**iscovery, which improves cross-task generalization through context-conditioned modeling and transferable causal augmentation. To model complex temporal causal dependencies, PTCD employs a dual-scale iterative attention mechanism to capture window-level causal relationships, and a Gaussian mixture with a context-level routing mechanism to handle heterogeneous exogenous distributions. To further address distribution shifts across causal graphs, PTCD adopts a pretraining paradigm on synthetic datasets that integrates intervention-based learning and a causal mixup strategy, promoting stable causal discovery and stronger generalization. Extensive experiments on multiple real-world out-of-distribution (OOD) datasets demonstrate that PTCD excels in both causal discovery and root cause identification.

1 INTRODUCTION

Accurate causal discovery from observed time series is fundamental to many real-world applications. It can reveal the drivers of stock price fluctuations (Li et al., 2024) or the factors influencing river water levels (Stein et al., 2025), thereby enabling actionable insights and more robust decision-making. In anomaly detection (Liu et al., 2025; Kim et al., 2025), monitoring changes in causal relationships can effectively surface anomalous events. In root cause diagnosis (Han et al., 2025; Nagalapatti et al., 2025), a causal graph helps localize the source of system failures and trace their propagation.

Existing methods (Han et al., 2025) typically use end-to-end optimization to infer causal structures from individual time series. However, as shown in Figure 1a, such end-to-end causal discovery methods generalize poorly across different causal mechanisms and require costly per-dataset optimization, especially for time series with complex temporal dependencies. We therefore propose a pretraining paradigm that learns transferable causal discovery capabilities from diverse time series and causal graphs, enabling stronger generalization to new time series and supporting accurate causal graph recovery in both zero-shot settings and with lightweight fine-tuning. Nevertheless, building a generalizable pretraining framework for time series causal discovery faces two main challenges.

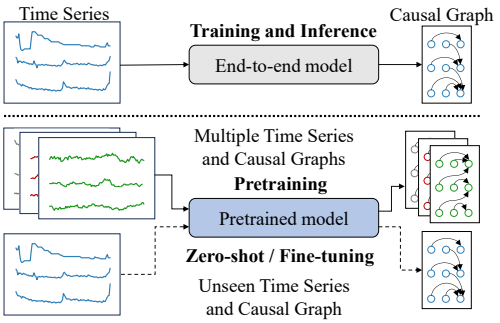
Challenge 1: *Inherent complexity of temporal causal dependencies.* As shown in Figure 1b, channel 3 at time t is influenced by channel 1 at $t-2$, which in turn is influenced by channel 2 at $t-4$. This reveals an iterative, cascading dependency across time steps. Existing methods (Cheng et al., 2023; Han et al., 2025) inadequately capture such iterative cross-window interactions. This suggests that reliable temporal causal discovery requires modeling dependencies beyond individual windows. Furthermore, in structural causal models, observed time series (endogenous variables) are influenced not only by other endogenous variables but also by unobserved factors (exogenous variables). Since exogenous distributions vary across time series and causal structures, existing methods (Pamfil et al., 2020; Han et al., 2025) assuming a fixed prior lack the flexibility to model diverse causal mechanisms.

Challenge 2: *Lack of causality-aware generalization under distribution shifts.* Real-world time series often undergo distribution shifts caused by changing environments or system dynamics, leading to

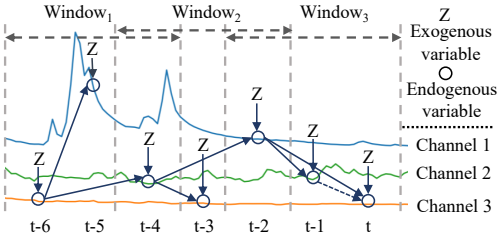
Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

out-of-distribution (OOD) scenarios. Therefore, a robust causal discovery model should generalize beyond the training distribution. However, existing methods fail to account for both the diversity of individual series and the variability of causal structures across different series. Additionally, in causal discovery, distinguishing between correlation and causation is crucial (Pearl & Mackenzie, 2018). As shown in Figure 1b, both channel 3 at time t ($x_{3,t}$) and channel 2 at $t-1$ ($x_{2,t-1}$) are influenced by channel 1 at $t-2$, which induces correlation between them. However, this does not imply a direct causal link from $x_{2,t-1}$ and $x_{3,t}$. What truly matters is determining whether a causal relationship exists. Unfortunately, existing deep learning methods struggle to transfer this causal reasoning capability to unseen time series and causal graphs.

For **Challenge 1,** we propose a *hierarchical context-conditioned temporal causal discovery modeling approach*, which learns window-level representations at multiple scales and captures the distributions of exogenous variables in time series. Our framework features two key components: (1) *Dual-scale iterative representation enhancement*, which employs an attention-based mechanism to iteratively model both intra-window and inter-window dependencies, capturing fine-grained causal dynamics within windows while accounting for coarse-grained long-range causal propagation across windows; and (2) *Context-conditioned exogenous variable estimation*, which employs an adaptive mixture-of-Gaussians model with routing to approximate exogenous variable distributions in a context-conditioned manner, thereby enhancing the robustness of causal discovery across diverse time series.



(a) End-to-end method v.s. Pretraining method.



(b) Windowed causal dependencies.

Figure 1: **(a)** End-to-end vs. pretraining paradigm: end-to-end models adopt dataset-specific optimization, while pretrained models learn from multiple time series and causal graphs, enabling zero-shot inference or fine-tuning on unseen time series/graphs. **(b)** Intra- and inter-window causal dependencies in multivariate time series x. Solid lines denote causal dependencies, while dashed lines indicate spurious correlations.

For **Challenge 2,** we propose *generalizable causal learning with causal augmentation* to strengthen the model's perception of causal relationships and improve its OOD robustness. Specifically, the *causal pretext task* introduces interventions into time series, breaking spurious correlations and providing strong signals for learning more generalizable causal structures. By treating the prediction of the intervened window as a pretext task, the model is encouraged to exploit discrepancies across varying time series and acquire transferable causal discovery capabilities. Furthermore, *Time series causal mixup* operates by jointly mixing latent representations and their associated causal graphs, generating augmented sequences with diverse causal strengths and dependency patterns. This not only enriches the training data with diverse causal scenarios but also compels the model to generalize beyond previously observed correlation structures. In summary, our contributions are as follows:

- We propose a novel pretraining paradigm for time-series causal discovery that can rapidly adapt to new time series and causal graphs.
- We introduce a hierarchical context-conditioned temporal causal discovery modeling that captures both inter-window and intra-window dependencies in individual time series, while flexibly handling context-conditioned exogenous noises across multiple series.
- We develop a generalizable causal pretraining strategy that employs intervention-based pretext tasks to break spurious correlations and reveal causal invariance in time series, alongside a causal mixup to achieve smoother gradients and improved robustness.
- Experiments on multiple real-world datasets show that PTCD consistently outperforms existing methods in both causal discovery and root cause identification.

2 Related Work

Causal Discovery for Time Series. Granger causality-based methods (Granger, 1969; Dahlhaus & Eichler, 2003) assume that if past values of X improve the prediction of future values of Y, then X is a Granger cause of Y. This idea has been extended with various neural architectures. cLSTM (Tank et al., 2021) leverages RNNs to infer Granger causal structures. TCDF (Nauta et al., 2019) utilizes attention-based CNNs for interpretable causal discovery. GVAR (Marcinkevics & Vogt, 2021) adopts a vector autoregressive model with generalized coefficient matrices to increase modeling flexibility. CUTS (Cheng et al., 2023; 2024) learns a sparse causal adjacency matrix directly from data. However, these methods share a key limitation: they do not explicitly model endogenous errors or exogenous noise, which restricts their applicability in practical scenarios. In contrast, Structural Causal Model (SCM) approaches explicitly characterize functional relationships among endogenous variables and account for exogenous influence. Varlingam (Hyvärinen et al., 2010) combines a non-Gaussian instantaneous causality with vector autoregressive dynamics. TiMINo (Peters et al., 2013) assumes that exogenous variables are independent over time. AERCA (Han et al., 2025) uses an autoencoder to model both causal relationships and the distribution of exogenous variables. However, existing methods largely ignore distribution shifts, limiting their ability to generalize time series causal discovery across multiple datasets. To address these gaps, we leverage a causal pretext task to uncover causal relationships in temporal sequences and causal mixup to generate diverse dependency patterns, thereby improving OOD generalization for causal discovery.

Table 1: Comparison of causal discovery methods. The symbol ✓ indicates that a component is incorporated, while × denotes its absence. Fine-grained and coarse-grained denote intra- and inter-window dependency, respectively. Exogenous refers to modeling of exogenous variables, and Generalization indicates the ability to handle unseen time series under different causal mechanisms.

Methods	Fine-grained	Coarse-grained	Exogenous	Generalization
CDMI (Ahmad et al., 2024)	✓	×	×	×
CUTS (Cheng et al., 2023)	✓	×	×	×
TCDF (Nauta et al., 2019)	✓	✓	×	×
CP (Stein et al., 2024)	×	✓	×	✓
AERCA (Han et al., 2025)	✓	×	✓	×
PTCD	✓	✓	✓	✓

Pre-trained Causal Discovery. Recent efforts have integrated causal discovery into pre-trained models (Lorch et al., 2022; Ke et al., 2023). CaML (Nilforoshan et al., 2023) formulates personalized effect prediction as a meta-task for zero-shot generalization. Cond-FiP (Mahajan et al., 2024) infers structural causal models via a Fixed-Point Approach (Scetbon et al., 2024). CInA (Zhang et al., 2024) learns transferable causal representations from unlabeled data, enabling zero-shot inference without fine-tuning. However, these methods do not explicitly capture temporal causality. CP (Stein et al., 2024) mitigates this by learning window-based causal graphs via supervised training across four architectures, yet it overlooks key temporal properties such as multi-scale dependencies and exogenous variable effects (see Table 1). To overcome these limitations, we leverage dual-scale attention to capture inter-window and intra-window temporal dependencies, and context-conditioned exogenous estimation to model the distribution of exogenous variables in each time series.

3 Problem Formulation

We consider a multivariate time series $\boldsymbol{x} = \{x_{1:T,i}\}_{i=1}^{C}$ with C channels, where $x_{1:T,i}$ denotes the length-T sequence for channel i. Following AERCA (Han et al., 2025), we assume causal sufficiency, no instantaneous effects, and time-invariant causal relations. For structural identifiability, we assume that the system satisfies the causal minimality (Varambally et al., 2024) and follows a continuous nonlinear Additive Noise Model (Peters et al., 2010). Under these assumptions, the data generation process is modeled as a time-invariant structural causal model (SCM):

$$x_{t,i} = g_i\left(\sum_{k=0}^{n-1}\sum_{j=1}^{C} G_{t-n+k,j,i} \cdot f_{k,j,i}(x_{t-n+k,j})\right) + Z_{i,t}, \tag{1}$$

where n is the maximum time lag, $\boldsymbol{G} \in \mathbb{R}^{n\times C\times C}$ is the causal graph, and $G_{t-n+k,j,i}$ indicates the impact of channel j at time $t-n+k$ on $x_{t,i}$. The operator $\cdot$ denotes scalar multiplication, and the functions $g_i(\cdot)$ and $f_{k,j,i}(\cdot)$ are nonlinear transformations of the past observations, and $Z_{i,t}$ is an exogenous noise term for channel i at time t. In this work, we focus on pretraining a causal discovery model $G = F_\theta(\boldsymbol{x})$ that acquires transferable causal discovery capability. The learned

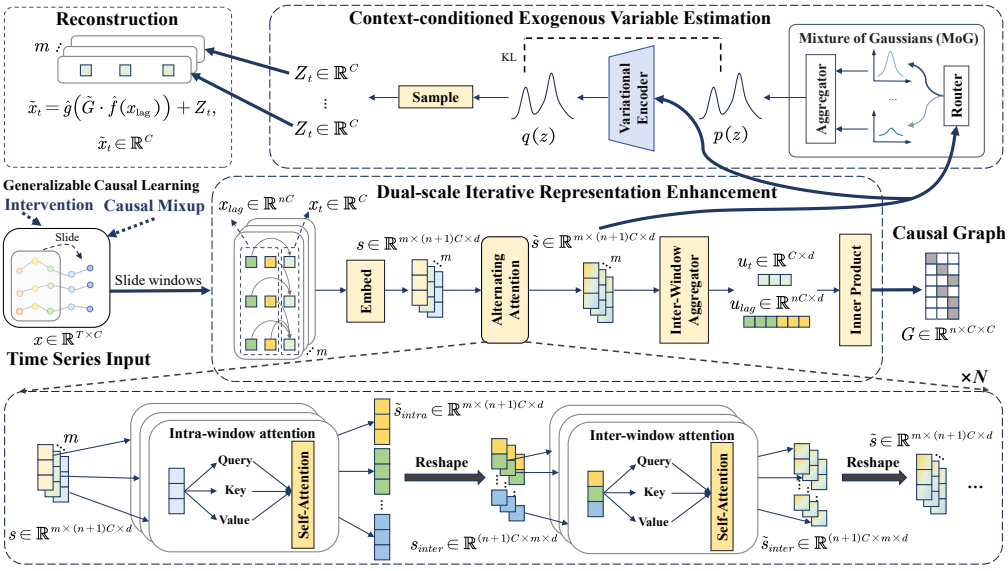


Figure 2: Overview of the hierarchical context-conditional temporal causal discovery framework, comprising: (1) Dual-scale Iterative Representation Enhancement, and (2) Context-conditioned exogenous variable estimation. The model's generalization is further improved via transferable causality-augmented pretraining, incorporating intervention pretext task and causal mixup strategies.

model is expected to generalize to unseen time series and enable the discovery of causal graphs that are not observed during pretraining. Moreover, we further transfer the pretrained model to real-world settings for causal discovery and root cause identification.

4 METHODOLOGY

We propose **PTCD**, a generalizable framework for time series causal discovery with context-conditioned modeling and causality-augmented pretraining (Figure 2). To capture complex temporal dependencies and diverse exogenous influences, we introduce a **hierarchical context-conditioned causal discovery framework**. The dual-scale iterative representation enhancement module segments time series into sliding windows and alternates intra-window and inter-window attention to refine temporal representations. An inter-window aggregator then produces past and current node embeddings, whose inner products estimate the window causal graph. To model heterogeneous exogenous variables, the context-conditioned exogenous variable estimation module adopts a routing-based mixture-of-Gaussians prior and reconstructs the current series from the estimated causal graph and past observations. To improve generalization, we introduce **transferable causal learning with causal augmentation**. The intervention pretext task performs interventions on the generative process of time series and formulates a prediction task for the intervened window, thereby eliminating spurious correlations and enabling stable causal discovery across different environments. causal mixup regularizes causal dependencies by jointly interpolating latent representations and causal graphs. Finally, PTCD reconstructs time series by jointly modeling causal structure, endogenous dynamics, and exogenous factors, enabling interpretable and transferable causal discovery.

4.1 HIERARCHICAL CONTEXT-CONDITIONAL TEMPORAL CAUSAL DISCOVERY FRAMEWORK

Dual-scale Iterative Representation Enhancement. To model complex causal dependencies over both short and long time scales, as shown in Figure 2, given a multivariate time series $x \in \mathbb{R}^{T\times C}$, we divide it into m windows of size $(n+1)$, yielding $x_{window} \in \mathbb{R}^{m\times(n+1)C}$. To model the data generation process of the current time step x_t, we represent each window as the concatenation of past observations $x_{lag} \in \mathbb{R}^{nC}$ and the current step $x_t \in \mathbb{R}^{C}$. We then embed x_{window} to obtain the window representation $s_{intra} \in \mathbb{R}^{m\times(n+1)C\times d}$, where d represents the dimension of the embedding. Intra-window features capture local temporal dynamics, while inter-window relations provide global context. To jointly model these complementary scales, we propose an alternating

attention mechanism that iteratively integrates features within and across windows. First, we apply self-attention within each window. The sequence s is projected into Q^s_{intra}, K^s_{intra}, and V^s_{intra}, and compute the intra-window representation as:

$$\tilde{s}_{intra} = \text{Softmax}\Big(Q^s_{intra}(K^s_{intra})^\top / \sqrt{d_k}\Big)\, V^s_{intra}. \tag{2}$$

Next, we perform inter-window self-attention to capture dependencies across windows. Specifically, the $\tilde{s}_{intra} \in \mathbb{R}^{m\times(n+1)C\times d}$ is reshaped as $s_{inter} \in \mathbb{R}^{(n+1)C\times m\times d}$, projected into Q^s_{inter}, K^s_{inter}, and V^s_{inter}, and processed as:

$$\tilde{s}_{inter} = \text{Softmax}\Big(Q^s_{inter}(K^s_{inter})^\top / \sqrt{d_k}\Big)\, V^s_{inter}. \tag{3}$$

The inter-window output $\tilde{s}_{inter}$ is reshaped back and fed into the next iteration. Repeating this alternating process for N iterations progressively refines the representation, yielding the dual-scale enhanced representation $\tilde{s} \in \mathbb{R}^{m\times(n+1)C\times d}$. We then apply an inter-window aggregator over $\tilde{s}$ to obtain node features $u \in \mathbb{R}^{(n+1)C\times d}$ for window causal graph construction. Following the partition of x_{lag} and x_t, we split u into past nodes u_{lag} and the current nodes u_t. Finally, we encode u_{lag} and u_t separately, and compute their inner product to estimate the link probability matrix $\tilde{G}$:

$$\begin{aligned} u = Aggregator(\tilde{s}), \quad \tilde{G} = \text{MLP}(u_{lag})\text{MLP}(u_t)^T, \\ u \in \mathbb{R}^{(n+1)C\times d}, \quad u_{lag} \in \mathbb{R}^{nC\times d}, \quad u_t \in \mathbb{R}^{C\times d}. \end{aligned} \tag{4}$$

Here, $\tilde{G}$ is a binary matrix indicating the presence or absence of causal links. The causal graph is trained to align the estimated link probability matrix $\tilde{G}$ with its ground-truth matrix G as:

$$\text{BCE}(G, \tilde{G}) = -G \cdot \log(\tilde{G}) - (1 - G) \cdot \log(1 - \tilde{G}). \tag{5}$$

Context-conditioned Exogenous Variable Estimation. Considering the diversity of time series and their underlying causal mechanisms, the distributions of exogenous variables can vary substantially across datasets. Therefore, we propose a context-conditioned approach that leverages the dual-scale enhanced representation $\tilde{s}$ as contextual input to estimate exogenous variable distributions. According to Eq. 1, we define a data-generating distribution $p(D)$ that jointly encapsulates the underlying causal structure and the influence of exogenous factors:

$$x \sim p(D) = \int_G \int_z p(D \mid G, z) \cdot p(G) \cdot p(z) dz dG. \tag{6}$$

Here, a sample x is generated by first drawing from the prior distribution of the exogenous variable $p(z)$ and the causal structure distribution $p(G)$, and then iteratively generating data through the data-generating mechanism $p(D|G, z)$.

In time series, noise and external disturbances often exhibit diverse patterns, reflecting heterogeneous exogenous influences. Therefore, a single shared Gaussian prior is insufficient to model such variability. To address this, we model the exogenous prior as a context-conditioned Gaussian mixture, where the mixing coefficients $\boldsymbol{\pi} \in \mathbb{R}^K$ are generated from the enhanced representation $\tilde{s}$ via a routing mechanism. The means μ and variances Σ of the K Gaussian components are learnable parameters:

$$\boldsymbol{\pi} = \text{Router}(\tilde{s}), p(z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(z \mid \mu_k, \Sigma_k), \sum_{k=1}^{K} \pi_k = 1, \tag{7}$$

where Router$(\cdot)$ maps the context $\tilde{s}$ to mixing weights, and $\mathcal{N}$ denotes the Gaussian probability density function. We employ a variational encoder to approximate the posterior distribution $q(z)$ of the exogenous variables conditioned on the temporal representation $\tilde{s}$. Using the reparameterization trick, we sample the exogenous variable Z_t from $q(z)$ at each time step t. According to Eq. 1, we then incorporate the sampled variable Z_t to reconstruct the current time series x_t, together with the predicted causal graph $\tilde{G}$ and the lagged endogenous variables x_{lag}. In this manner, both exogenous influences and the causal dependencies are explicitly modeled, ensuring that the learned representations are consistent with a structural causal model. The reconstruction of x_t is given by:

$$\tilde{x}_t = \tilde{g}\left(\tilde{G}_p \cdot \tilde{f}(x_{lag})\right) + Z_t, \tag{8}$$

where $\tilde{g}(\cdot)$ and $\tilde{f}(\cdot)$ denote multi-layer perceptrons. We use the reconstruction error $\|x_t - \tilde{x}_t\|_2$ as the supervision signal for both pre-training and fine-tuning.

To regularize the latent space, we introduce a KL divergence term between the learned posterior distribution $q(z)$ and the assumed prior mixture Gaussian distribution $p(z)$:

$$\mathcal{L}_{KL}(p(z) \parallel q(z)) = \int p(\mathbf{z}) \log \frac{p(\mathbf{z})}{q(\mathbf{z})} \, d\mathbf{z}, \tag{9}$$

4.2 Transferable Causal Learning with Causal Augmentation

Intervention Pretext Task. Time series often exhibit autocorrelation and interdependencies, where past values influence the present, and multivariate series may involve cross-channel causality. Relying solely on observational data can confuse correlation and causality, making it difficult to distinguish between causal graphs within the same Markov equivalence class (Li et al., 2023; Tian & Pearl, 2001). To improve causal identifiability, we introduce an intervention pretext task that leverages synthetic interventions for causal structure learning. Specifically, we randomly select a time interval $[t_1, t_2]$ from the time series x and replace the observation at t_1 with a randomly sampled value. Based on the underlying causal dynamics, we then iteratively propagate this intervention to compute the intervened segment $(x'_{t_1}, \cdots, x'_{t_2})$, replacing the original segment $(x_{t_1}, \cdots, x_{t_2})$ to obtain a modified series x^{do}. By examining segmented windows of x^{do}, we identify those inconsistent with the causal structure to derive binary ground-truth intervention labels. To enhance generalization and encourage the model to capture stable causal dependencies, we train it to predict the window in which an intervention occurs. We formulate intervention detection as a classification task for intervention detection:

$$\mathcal{L}_{do_window} = \sum_{i=1}^{m} \Big(- h_i \log(\hat{h}_i) - (1 - h_i) \log(1 - \hat{h}_i) \Big), \tag{10}$$

where h_i and $\hat{h}_i$ denote the ground-truth and predicted intervention labels for window i, respectively.

Furthermore, Theorem 4.1 shows that optimizing the proposed intervention detection objective is equivalent to identifying the true causal mechanisms. The proof is provided in Appendix A.1.

Theorem 4.1. ***(Causal Identifiability via Intervention Detection)*** *Let $P_{obs}(\mathbf{x})$ denote the observational distribution induced by the true SCM, and $P_{int}(\mathbf{x})$ denote the interventional distribution where a subset of variables $\mathbf{x}_\mathcal{I}$ (e.g., a time window) is replaced by samples from a noise or marginal proposal distribution $Q(\mathbf{x}_\mathcal{I})$. The optimal discriminator D^* that minimizes the intervention classification loss is monotonically dependent on the true causal mechanism $P(\mathbf{x}_\mathcal{I} \mid Pa(\mathbf{x}_\mathcal{I}))$. Consequently, optimizing $\mathcal{L}_{do_window}$ necessitates the model to learn the true underlying causal structure and the associated conditional dependencies.*

To preserve semantic consistency under interventions, we apply contrastive learning to align the representation u^{do} of the intervened series x^{do} with the representation u of the original x using Eq. 4:

$$\mathcal{L}_{do_contrast} = - \sum_{i=1}^{(n+1)C} \log \frac{\exp\big(\mathrm{sim}(u^{do,+}, u_i)/\tau\big)}{\sum_{j=1}^{(n+1)C} \exp\big(\mathrm{sim}(u_j^{do}, u_i)/\tau\big)}, \tag{11}$$

where τ is a temperature parameter, $sim(\cdot, \cdot)$ is the similarity function, and $u^{do,+}$ denotes the positive pair of u_i from the same node in the window causal graph. We jointly optimize the intervention detection objective and contrastive objective, *i.e.,* $\mathcal{L}_{do} = \mathcal{L}_{do_window} + \mathcal{L}_{do_contrast}$ to enhance causal discovery capability. The reconstruction loss on x^{do} is computed following Eq. 8.

Time Series Causal Mixup. To improve the generalization of the causal discovery framework, we propose a time series causal mixup strategy for temporal causal discovery. This operation smooths the solution space, yielding more stable and continuous estimates of edge existence. Specifically, we randomly sample two time series from the training dataset, each with latent embedding $\tilde{s}^i$ and its corresponding causal relation G^i. We then construct mixed samples as:

$$\tilde{s}^m = \lambda_1 \tilde{s}^1 + \lambda_2 \tilde{s}^2, \quad G^m = \lambda_1 G^1 + \lambda_2 G^2, \tag{12}$$

where the mixing coefficient λ_i is drawn from a Beta distribution. By blending samples from diverse time series and varying causal graphs, time series causal mixup achieves a smoother optimization landscape during training, enhancing the transferability of causal dependency modeling.

We further present Theorem 4.2, which suggests that causal mixup enforces the learned causal probabilities to be transferable and physically consistent, avoiding the artifacts introduced by input-space interpolation. The proof is provided in Appendix A.2. Since function smoothness is closely related to generalization (Sokolic et al., 2017), PTCD achieves transferable causal discovery capability.

Theorem 4.2. ***(Geodesic Consistency of Causal Probability)*** *Under Assumption A.2, linear interpolation in the latent space $\mathcal{S}$ approximates a geodesic interpolation on the data manifold $\mathcal{M}$. Minimizing the Mixup empirical risk objective, which penalizes the discrepancy between the network prediction $\hat{G}(s^{mix})$ and the interpolation target $\lambda G(x_1) + (1-\lambda)G(x_2)$ under a bounded loss ϵ, inherently imposes an implicit gradient penalty proportional to the Frobenius norm of the Jacobian, $\mathcal{O}(\|\nabla_s \hat{G}(s)\|_F^2)$. This bounds the local Lipschitz constant of the causal mapping, regularizing the Jacobian field to be smooth along the transport path and ensuring stable causal graph discovery.*

Following Eq. 5, we compute the binary cross-entropy loss $\text{BCE}(G^m, \tilde{G}^m)$ between the estimated link probability matrix $\tilde{G}$ and the mixup link probability matrix G^m.

4.3 Causal Learning and Root Cause Identification

Pretraining Phase. During pretraining, we generate synthetic datasets by iteratively simulating time series from causal graphs (see Appendix B.1) to train the model. Given a time series x, the objective function consists of several components: the reconstruction loss $\mathcal{L}_{recon}$, the causal graph loss $\mathcal{L}_G$, the KL divergence loss L_{KL}, the intervention loss $\mathcal{L}_{do}$. The overall pre-training loss is defined as:

$$\mathcal{L}_{pre} = \mathcal{L}_{recon} + \lambda_G \mathcal{L}_G + \lambda_{KL} \mathcal{L}_{KL} + \lambda_{do} \mathcal{L}_{do}, \tag{13}$$

where λ_G, λ_{KL}, λ_{do} are the corresponding weighting coefficients. The reconstruction loss $\mathcal{L}_{\text{recon}}$ is computed over two types of input, the original series, the intervened series, and is defined as $\mathcal{L}_{recon} = \|x_t - \tilde{x}_t\|_2 + \left\|x_t^{do} - \tilde{x}_t^{do}\right\|_2$. The causal graph loss $\mathcal{L}_G$ is computed over two types of input, the original series, the mixup series, and is defined as $\mathcal{L}_G = \text{BCE}(G, \tilde{G}) + \text{BCE}(G^m, \tilde{G}^m)$.

Fine-tuning Phase. During fine-tuning, we address two downstream tasks: *Causal Discovery* and *Root Cause Identification*. In real-world scenarios, the true data generation process is typically unavailable. Hence, we focus solely on the reconstruction loss with s_t, and the KL divergence loss L_{KL}. The overall fine-tuning loss is as follows:

$$\mathcal{L}_{ft} = \|\tilde{s}_t - s_t\|_2 + \lambda_{KL} \mathcal{L}_{KL}. \tag{14}$$

For *Root Cause Identification*, we score exogenous variable abnormality to pinpoint root causes. We compute the mean (μ) and standard deviation (σ) of the exogenous distribution from normal data, and compute the root cause score as the z-score, $\text{score}_t = \min_{k=1}^{K} \left(\frac{Z_t - \mu_k}{\sigma_k} \right)$. We then adopt Streaming Peaks-Over-Threshold (Siffer et al., 2017) for threshold selection.

5 Experiments

5.1 Experimental Design

Datasets. We construct a synthetic time series causal dataset (see Appendix B.1) for pretraining and fine-tune the model on two real-world tasks. *Causal Discovery.* On the German river benchmark (Stein et al., 2025), which includes two variable types and five causal relation types, we fine-tune on Bavaria and evaluate on Eastern Germany. *Root Cause Identification.* We fine-tune and evaluate on SWaT (Mathur & Tippenhauer, 2016), a water treatment testbed dataset, and MSDS (Nedelkoski et al., 2020), an OpenStack infrastructure dataset. See Appendix B.1 for details.

Evaluation Metrics. *Causal Discovery.* To avoid threshold tuning, we report the AUROC score (Moraffah et al., 2021; Hasan et al., 2023; Sun et al., 2023; Nauta et al., 2019). *Root Cause Identification.* Following prior work (Ikram et al., 2022; Yu et al., 2021; Ma et al., 2020; Han et al., 2025), we evaluate using recall at top-k, denoted as $Recall@K$. Further details are in Appendix B.3.

Table 2: Causal discovery results of AUROC on Eastern Germany river datasets, best in bold and second-best underlined.

Model	TCDF	CUTS+	VAR	Varlingam	Dynotears	PCMCI	CDMI	CP(Gru)	CP(Transformer)	PTCD
Close (3)	0.57	0.56	<u>0.81</u>	0.79	0.50	0.64	<u>0.81</u>	0.79	0.75	**0.84**
Close (5)	0.51	0.54	0.81	0.77	0.50	0.62	0.81	0.81	<u>0.83</u>	**0.85**
Root cause (3)	0.55	0.56	0.79	0.77	0.56	0.70	0.75	0.78	<u>0.84</u>	**0.88**
Root cause (5)	0.54	0.59	0.75	0.77	0.56	0.74	0.65	0.81	<u>0.84</u>	**0.86**
Random+1 (3)	0.72	0.56	0.80	<u>0.84</u>	0.52	0.83	0.82	0.82	0.82	**0.87**
Random+1 (5)	0.56	0.57	0.79	0.79	0.61	0.74	0.80	<u>0.84</u>	**0.85**	**0.85**
Confounder (3)	0.62	0.55	<u>0.71</u>	0.68	0.53	0.66	0.63	0.64	0.65	**0.81**
Confounder (5)	0.56	0.51	<u>0.72</u>	0.70	0.53	0.64	0.71	0.71	0.71	**0.79**
Random (3)	0.55	0.57	<u>0.82</u>	0.79	0.50	0.65	0.80	0.77	0.81	**0.85**
Random (5)	0.52	0.54	0.80	0.75	0.51	0.65	0.78	<u>0.83</u>	**0.86**	**0.86**
avg	0.57	0.56	0.78	0.77	0.53	0.69	0.76	0.78	0.80	**0.85**

Table 3: Results of root cause identification, best in bold.

Dataset	Model	Recall@1	Recall@3	Recall@5	Recall@10	Avg@10
MSDS	ϵ-Diagnosis	0.004±0.004	0.266±0.002	0.452±0.009	**1.000±0.000**	0.492±0.001
	RCD	0.412±0.048	0.573±0.010	0.984±0.001	**1.000±0.000**	0.821±0.012
	CIRCA	0.454±0.238	0.860±0.140	0.917±0.084	**1.000±0.000**	0.809±0.035
	AERCA	0.381±0.408	0.908±0.062	0.974±0.027	**1.000±0.000**	0.896±0.037
	PTCD	**0.515±0.252**	**0.993±0.004**	**0.996±0.003**	**1.000±0.000**	**0.929±0.013**
SWaT	ϵ-Diagnosis	0.075±0.179	0.125±0.217	0.125±0.217	0.375±0.383	0.180±0.194
	RCD	0.000±0.000	0.000±0.000	0.000±0.000	0.300±0.458	0.100±0.161
	CIRCA	0.000±0.000	0.000±0.000	0.000±0.000	0.300±0.458	0.100±0.161
	AERCA	0.220±0.111	0.290±0.088	0.330±0.048	0.455±0.044	0.342±0.052
	PTCD	**0.300±0.132**	**0.450±0.107**	**0.450±0.137**	**0.475±0.141**	**0.440±0.105**

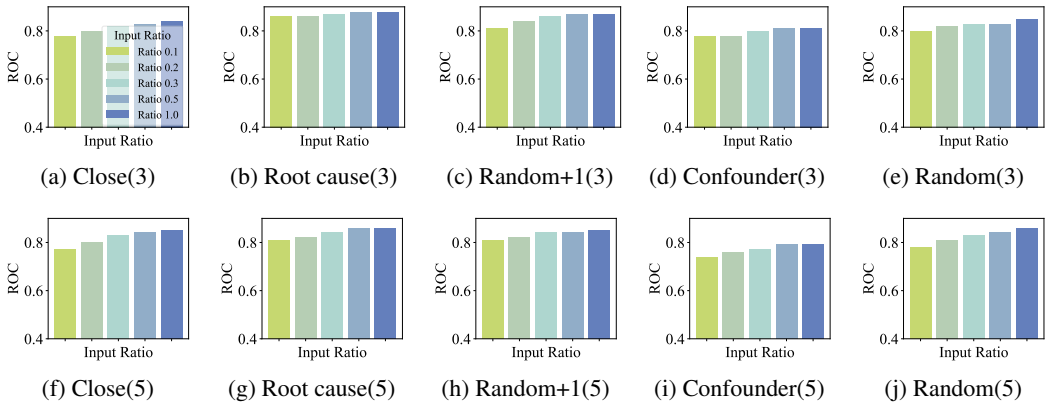
Baselines. *Causal Discovery.* We evaluate methods spanning classical and modern paradigms. Classical methods include PCMCI (Runge et al., 2019), Varlingam (Hyvärinen et al., 2010), Dynotears (Pamfil et al., 2020), VAR (Assaad et al., 2023), CDMI (Ahmad et al., 2024), TCDF (Nauta et al., 2019), CUTS+ (Cheng et al., 2024). Among pretraining methods, we compare with Causal Pretraining (CP) (Stein et al., 2024), which is implemented through both Transformer-based and GRU-based architectures. *Root Cause Identification.* We compare PTCD with four baselines: ε-Diagnosis (Shan et al., 2019), RCD (Ikram et al., 2022), CIRCA (Li et al., 2022), and AERCA (Han et al., 2025).

5.2 Experimental Results

Causal Discovery. Table 2 presents the results for time series causal graph discovery, demonstrating that PTCD consistently outperforms all baselines across various types of causal relationships. This highlights the effectiveness of its dual-scale iterative representation enhancement and context-conditioned exogenous variables estimation in capturing complex causal dependencies. This approach enables the model to accurately identify true causal dependencies among time series while filtering out spurious correlations caused by confounders or environmental biases. For example, we observe that PTCD achieves strong causal discovery performance on the Eastern Germany river dataset, despite not being directly trained on it. Notably, compared to CP, our method demonstrates a significant improvement, with an average gain of 5%. These results confirm that PTCD acquires a generalizable mechanism for causal discovery through intervention-based modeling and causal mixup, ensuring robustness even under real-world distribution shifts.

Root Cause Identification. Table 3 shows that PTCD achieves the best performance on real-world datasets in the root cause diagnosis task. Unlike AERCA, which primarily focuses on intra-window modeling, PTCD leverages a dual-scale iterative representation enhancement to capture both coarse-grained and fine-grained temporal patterns. This approach enhances the model's ability to detect subtle variations over time, leading to superior performance in identifying the root cause of anomalies. Furthermore, the context-conditioned exogenous variable estimation boosts PTCD to adapt to the complex and often noisy conditions found in real-world data. According to the AC@1 metric, PTCD excels in accurately identifying the time series with the highest root cause score, demonstrating that incorporating intervention-based pre-training enables more stable detection of anomalous root causes even in the presence of dynamic shifts and external disturbances.

Ablation Studies. To assess the impact of different components in PTCD, we conduct an ablation



(a) Close(3) (b) Root cause(3) (c) Random+1(3) (d) Confounder(3) (e) Random(3)

(f) Close(5) (g) Root cause(5) (h) Random+1(5) (i) Confounder(5) (j) Random(5)

Figure 3: Results on the Eastern Germany river dataset (avg. length 3,143) under test input ratios of 10%, 20%, 30%, 50%, and 100%.

Table 4: Ablation study of causal discovery using the Eastern Germany river dataset. w/o inter, w/o intra, w/o exogenous, w/o causal mixup, w/o intervention represent removing the inter-window attention, intra-window attention, exogenous, causal mixup, and intervention, respectively.

Model	w/o inter	w/o intra	w/o exogenous	w/o causal mixup	w/o intervention	PTCD
Close (3)	0.83	0.82	0.80	0.83	0.80	**0.84**
Close (5)	0.84	0.82	0.81	0.83	0.81	**0.85**
Root cause (3)	0.86	0.84	0.82	0.86	0.83	**0.88**
Root cause (5)	0.85	0.84	0.82	0.85	0.83	**0.86**
Random+1 (3)	0.85	0.84	0.83	0.86	0.84	**0.87**
Random+1 (5)	0.83	0.83	0.82	0.83	0.81	**0.85**
Confounder (3)	0.79	0.79	0.78	0.80	0.75	**0.81**
Confounder (5)	0.78	0.76	0.75	0.77	0.74	**0.79**
Random (3)	0.84	0.84	0.81	0.83	0.82	**0.85**
Random (5)	0.85	0.84	0.83	0.84	0.83	**0.86**
avg	0.83	0.82	0.81	0.83	0.81	**0.85**

study on inter-window attention, intra-window attention, mixture of Gaussians exogenous valuables, causal mixup, and intervention. As shown in Table 4, each module contributes to the overall performance. Notably, removing exogenous variable estimation significantly reduces performance, highlighting the importance of learning context-conditioned distributions. Furthermore, the inclusion of intervention yields notable performance improvements, particularly in datasets where the root cause is confounded (Confounder (3) and Confounder (5)). This demonstrates that interventions help the model distinguish between spurious correlations and true causal relationships, enhancing the accuracy of causal graph prediction. Further ablation studies are in Appendix B.5.

Different Time Series Ratio. Figure 3 shows the impact of varying lengths of the input time series on performance. We find that using only 30% of the full sequence length yields results comparable to the complete input, while even 10% still provides reasonable results across all datasets. This shows that PTCD remains effective under limited resources and efficient for causal discovery using reduced temporal inputs, making it suitable for real-world applications.

6 Conclusions

This paper introduces PTCD, a novel pretraining framework for generalizable time series causal discovery across diverse downstream tasks. To capture complex temporal dependencies, PTCD employs a dual-scale attention mechanism for both fine-grained local dynamics and long-range causal relationships, complemented by a context-conditioned Gaussian mixture to accommodate heterogeneous exogenous variable distributions flexibly. To handle distribution shifts, PTCD incorporates a causal mixup and intervention-based learning to achieve causal augmentation, promoting stable causal discovery and stronger generalization. Experiments on real-world OOD datasets show that PTCD outperforms existing methods in causal discovery and root cause identification, establishing it as a robust foundation for time series analysis.

AI USE STATEMENT

In this work, we used generative AI tools to improve the clarity, grammar, and overall quality of the writing. We have not used generative AI tools for generating the core ideas, formulating the research questions, and developing the methodology. We have reviewed and edited all AI-assisted work. In particular, we examined the correctness, logical coherence, and contextual appropriateness of any AI-generated text to ensure that it accurately reflected our intended meaning and met academic standards. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is conducted entirely on publicly available benchmark datasets, as detailed in the paper, and does not involve the release of any personal or sensitive information. No human subjects are involved in this research, ensuring that our work complies with ethical standards in research integrity.

REPRODUCIBILITY STATEMENT

All experiments in this work are conducted on publicly available benchmark datasets, and the reported performance of PTCD is fully reproducible. Details of model architecture, training procedures, and evaluation protocols are provided in the main text and appendix. To further facilitate reproducibility, we release the code and datasets at https://anonymous.4open.science/r/PTCD-D4E3.

REFERENCES

Wasim Ahmad, Valentin Kasburg, Nina Kukowski, Maha Shadaydeh, and Joachim Denzler. Deep-learning based causal inference: A feasibility study based on three years of tectonic-climate data from moxa geodynamic observatory. *Earth and Space Science*, 11(10):e2023EA003430, 2024.

Charles K. Assaad, Emilie Devijver, and Éric Gaussier. Survey and evaluation of causal discovery methods for time series (extended abstract). In *IJCAI*, pp. 6839–6844, 2023.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286 (5439):509–512, 1999.

Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE TPAMI*, 35(8):1798–1828, 2013.

Yuxiao Cheng, Runzhao Yang, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. CUTS: neural causal discovery from irregular time-series data. In *ICLR*, 2023.

Yuxiao Cheng, Lianglong Li, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. Cuts+: High-dimensional causal discovery from irregular time-series. In *AAAI*, volume 38, pp. 11525–11533, 2024.

Rainer Dahlhaus and Michael Eichler. Causality and graphical models in time series analysis. *Highly structured stochastic systems*, pp. 115–137, 2003.

Paul Erdős and Alfréd Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

Wenbo Gong, Joel Jennings, Cheng Zhang, and Nick Pawlowski. Rhino: Deep causal temporal relationship learning with history-dependent noise. In *ICLR*, 2023.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *nips*, pp. 2672–2680, 2014.

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

Xiao Han, Saima Absar, Lu Zhang, and Shuhan Yuan. Root cause analysis of anomalies in multivariate time series through granger causal discovery. In *ICLR*, 2025.

Uzma Hasan, Emam Hossain, and Md. Osman Gani. A survey on causal discovery methods for I.I.D. and time series data. *Trans. Mach. Learn. Res.*, 2023, 2023.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *J. Mach. Learn. Res.*, 11:1709–1731, 2010.

Azam Ikram, Sarthak Chakraborty, Subrata Mitra, Shiv Kumar Saini, Saurabh Bagchi, and Murat Kocaoglu. Root cause analysis of failures in microservices through causal discovery. In *NeurIPS*, 2022.

Nan Rosemary Ke, Silvia Chiappa, Jane X. Wang, Jörg Bornschein, Anirudh Goyal, Mélanie Rey, Theophane Weber, Matthew M. Botvinick, Michael Curtis Mozer, and Danilo Jimenez Rezende. Learning to induce causal structure. In *ICLR*, 2023.

HyunGi Kim, Jisoo Mok, Dongjun Lee, Jaihyun Lew, Sungjae Kim, and Sungroh Yoon. Causality-aware contrastive learning for robust multivariate time-series anomaly detection. *CoRR*, abs/2506.03964, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Adam Li, Amin Jaber, and Elias Bareinboim. Causal discovery from observational and interventional data across multiple environments. *Advances in Neural Information Processing Systems*, 36: 16942–16956, 2023.

Mingjie Li, Zeyan Li, Kanglin Yin, Xiaohui Nie, Wenchi Zhang, Kaixin Sui, and Dan Pei. Causal inference-based root cause analysis for online service systems with intervention recognition. In *KDD*, pp. 3230–3240, 2022.

Shuqi Li, Yuebo Sun, Yuxin Lin, Xin Gao, Shuo Shang, and Rui Yan. Causalstock: Deep end-to-end causal discovery for news-driven multi-stock movement prediction. In *NeurIPS*, 2024.

Zehao Liu, Mengzhou Gao, and Pengfei Jiao. GCAD: anomaly detection in multivariate time series from the perspective of granger causality. In *AAAI*, pp. 19041–19049, 2025.

Lars Lorch, Scott Sussex, Jonas Rothfuss, Andreas Krause, and Bernhard Schölkopf. Amortized inference for causal structure learning. In *nips*, 2022.

Meng Ma, Jingmin Xu, Yuan Wang, Pengfei Chen, Zonghua Zhang, and Ping Wang. Automap: Diagnose your microservice-based web applications automatically. In *WWW*, pp. 246–258, 2020.

Divyat Mahajan, Jannes Gladrow, Agrin Hilmkil, Cheng Zhang, and Meyer Scetbon. Zero-shot learning of causal models. *CoRR*, abs/2410.06128, 2024.

Ricards Marcinkevics and Julia E. Vogt. Interpretable models for granger causality using self-explaining neural networks. In *ICLR*, 2021.

Aditya P. Mathur and Nils Ole Tippenhauer. Swat: a water treatment testbed for research and training on ICS security. In *2016 International Workshop on Cyber-physical Systems for Smart Water Networks, CySWater@CPSWeek 2016*, pp. 31–36, 2016.

Raha Moraffah, Paras Sheth, Mansooreh Karami, Anchit Bhattacharya, Qianru Wang, Anique Tahir, Adrienne Raglin, and Huan Liu. Causal inference for time series analysis: problems, methods and evaluation. *Knowl. Inf. Syst.*, 63(12):3041–3085, 2021.

Lokesh Nagalapatti, Ashutosh Srivastava, Sunita Sarawagi, and Amit Sharma. Robust root cause diagnosis using in-distribution interventions. In *ICLR*, 2025.

Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Mach. Learn. Knowl. Extr.*, 1(1):312–340, 2019.

Sasho Nedelkoski, Jasmin Bogatinovski, Ajay Kumar Mandapati, Soeren Becker, Jorge Cardoso, and Odej Kao. Multi-source distributed system data for ai-powered analytics. In *ESOCC*, volume 12054 of *Lecture Notes in Computer Science*, pp. 161–176, 2020.

Hamed Nilforoshan, Michael Moor, Yusuf H. Roohani, Yining Chen, Anja Surina, Michihiro Yasunaga, Sara Oblak, and Jure Leskovec. Zero-shot causal learning. In *NeurIPS*, 2023.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. DYNOTEARS: structure learning from time-series data. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1595–1605, 2020.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, 2018.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Identifying cause and effect on discrete data using additive noise models. In *AISTATS*, volume 9 of *JMLR Proceedings*, pp. 597–604, 2010.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. In *NeurIPS*, pp. 154–162, 2013.

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

Meyer Scetbon, Joel Jennings, Agrin Hilmkil, Cheng Zhang, and Chao Ma. Fip: a fixed-point approach for causal generative modeling. *CoRR*, abs/2404.06969, 2024.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proc. IEEE*, 109(5): 612–634, 2021.

Huasong Shan, Yuan Chen, Haifeng Liu, Yunpeng Zhang, Xiao Xiao, Xiaofeng He, Min Li, and Wei Ding. ε-diagnosis: Unsupervised and real-time diagnosis of small- window long-tail latency in large-scale microservice platforms. In *WWW*, pp. 3215–3222, 2019.

Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouët. Anomaly detection in streams with extreme value theory. In *SIGKDD*, pp. 1067–1075. ACM, 2017.

Jure Sokolic, Raja Giryes, Guillermo Sapiro, and Miguel R. D. Rodrigues. Robust large margin deep neural networks. *IEEE Trans. Signal Process.*, 65(16):4265–4280, 2017.

Gideon Stein, Maha Shadaydeh, and Joachim Denzler. Embracing the black box: Heading towards foundation models for causal discovery from time series data. *CoRR*, abs/2402.09305, 2024.

Gideon Stein, Maha Shadaydeh, Jan Blunk, Niklas Penzel, and Joachim Denzler. Causalrivers - scaling up benchmarking of causal discovery for real-world time-series. In *ICLR*, 2025.

Xiangyu Sun, Oliver Schulte, Guiliang Liu, and Pascal Poupart. NTS-NOTEARS: learning nonparametric dbns with prior knowledge. In *International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 1942–1964, 2023.

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

Jin Tian and Judea Pearl. Causal discovery from changes. In *UAI*, pp. 512–521, 2001.

Sumanth Varambally, Yian Ma, and Rose Yu. Discovering mixtures of structural causal models from time series data. In *ICML*, 2024.

Guangba Yu, Pengfei Chen, Hongyang Chen, Zijie Guan, Zicheng Huang, Linxiao Jing, Tianjun Weng, Xinmeng Sun, and Xiaoyun Li. Microrank: End-to-end latency issue localization with extended spectrum analysis in microservice environments. In *WWW*, pp. 3087–3098, 2021.

Jiaqi Zhang, Joel Jennings, Agrin Hilmkil, Nick Pawlowski, Cheng Zhang, and Chao Ma. Towards causal foundation model: on duality between optimal balancing and attention. In *ICML*, 2024.

A PROOFS

A.1 PROOFS OF INTERVENTION PRETEXT TASK

In this section, we provide the theoretical guarantee for the effectiveness of our proposed Intervention Pretext Task. We formally prove that minimizing the intervention detection loss is equivalent to estimating the true conditional causal mechanism, thereby enabling the model to identify the true parents in the causal graph.

Proof. The Intervention Pretext Task is formulated as a binary classification problem. Let $y = 1$ denote samples from the observational distribution P_{obs}, and $y = 0$ denote samples from the interventional distribution P_{int}. The objective is to minimize the binary cross-entropy loss $\mathcal{L}_{do}$:

$$\mathcal{L}_{do} = -\mathbb{E}_{\mathbf{x}\sim P_{\text{obs}}}[\log D(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim P_{\text{int}}}[\log(1 - D(\mathbf{x}))], \tag{15}$$

where $D(\mathbf{x})$ represents the probability that $\mathbf{x}$ is from the observational distribution.

According to the classical result in density ratio estimation (Goodfellow et al., 2014), for fixed distributions P_{obs} and P_{int}, the optimal discriminator $D^*(\mathbf{x})$ that minimizes this loss is given by:

$$D^*(\mathbf{x}) = \frac{P_{\text{obs}}(\mathbf{x})}{P_{\text{obs}}(\mathbf{x}) + P_{\text{int}}(\mathbf{x})}. \tag{16}$$

Now, consider the factorization of the joint probability based on the causal graph G. Let the intervention be performed on a specific variable (in window) x_i at time t, replacing it with a value drawn from $Q(x_i)$. The observational distribution factors according to the causal Markov condition:

$$P_{\text{obs}}(\mathbf{x}) = P(x_i \mid Pa(x_i)) \cdot \prod_{j\neq i} P(x_j \mid Pa(x_j)). \tag{17}$$

The interventional distribution, where the causal mechanism of x_i is replaced by $Q(x_i)$ while other mechanisms remain invariant, factors as:

$$P_{\text{int}}(\mathbf{x}) = Q(x_i) \cdot \prod_{j\neq i} P(x_j \mid Pa(x_j)). \tag{18}$$

Substituting these factorizations into Eq. (16), the terms for all non-intervened variables $\prod_{j\neq i} P(x_j \mid Pa(x_j))$ are common to both the numerator and the denominator and thus cancel out. The optimal discriminator simplifies to:

$$D^*(\mathbf{x}) = \frac{P(x_i \mid Pa(x_i))}{P(x_i \mid Pa(x_i)) + Q(x_i)} = \sigma\left(\log \frac{P(x_i \mid Pa(x_i))}{Q(x_i)}\right), \tag{19}$$

where $\sigma(\cdot)$ is the sigmoid function.

Since $Q(x_i)$ is a known proposal distribution (e.g., mixture Gaussian distribution), and $\sigma(\cdot)$ is a strictly monotonic function, the output of the optimal discriminator $D^*(\mathbf{x})$ is determined solely by the true conditional probability $P(x_i \mid Pa(x_i))$.

Therefore, to minimize the loss $\mathcal{L}_{do}$, the model D is forced to approximate D^*. This implies that the model must accurately estimate the conditional likelihood $P(x_i \mid Pa(x_i))$. An accurate estimation of this term is only possible if the model correctly identifies the true parent set $Pa(x_i)$ and models the correct functional relationship. Relying on spurious correlations would yield a suboptimal estimate of the conditional density, resulting in a higher loss. This proves that the intervention pretext task provides a sufficient signal for recovering the causal parents and discovering the underlying causal relationships. □

A.2 PROOFS OF TIME SERIES CAUSAL MIXUP

Let $x_t \in \mathcal{X} \subseteq \mathbb{R}^d$ denote the observed system state. The dynamics are governed by a structural equation $x_{t+1} = f^*(x_t) + \epsilon$, where f^* is the ground-truth mechanism. We focus on estimating the **Causal Probability Matrix**, defined as the Jacobian $G(x) := J_{f^*}(x) = \frac{\partial f^*}{\partial x} \in \mathbb{R}^{d\times d}$.

Standard mixup operates in the input space $\mathcal{X}$. However, physical observations usually lie on a low-dimensional manifold $\mathcal{M} \subset \mathcal{X}$ embedded in the high-dimensional ambient space. A convex combination $\lambda x_1 + (1-\lambda)x_2$ often falls off the manifold $\mathcal{M}$, leading to physically invalid states where the causal mechanism f^* is undefined.

To resolve the geometric inconsistency, we leverage the concept of causal representation learning.

Definition A.1 (Latent Causal Embedding)**.** We assume there exists a latent space $\mathcal{S} \subseteq \mathbb{R}^k$ and an encoder function $\phi : \mathcal{X} \to \mathcal{S}$. While ϕ is implemented as a neural network, we assume it acts as a smooth embedding on the support of the data manifold $\mathcal{M}$. Specifically, $\phi|_{\mathcal{M}}$ is injective and preserves the topological structure of the causal mechanisms.

This definition aligns with the *Manifold Hypothesis*, where deep networks learn to unfold complex curved manifolds into flatter representations Bengio et al. (2013).

Assumption A.2 (Linearity of Mechanisms in Disentangled Space)**.** Following the principle of Independent Causal Mechanisms (ICM) Schölkopf et al. (2021), we assume that in the ideal disentangled latent space $\mathcal{S}$, the dynamics $g(s)$ (where $s = \phi(x)$) are governed by simpler, smoother mechanisms compared to the ambient space. Locally, the evolution in $\mathcal{S}$ can be well-approximated by linear transitions.

Geodesic Consistency via Causal Mixup. We propose performing mixup in $\mathcal{S}$. Let $s_1 = \phi(x_1)$ and $s_2 = \phi(x_2)$. The mixed state is $s^{mix} = \lambda z_1 + (1-\lambda)z_2$.

1. Geodesic Interpretation. Since ϕ acts as a smooth embedding (Definition A.1), the image of the data manifold $\phi(\mathcal{M})$ in $\mathcal{S}$ is locally homeomorphic to Euclidean space. A straight line segment in $\mathcal{S}$ defined by $\gamma_s(\lambda) = \lambda s_1 + (1-\lambda)s_2$ corresponds to a curve $\gamma_x(\lambda) = \phi^\dagger(\gamma_z(\lambda))$ in the original space $\mathcal{X}$, where $\phi^\dagger$ denotes the pseudo-inverse or decoder mapping. Because ϕ unfolds the manifold, the curve $\gamma_x(\lambda)$ approximates the **geodesic** (shortest path) between x_1 and x_2 constrained to lie on $\mathcal{M}$. Unlike input mixup, which creates off-manifold samples, Causal Mixup generates valid physical states along this geodesic.

2. Jacobian Dynamics and Chain Rule. Let the dynamics in the latent space be described by $s_{t+1} = g(s_t)$. The relationship between the observed causal probability $G(x)$ and the latent mechanism is given by the chain rule:

$$G(x) = \frac{\partial f^*}{\partial x} = J_{\phi^\dagger}(s) \cdot \frac{\partial g}{\partial s} \cdot J_\phi(x) \tag{20}$$

where J_ϕ and $J_{\phi^\dagger}$ are the Jacobians of the encoder and decoder, respectively.

3. Regularization via Mixup and Jacobian Smoothing. Causal Mixup optimizes the model by minimizing the mixup loss between the network's predicted causal probability and the linear interpolation target. The empirical risk objective can be formally defined as:

$$\mathcal{L}_{mixup} = \mathbb{E}_\lambda \left[\left\| \hat{G}(s^{mix}) - (\lambda G(x_1) + (1-\lambda)G(x_2)) \right\|_F^2 \right] \tag{21}$$

where $s^{mix} = s_2 + \lambda(s_1 - s_2)$ is the mixed latent state, $\hat{G}(\cdot)$ is the neural network predictor, and $G(\cdot)$ represents the target causal mechanism.

To rigorously quantify the implicit regularization effect of this objective, we employ Taylor's Theorem with the Lagrange remainder. By defining the perturbation direction as $\Delta s = s_1 - s_2$, the mixed state can be rewritten as $s^{mix} = s_2 + \lambda \Delta s$. Applying a first-order Taylor expansion to the network output yields:

$$\hat{G}(s_2 + \lambda \Delta s) = \hat{G}(s_2) + \lambda \nabla_s \hat{G}(s_2) \Delta s + R_2 \tag{22}$$

where R_2 represents the Lagrange remainder. Minimizing the objective $\mathcal{L}_{mixup}$ forces the Taylor expansion of the network output to match the linear interpolation of the targets. This optimization process yields an implicit gradient penalty proportional to the Frobenius norm of the Jacobian, denoted as $\mathcal{O}(\|\nabla_s \hat{G}(s)\|_F^2)$, with the approximation error tightly controlled by the spectral norm of the Hessian.

Formally, if the mixup loss is bounded by a small value ϵ, the Hessian norm is bounded by $\mathcal{O}(\epsilon)$. This implies that the Jacobian $\nabla_s \hat{G}(s)$ is locally constant up to an $\mathcal{O}(\epsilon)$ variation along the interpolation

path. Consequently, minimizing the mixup loss inherently bounds the local Lipschitz constant of the causal mapping $\hat{G}$. Under Assumption A.2, this mathematical equivalence ensures the global smoothness of the causal field and effectively prevents structural instability in the inferred causal graphs.

4. Handling ReLU Networks (Piecewise Linearity). While we assumed smoothness, modern neural networks (MLPs with ReLU) are piecewise linear functions. The input space is partitioned into polytopes where the network is an affine transformation. Within each polytope, ϕ is strictly linear, and its Jacobian J_ϕ is constant. Causal Mixup interpolates across these regions. By enforcing the linearity of the output G with respect to the convex combination in $\mathcal{S}$, we effectively regularize the transitions across the boundaries of these linear regions, ensuring that the global causal field remains continuous and smooth.

Causal Mixup acts as a regularizer that enforces the *Parallel Transport* of the causal mechanism along the data manifold. This minimizes the curvature of the learned causal field, leading to better generalization and structural identifiability compared to input-space interpolation. Therefore, time series causal Mixup not only enables a valid mixup process but also boosts the model's generalization performance.

B Experiments

B.1 Details of Datasets

Synthetic Datasets. Referring the data generation process of RHINO (Gong et al., 2023), we generate diverse time series and causal graphs. For the synthetic data pipeline, we construct the synthetic datasets using a four-step simulation process designed to capture diverse causal mechanisms. The procedure is defined as follows:

1. **Topology Generation:** We generate random Erdős–Rényi (ER) or Scale-Free (SF) graphs (Erdős & Rényi, 1959; Barabási & Albert, 1999) to specify the underlying lagged causal relationships between variables.
2. **Exogenous Noise Sampling:** For each time series, the exogenous variables were drawn from random Gaussian Mixture Models (GMMs) to introduce non-Gaussian complexity.
3. **Temporal Simulation:** We sample initial starting conditions and simulate the temporal progression of the system following Eq. 1 with additive noise. The functional relationships governing the dynamics were modeled using MLPs with two hidden layers (64 units each) and ReLU activation functions.
4. **Stabilization:** A burn-in period was simulated and subsequently discarded to ensure the time series reached a stable, stationary regime.

For dataset specifications, we generate datasets across multiple axes of variation. The specific configurations are detailed below:

- **System Size and Topology:** The number of nodes varies from 5 to 40. Both ER and SF graph structures are utilized.
- **Sequence Length:** Each generated time series has a length of $T = 100$ steps, obtained after removing an initial burn-in period of 100 steps.
- **Training Dataset:** The training corpus is designed to cover a wide range of structural variations. We generate 1,000 distinct causal graphs for each node count. For each causal graph, we simulate 5 independent multivariate time series. This yields a total dataset size of 180,000 time series.
- **Validation and Test Datasets:** The validation and test datasets each contain 370 time series, along with their corresponding ground-truth causal graphs for evaluation. Notably, these causal graphs are strictly disjoint from the training dataset.
- **Interventional Data:** To facilitate the learning of the intervention pretext task, we generate interventional samples for the training dataset. Specifically, we first sample a start time t_1 and an end time t_2. We then randomly resample the specific treatment variables at time

t_1. Conditioned on this intervention, we iteratively generate the sequence over the interval $(t_1, t_2]$. Finally, we assign a binary label to each window indicating whether it contains an intervention.

Real-world Benchmark. CausalRiver (Stein et al., 2025) treats river flow measurements from monitoring stations as a time series dataset with inherent causal structure. It covers two regions in Germany: the eastern German territory (666 measuring stations) and Bavaria (494 stations). The dataset spans from 2019 to 2023 with a temporal resolution of 15 minutes. Importantly, CausalRiver includes both normal hydrological conditions and extreme events such as heavy rainfall and large-scale precipitation, enabling the study of causal dynamics under diverse environmental scenarios. The dataset is categorized into five types based on the characteristics of their underlying causal structures:

- Random: All connected subgraphs with three or five nodes, covering the entire dataset and full diversity of benchmark conditions.
- Close: Connected subgraphs whose edges have a maximum geographic (Euclidean) distance of 0.3; by excluding long-range connections, causal effects are expected to be more pronounced. This set is fully contained within Random.
- Random + 1: Connected subgraphs with two or four nodes, combined with one additional isolated node. To avoid confounding, the isolated nodes are drawn from coastal or border regions where disconnected nodes naturally occur.
- Root cause: Connected subgraphs with three or five nodes in which each node has at most one parent, forming chain-like structures. This setting is useful for root-cause analysis (Ikram et al., 2022) and is fully contained within Random.
- Confounder: Subgraphs with four or six nodes containing a single node with multiple children (rarely observed in cases such as river splits). The multi-child node is then removed to simulate permanent hidden confounding.

Root Cause Identification Datasets. SWaT (Mathur & Tippenhauer, 2016) is a dataset collected from a testbed that simulates a real-world water treatment plant. It comprises data from 51 sensors in the critical infrastructure system during continuous operation, including both normal operating conditions and attack scenarios within the water treatment process. MSDS (Multi-Source Distributed System) (Nedelkoski et al., 2020) is developed on an OpenStack testbed and serves as a dataset for AIOps (Artificial Intelligence for IT Operations). Instances of fault injections in this system are labeled as anomalies. More detail are provided in Table 5.

Table 5: Details of Root Cause Identification Datasets.

Dataset	Training Time Steps	Test Sequences	Avg. Sequence Length	Avg. # of Root Causes
SWaT (51)	49,500	20	51	13.35
MSDS (10)	29,268	4,255	21	3.05

B.2 More Details of Baselines

We provide detailed descriptions of the baseline methods used for comparison in both Causal Discovery and Root Cause Identification tasks.

Causal Discovery Baselines. We evaluate representative methods spanning classical statistical approaches and modern deep learning paradigms.

- **PCMCI** (Runge et al., 2019): A constraint-based algorithm optimized for time-series data. It first selects preliminary parents using PC-algorithm-based selection and then estimates the causal graph using momentary conditional independence tests to control the false positive rate. It effectively handles time lags and strong autocorrelation.
- **VarLiNGAM** (Hyvärinen et al., 2010): A linear non-Gaussian acyclic model extended for autoregressive processes. It assumes the data generating process is linear with non-Gaussian noise and utilizes independent component analysis to identify the causal structure and time-lagged causal effects.

- **DYNOTEARS** (Pamfil et al., 2020): A score-based method that formulates structural learning as a continuous optimization problem. It extends the NOTEARS framework to dynamic causal networks by minimizing a penalized loss function subject to acyclicity constraints, allowing for the simultaneous learning of intra-slice and inter-slice dependencies.
- **VAR** (Assaad et al., 2023): A classic statistical model that captures the linear interdependencies among multiple time series. In the context of causal discovery, Granger Causality is inferred based on the non-zero coefficients of the fitted autoregressive matrices.
- **CDMI** (Ahmad et al., 2024): A method based on Causal Discovery with Mutual Information. It estimates the causal probability between time series variables by quantifying the mutual information shared between past values of potential parents and current values of the target, often utilizing non-parametric estimators to capture non-linear dependencies.
- **TCDF** (Nauta et al., 2019): A deep learning framework that uses attention-based CNNs. It learns a predictive model for each time series and interprets the internal attention parameters and variable importance scores to discover causal links and time delays.
- **CUTS+** (Cheng et al., 2024): An improved version of the CUTS framework designed for non-stationary time series. It employs a latent causal discovery approach where a Granger-causality-inspired graph is learned jointly with a data augmentation strategy to handle distribution shifts and improve robustness.
- **Causal Pretraining (CP)** (Stein et al., 2024): A self-supervised pretraining framework designed to learn causal representations from time series. It reconstructs masked sub-series to capture temporal dependencies. To evaluate the impact of the backbone architecture, we compare two variants:
  - **CP-Trans**: Implements the standard CP framework using a Transformer-based backbone to capture long-range dependencies via self-attention mechanisms.
  - **CP-GRU**: Implements the CP framework using a gated recurrent unit backbone, focusing on sequential recurrent modeling.

Root Cause Identification Baselines. To evaluate the effectiveness of PTCD in identifying root causes of anomalies, we compare it with four established methods:

- ϵ-**Diagnosis** (Shan et al., 2019): A statistical approach that locates root causes by analyzing the correlation changes between variables. It performs pairwise significance tests (e.g., standard deviation and correlation shifts) between normal and abnormal periods to construct an anomaly propagation graph and rank potential root causes.
- **RCD** (Ikram et al., 2022): A framework that integrates causal structure learning with anomaly attribution. It learns a partial causal graph from data and analyzes the invariance of conditional distributions. Nodes corresponding to intervention targets are identified as root causes.
- **CIRCA** (Li et al., 2022): A method that leverages domain knowledge to construct structural causal graphs. It identifies root causes by modeling the residuals of causal mechanisms; specifically, it detects nodes that exhibit significant distribution shifts in their conditional probabilities during the anomaly window.
- **AERCA** (Han et al., 2025): A reconstruction-based method that models causal dependencies and exogenous variables using an Autoencoder framework. It posits that anomalies are generated by perturbations in exogenous factors. AERCA attributes root causes by measuring the magnitude of the reconstruction error associated with the exogenous noise terms of each variable.

B.3 More Details of Metrics on Root Cause Identification

Given a multivariate time series $\mathcal{X}$, the $Recall@K$ is defined as:

$$Recall@K = \frac{1}{|\mathcal{X}|} \sum_{x_i \in \mathcal{X}} \frac{\left| V_{x_i}^{(RC)} \cap \{R_{x_i}[k] \mid k = 1, 2, \dots K\} \right|}{\min\left(K, \left| V_{x_i}^{(RC)} \right|\right)}, \tag{23}$$

where $R_{x_i}[k]$ indicates the time series at the k-th rank for the channel x_i, and $V_{x_i}^{(RC)}$ indicates a set of root cause variables over the whole channel x_i. Note that if a time series receives multiple exogenous interventions, it only counts as one root cause time series in $V_{x_i}^{(RC)}$. We further compute the overall performance by computing the average $Recall@K$, denoted as $Avg@K = \frac{1}{K}\sum_{k=1}^{K} Recall@k$.

B.4 Implementation Details

We first pretrain PTCD using the Adam optimizer (Kingma & Ba, 2015) with a learning rate of 10^{-3}. We apply early stopping with a patience of 30 epochs to prevent overfitting. The loss weights λ_G, λ_{KL} and λ_{do} are set to 0.6. The model consists of 4 alternating attention blocks, and the number of Gaussian components K is set to 10. For fine-tuning in both the causal discovery task and the root cause identification task, we use a learning rate of 10^{-4}, and keep λ_{KL} fixed at 0.6. All experiments are implemented in PyTorch and conducted on an NVIDIA A800 80GB GPU. Following AERCA (Han et al., 2025), data preprocessing is standardized across datasets using a MinMax scaler. To improve computational efficiency, we downsample the SWaT dataset every 10 seconds and the MSDS dataset every 5 time steps. The code and datasets are available at https://anonymous.4open.science/r/PTCD-D4E3.

B.5 More Ablation Study

To investigate the impact of different components on root cause identification, we conduct an ablation study focusing on inter-window attention, intra-window attention, mixture-of-Gaussians modeling of exogenous variables, causal mixup, and intervention. As shown in Table 6, we find that the intervention mechanism achieves superior performance in identifying root causes. This gain is attributed to the fact that interventions provide the model with information across diverse environments, enabling it to filter out spurious correlations, learn more accurate causal relationships, and enhance overall robustness. Furthermore, when the mixture-of-Gaussians constraint on exogenous variables is removed, PTCD exhibits a notable drop in performance, demonstrating the importance of instance-specific exogenous variable estimation for effective root cause identification.

Table 6: Ablation studies on Root Cause Localization. w/o inter, w/o intra, w/o exogenous, w/o causal mixup, w/o intervention represent removing the inter-window attention, intra-window attention, exogenous, causal mixup, and intervention, respectively.

Dataset	Model	Recall@1	Recall@3	Recall@5	Recall@10	Avg@10
MSDS	w/o inter	0.452	0.792	0.910	**1.000**	0.492
	w/o intra	0.498	0.801	0.910	**1.000**	0.835
	w/o exogenous	0.227	0.697	0.893	**1.000**	0.694
	w/o causal mixup	0.500	0.792	0.993	**1.000**	0.842
	w/o intervention	0.296	0.798	0.900	**1.000**	0.728
	PTCD	**0.515**	**0.993**	**0.996**	1.000	**0.929**
SWaT	w/o inter	0.200	0.350	0.350	0.455	0.351
	w/o intra	0.250	0.375	0.375	0.450	0.360
	w/o exogenous	0.100	0.150	0.250	0.350	0.205
	w/o causal mixup	0.250	0.330	0.450	0.455	0.342
	w/o intervention	0.150	0.220	0.330	0.375	0.252
	PTCD	**0.300**	**0.450**	**0.450**	**0.475**	**0.440**

B.6 Efficiency Analysis

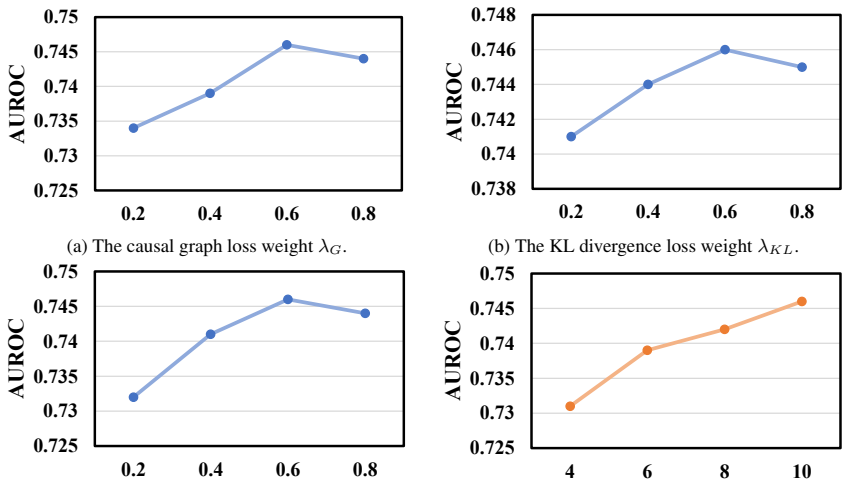
Table 7: Efficiency comparison across different paradigms on the causalriver dataset.

Model	Paradigm	Params (K)	Inference Time	GPU Mem (MB)	CPU Mem (MB)	**avg. AUROC**
TCDF	End-to-End Causal	0.12	17.19 s	2.00	460.45	0.57
CUTS+	End-to-End Causal	8.23	14.19 s	2.00	293.36	0.56
PTCD (zero-shot)	Pre-trained Causal	877.24	0.01 s	1236.00	504.08	0.74
PTCD (fine-tune)	Pre-trained Causal	877.24	3.59 s	1236.00	504.08	**0.85**

To provide a comprehensive efficiency profile, we evaluate the parameters, inference time, and memory usage across different paradigms. As shown in Table 7, although PTCD introduces a moderate parameter overhead due to pre-training, its fine-tuned variant achieves the best overall performance, at the expense of increased inference time compared to the zero-shot setting. This demonstrates a favorable trade-off: zero-shot PTCD offers exceptional efficiency, while the fine-tuned variant delivers superior predictive accuracy, all while retaining rapid inference capabilities.

B.7 Parameter Sensitivity Analysis

We conduct a hyperparameter sensitivity analysis with respect to the loss weights λ_G, λ_{KL}, and λ_{do}, which correspond to the causal graph loss, the KL-divergence loss, and the intervened loss, respectively, as well as the number of Gaussian components K. All experiments are performed on synthetic datasets. As shown in Figure 4, PTCD achieves optimal causal discovery performance when λ_G, λ_{KL}, and λ_{do} are set to 0.6. This indicates that causal graph prediction, the modeling of exogenous variables, and intervention mechanisms play a significant role in the generalization capability of the learned causal discovery. Additionally, the number of Gaussian components also influences the model's generalization performance to a certain degree.



(a) The causal graph loss weight λ_G.

(b) The KL divergence loss weight λ_{KL}.

(c) The intervention loss weight λ_{do}.

(d) The number of Gaussian components K.

Figure 4: The effect of parameter λ_G, λ_{KL}, λ_{do}, and K on synthetic datasets for time-series causal discovery.

C Limitations

While PTCD demonstrates strong generalization capabilities in time series causal discovery and root cause identification, it has a few limitations regarding its structural assumptions. The SCM assumes an Additive Noise Model (ANM). Although ANM is theoretically advantageous for guaranteeing causal identifiability and cleanly decoupling endogenous dynamics from exogenous anomalies for precise root cause localization, it may not capture complex, non-additive external disturbances (e.g., multiplicative noise or nonlinear entanglement) present in reality. Future work will explore flexible non-additive causal mechanisms to further enhance robustness against extreme out-of-distribution (OOD) noise.