

Table 1: Performance on TSCD task on CausalTime datasets. We utilize the performance of baseline TSCD algorithms reported in CausalTime paper (Cheng et al., 2024). **Bold text** means the best model and underlined text indicates second-best model

Methods	AUROC			AUPRC		
	AQI	Traffic	Medical	AQI	Traffic	Medical
GC	0.4538 $\pm 0.0377$	0.4191 $\pm 0.0310$	0.5737 $\pm 0.0338$	0.6347 $\pm 0.0158$	0.2789 $\pm 0.0018$	0.4213 $\pm 0.0281$
SVAR	0.6225 $\pm 0.0406$	<u>0.6329 <math>\pm 0.0047</math></u>	0.7130 $\pm 0.0188$	0.7903 $\pm 0.0175$	0.5845 $\pm 0.0021$	0.6774 $\pm 0.0358$
N.NTS	0.5729 $\pm 0.0229$	<b>0.6329 <math>\pm 0.0335</math></b>	0.5019 $\pm 0.0682$	0.7100 $\pm 0.0228$	0.5770 $\pm 0.0542$	0.4567 $\pm 0.0162$
PCMCI	0.5272 $\pm 0.0744$	0.5422 $\pm 0.0737$	0.6991 $\pm 0.0111$	0.6734 $\pm 0.0372$	0.3474 $\pm 0.0581$	0.5082 $\pm 0.0177$
Rhino	0.6700 $\pm 0.0983$	0.6274 $\pm 0.0185$	0.6520 $\pm 0.0212$	0.7593 $\pm 0.0755$	0.3772 $\pm 0.0093$	0.4897 $\pm 0.0321$
CUTS	0.6013 $\pm 0.0038$	0.6238 $\pm 0.0179$	0.3739 $\pm 0.0297$	0.5096 $\pm 0.0362$	0.1525 $\pm 0.0226$	0.1537 $\pm 0.0039$
CUTS+	<b>0.8928 <math>\pm 0.0213</math></b>	0.6175 $\pm 0.0752$	<b>0.8202 <math>\pm 0.0173</math></b>	<u>0.7983 <math>\pm 0.0875</math></u>	<u>0.6367 <math>\pm 0.1197</math></u>	0.5481 $\pm 0.1349$
NGC	0.7172 $\pm 0.0076$	0.6032 $\pm 0.0056$	0.5744 $\pm 0.0096$	0.7177 $\pm 0.0069$	0.3583 $\pm 0.0495$	0.4637 $\pm 0.0121$
NGM	0.6728 $\pm 0.0164$	0.4660 $\pm 0.0144$	0.5551 $\pm 0.0154$	0.4786 $\pm 0.0196$	0.2826 $\pm 0.0098$	0.4697 $\pm 0.0166$
LCCM	0.8565 $\pm 0.0653$	0.5545 $\pm 0.0254$	0.8013 $\pm 0.0218$	<b>0.9260 <math>\pm 0.0246</math></b>	0.5907 $\pm 0.0475$	<b>0.7554 <math>\pm 0.0235</math></b>
eSRU	0.8229 $\pm 0.0317$	0.5987 $\pm 0.0192$	0.7559 $\pm 0.0365$	0.7223 $\pm 0.0317$	0.4886 $\pm 0.0338$	0.7352 $\pm 0.0600$
SCGL	0.4915 $\pm 0.0476$	0.5927 $\pm 0.0553$	0.5019 $\pm 0.0224$	0.3584 $\pm 0.0281$	0.4544 $\pm 0.0315$	0.4833 $\pm 0.0185$
TCDF	0.4148 $\pm 0.0207$	0.5029 $\pm 0.0041$	0.6329 $\pm 0.0384$	0.6527 $\pm 0.0087$	0.3637 $\pm 0.0048$	0.5544 $\pm 0.0313$
CALAS	<u>0.8772 <math>\pm 0.0287</math></u>	0.6312 $\pm 0.0461$	<u>0.8124 <math>\pm 0.0125</math></u>	0.6788 $\pm 0.0512$	<b>0.6701 <math>\pm 0.0980</math></b>	<u>0.7412 <math>\pm 0.0518</math></u>

## A ADDITIONAL EXPERIMENTAL RESULTS

To prove that CALAS actually finds the ground truth causality, we conduct experiments with three real-world datasets in CausalTime benchmark (Cheng et al., 2024) and one well-known Synthetic dataset for causal discovery (Suiz A. Baccalá, 2001). We quantitatively and qualitatively showcases CALAS’s superiority on causal discovery with Air-quality (AQI), Traffic, and Medical datasets, experimentally proving that CALAS can actually model the causal relationship. We compared CALAS with various baselines including: Granger causality (GC) (Granger, 1969), neural Granger causality (NGC) (Tank et al., 2022), economy-SRU (eSRU) (Khanna & Tan, 2020), scalable causal graph learning (SCGL) (Xu et al., 2019), temporal causal discovery framework (TCDF) (Nauta et al., 2019), CUTS (Cheng et al., 2023b), CUTS+ (Cheng et al., 2023a), PCMCI (Runge et al., 2019), SVAR, NTS-NOTEARS (shown as N.NTS) (Sun et al., 2023), Rhino (Gong et al., 2023), latent convergent cross mapping (LCCM) (Brouwer et al., 2021), and neural graphical model (NGM) (Bellot et al., 2022). For the synthetic dataset, we only provide visual comparison among Granger causality test, LIFT (i.e., cross-correlation), and CALAS. In the causal discovery experiments, we stick to the our MTS forecasting setting with input length 336 but with output length 1. As a backbone, we utilize one layer linear model.

### A.1 QUANTITATIVE EVALUATION ON CAUSAL DISCOVERY

Table 1 indicates the experimental results for the time series causal discovery (TSCD) task with three real-world datasets. Even though CALAS focuses on the dynamic causality discovery, it exhibits competitive results across all three datasets in traditional TSCD task, achieving the best performance in AUPRC for the Traffic dataset, and competitive performance to the state-of-the-art models, such as CUTS+, in AUROC and AUPRC across Medical and AQI datasets. It experimentally proves that CALAS successfully models the causal relationships during its optimization. Furthermore, CALAS is the one of two algorithms that simultaneously models the propagation delay and causal strength, however, the other one (i.e., TCDF) indicates its limitation to properly model both characteristics. Lastly, despite CALAS has only one the simple hyperparameter, the maximum delay  $k$ , for the causal modeling, it outperforms the algorithms requiring sophisticated, data-specific hyperparameter settings. It further reduces the difficulty to introduce the algorithm to unseen datasets. However, as we can depict in Figure 1, CALAS refers to improper cause signals, which could be improved by introducing contrastive learning methods or regularization term in optimization.

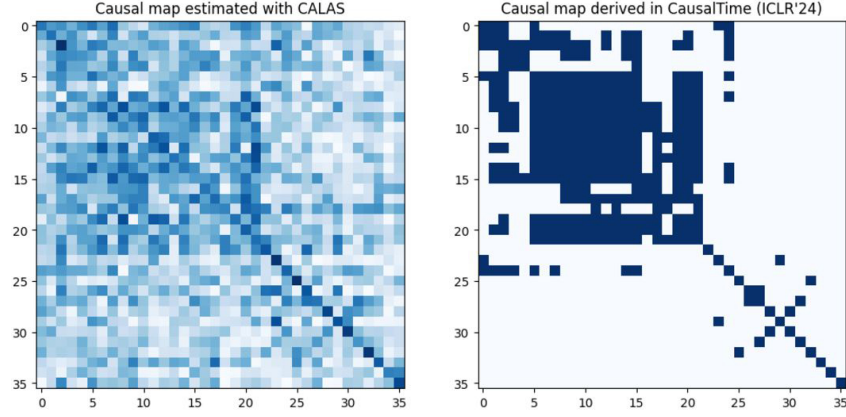


Figure 1: Causality map estimated via CALAS (left) and ground truth causal graph (right).

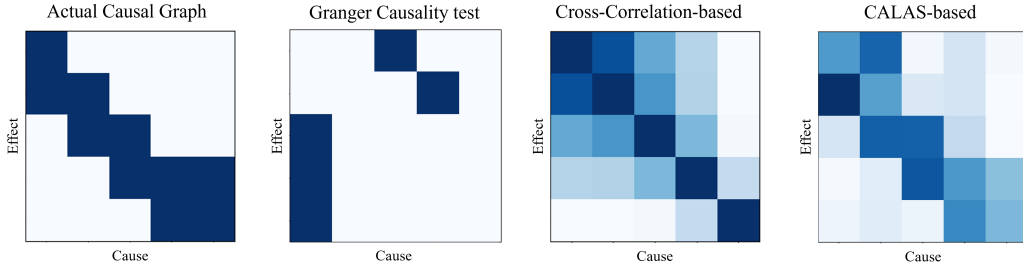


Figure 2: Visualization of actual causal graph and those calculated with Granger causality test, cross-correlation, and CALAS without any hyperparameter tuning or additional techniques like weight decay or  $L_2$  normalization.

## A.2 QUALITATIVE EVALUATION ON CAUSAL DISCOVERY

To prove that CALAS actually finds the ground truth causality, we conduct experiments with a well-known Synthetic dataset for causal discovery (Suiz A. Baccalá, 2001). We have compared CALAS with Granger Causality test and LIFT (i.e., cross-correlation). We have excluded other CD modeling methods in MTS forecasting, because 1) they are unable to model propagation delay, which is unsuitable for causal discovery, as we depicted in main paper. Furthermore, to compare the performance in nature, that means, end-to-end manner without *any hyperparameter tuning*, we have excluded deep learning-based causal discovery methods, such as TCDF (Nauta et al., 2019), CUTS (Cheng et al., 2023b), cLSTM (Tank et al., 2022), or other methods. Please note that causal discovery models require sophisticate hyperparameter tuning to obtain a proper causality graph (Nauta et al., 2019; Li et al., 2023).

**Main results** Figure 2 indicates the actual causal graph and those calculated with each methods. It indicates that CALAS can successfully approximate actual causality graph without any hyperparameter tuning related to optimization or model parameter. It indicates both validity and importance of learning propagation delay, which has not yet been investigated in both causal discovery and MTS forecasting. It also provides another lesson-learned to the causal discovery domain that we do not actually need to data-specifically and manually conduct hyperparameter search and only need to optimize both propagation delay and causal strength. Please note that this phenomenon is also reported as one of the challenges in CD modeling for MTS forecasting—model tend to encounter overfitting issue without delay estimation or dynamic CD modeling (Han et al., 2023).

Table 2: Performance comparison in terms of forecasting errors. The **bold text** means the best results and underlined text means the second best results.

Method		TimesNet		CALAS +Linear		DLinear	
		MSE	MAE	MSE	MAE	MSE	MAE
Weather	12.5%	<b>0.025</b>	<b>0.045</b>	<b>0.025</b>	<b>0.045</b>	0.039	0.101
	25.0%	<b>0.029</b>	<u>0.052</u>	<u>0.030</u>	<b>0.051</b>	0.048	0.111
	37.5%	<b>0.031</b>	<b>0.057</b>	<u>0.033</u>	<u>0.065</u>	0.057	0.121
	50.0%	<b>0.034</b>	<b>0.062</b>	<u>0.040</u>	<u>0.072</u>	0.066	0.134
Electricity	12.5%	<u>0.085</u>	<u>0.202</u>	<b>0.063</b>	<b>0.170</b>	0.092	0.214
	25.0%	<u>0.089</u>	<u>0.206</u>	<b>0.077</b>	<b>0.190</b>	0.118	0.247
	37.5%	<u>0.094</u>	<u>0.213</u>	<b>0.093</b>	<b>0.206</b>	0.144	0.276
	50.0%	<b>0.100</b>	<b>0.221</b>	<u>0.106</u>	<u>0.230</u>	0.175	0.284

### A.2.1 DATA GENERATION

For the training, we utilize Synthetic dataset generated with following equations:

$$\begin{cases} x_1(n) = 0.95\sqrt{2}x_1(n-1) - 0.9025x_1(n-2) + w_1(n) \\ x_2(n) = -0.5x_1(n-1) + w_2(n) \\ x_3(n) = 0.4x_2(n-2) + w_3(n) \\ x_4(n) = -0.5x_3(n-1) + 0.25\sqrt{2}x_4(n-1) + 0.25\sqrt{2}x_5(n-1) + w_4(n) \\ x_5(n) = -0.25\sqrt{2}x_4(n-1) + 0.25\sqrt{2}x_5(n-1) + w_5(n) \end{cases}, \quad (5)$$

where  $n$  is  $n$ -th time steps,  $x_i$  means  $i$ -th variate, and  $w_i(n)$  means zero-mean uncorrelated white processes with identical variances.

### A.3 EXPERIMENTAL RESULTS FOR IMPUTATION TASK

## B DISCUSSION

**Discussion on other CNN- or RNN-based methods** The emergence of Mamba (Gu & Dao, 2024) played a significant role in shifting researchers' focus from Transformer models to State Space Models (SSMs). One main flow in the SSM research is the extension of 1D Mamba to the multidimensional state space models. In the MTS forecasting, there are RNN-based (Tank et al., 2022; Jia et al., 2023; Behrouz et al., 2024), CNN-based (Wu et al., 2023), and SSM-based approaches (Zeng et al., 2024; Hu et al., 2024). In the following paragraphs, we will discuss how they can achieve inter-variate dependency modeling and difference between CALAS and them.

*RNN-based approaches*, for example, WITRAN (Jia et al., 2023) or cRNN (Tank et al., 2022), need to propose additional channel dependency module to model the inter-variables relationships. In case of WITRAN, similar to the TimesNet (Wu et al., 2023), it folds the input 1D time series into 2D time series, enabling intra- and inter-periodicity modeling. However, it less focuses on the potential causal effects from the other time series. cRNN (Tank et al., 2022), one of the causal discovery algorithms, induces the inter-variables relationships via weights of trained RNN (i.e., projection layer of RNNs), however, it only captures the gradual changes, especially with the one-step time lag.

For the *CNN- and SSM-based approaches*, such as C-Mamba (Zeng et al., 2024) or TimeSSM (Hu et al., 2024), the main difference between CALAS and these approaches are how they dealt with receptive fields and introduces inductive biases into model. C-Mamba and TimeSSM considers the convolutional- or SSM-based state space as the range of information fusion, not introducing inductive biases into. For example, in case of 1D  $\mathbb{R}^{N_i \times N_o \times k}$  CNN kernel, where  $N_i, N_o, k$  are input and output channel and kernel size, respectively, aforementioned model optimize both causal strength and propagation delay into one kernel. In such design, fused with causal strength, propagation delay will be independent weights to each other. However, for the inter-variate relationship, there are only one unique propagation delay, which should be modeled with probability function among  $k$  possible delays. CALAS disentangles the propagation delay and approximate them with Gaussian probability

kernel, it introduces additional inductive bias—given two cause signal  $X$  and effect signal  $Y$ , there exists unique discrete delay  $d_{X,Y}$  such that the time gap between change of  $X$  and its actual influence to  $Y$ .

**Transformer-based methods vs. CALAS** Transformer-based methods, including iTransformer, mix the multivariate information regardless of propagation delay. This design may introduce misaligned or outdated information from lagged time series, resulting to degrade of model performance. To properly align the variates, previous models should borrow the modeling capacity from temporal dependency modeling components, which lowers the temporal dependency modeling quality. By facilitating proposed convolution, CALAS simultaneously conduct such alignment and CD modeling. Though LIFT (Zhao & Shen, 2024) achieves lead-lag relation modeling with cross-correlation, it requires additional computation and relies on the statistical methods that often be suboptimal.

**Layer-agnostic causality modeling is important in MTS forecasting** Distribution shifts of the statistical features are well-studied problem in MTS forecasting (Kim et al., 2022; Zeng et al., 2023; Liu et al., 2022). However, distribution shifts of channel dependency is not yet investigated deeply. Here, we derive a discussion for distribution shifts in short period, as we depicted in Figure 1. By introducing shared CALAS across multiple layers except the input-dependent parts, our model reduces the misalignment or over-reliance of previously generated causal maps when short term distribution shifts occur.

**Generalization for multi-periodicity modeling** Since CALAS stems from convolution mechanism, we can achieve the multi-periodicity decomposition by adjusting the stride. However, addressing this question is beyond the scope of this work, so we leave it for future exploration.