Stable Partial Order Constraints for Temporal Causal Structure Learning

Changxin Rong, Xiangyu Wang*, Taiyu Ban, Yanze Gao, Xingjian Lin, Huanhuan Chen*
School of Computer Science and Technology
University of Science and Technology of China, Hefei, China

Abstract

Temporal causal structure learning aims to recover lag-aware causal relations from multivariate time series. While prior knowledge has been widely used to improve static structure learning, its integration into lag-aware structure learning remains limited, because available priors are often lag-agnostic. Recent studies have explored lag-agnostic edge presence priors, but partial order priors, which can prune the ordering space and improve structure learning, remain underexplored. A key challenge is that variables may influence each other at different lags, leading to cyclic variable-level relations where ordering is no longer naturally defined. Moreover, directly applying static differentiable formulations of partial order constraints can be unstable due to repeated-walk accumulation on cycles. To address this gap, we represent partial order priors on the summary graph and follow the natural idea of the order relation that only allows a single direction of possible causality between two variables, which is equivalent to forbidding all reversed causality, direct or indirect. This converts partial order constraints into path prohibition constraints that can be extended to cyclic summary structures. For stable differentiable characterization, we propose a decoupled partial order constraint method. The method preserves the original lag-aware structure for data fitting, imposes partial order constraints through a normalized summary representation, and captures higher-order reachability with a closed-form connectivity characterization. Experiments further demonstrate its effectiveness in nonstationary settings.

1 Introduction

Temporal causal structure learning aims to recover lag-aware causal relations among variables from multivariate time series [Haufe et al., 2010]. The lag-aware relations capture both instantaneous causal relations within the same time step and cross-lag causal relations from past to the current time step. Recovering such structures is important for understanding dynamic interactions in real-world systems, including biological analysis, medical monitoring, and climate system modeling [Li et al., 2023, Runge et al., 2019]. Recently, this problem has been further formulated as a differentiable optimization task, in which instantaneous and cross-lag structures are represented by structural parameters. The structural parameters are then learned by optimizing data-fitting objectives, with a smooth DAG constraint imposed on instantaneous structure [Pamfil et al., 2020, Sun et al., 2021].

In practical applications, prior knowledge about partial causal relations provides an important way to recover more interpretable and insightful structures when data alone are insufficient [Runge et al., 2019, Assaad et al., 2022]. While prior knowledge has been widely used for static structure learning [Zheng et al., 2018, Ban et al., 2024, 2025, Chen et al., 2025], its integration into lag-aware structure learning remains limited. A central challenge is that available priors are usually lag-agnostic and summary, which do not contain specific lags of causal relationships that lag-aware structure learning

*Corresponding authors: Xiangyu Wang (sa312@ustc.edu.cn) and Huanhuan Chen (hchen@ustc.edu.cn).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

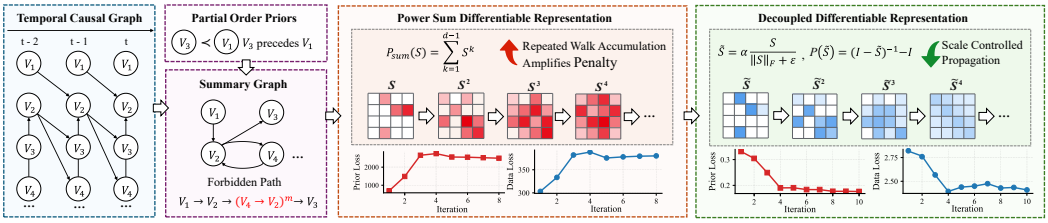

Figure 1: Illustration of temporal partial order integration. Partial order priors are converted into forbidden reachability constraints on the summary graph, while the proposed decoupled normalized method avoids unstable repeated-walk accumulation on cyclic summaries.

requires. Recent work has explored lag-agnostic edge presence priors, which specify whether causal edges exist between variables without requiring exact lag information [Ban et al., 2026].

However, experts often specify only coarse causal ordering among variables rather than exact causal edges. This motivates the use of partial order priors, which restrict admissible causal directions to prune the ordering space and improve structure learning [Niinimäki et al., 2016, Ban et al., 2024]. Nevertheless, applying partial order priors to lag-aware structure learning remains challenging. In a lag-aware context, variables may influence each other at different lags, leading to bidirectional causal relations at the variable level that make the notion of ordering no longer naturally defined [Assaad et al., 2022, Ferreira and Assaad, 2024]. Moreover, directly applying differentiable formulations of partial order constraints in static structure learning, which characterize variable reachability through a matrix power-sum, can introduce optimization instability in lag-aware structure learning [Ban et al., 2024]. When variable-level causal relations contain cycles, this formulation repeatedly accumulates walks along cycles, amplifies the prior penalty, and destabilizes optimization, as shown in Fig. 1.

To address this gap, we first represent partial order priors on the summary graph that compresses a lag-aware structure by recording whether a causal relation exists between variables at any lag. This provides a unified variable-level representation. We follow the natural idea of the order relation that only allows a single direction of possible causality between two variables, which is equivalent to forbidding all reversed causality, direct or indirect. Thus, partial order constraints are converted into a set of path prohibition constraints that can be extended to cyclic summary graph.

For stable differentiable characterization, we propose a decoupled partial order constraint method. The method preserves the original lag-aware structure for data fitting while imposing partial order constraints through a normalized summary representation that bounds the propagation scale and prevents unbounded cyclic accumulation. A closed-form connectivity characterization is further used to capture higher-order reachability without explicit matrix power-sum accumulation. In this way, lag-agnostic partial order priors can provide stable guidance for recovering lag-aware structures.

The proposed method is evaluated across synthetic and real-world benchmarks. The results show that lag-agnostic partial order priors can effectively guide the recovery of lag-aware structures. The successful extension to nonstationary temporal structure learning further highlights the generality and extensibility of the method. The main contributions of this work are summarized as follows:

- This paper formalizes temporal partial order constraints and converts them into path prohibition constraints extendable to cyclic summary structures.
- This work identifies the instability caused by matrix power-sum formulations of path prohibition on cyclic summary graphs and proposes a stable decoupled partial order constraint.
- The proposed method is successfully integrated into nonstationary temporal structure learning frameworks, showing its generality and extensibility.

2 Preliminaries

This section reviews differentiable temporal causal structure learning and partial order priors in the static setting, which provide the basis for the proposed temporal extension.

2.1 Temporal Causal Structure Learning

Let $X = \{x_t\}_{t=1}^{T}$ be a d-dimensional multivariate time series, where $x_t \in \mathbb{R}^d$ denotes the observation at time t. Given a maximum lag order p, temporal causal structure learning aims to recover a lag-aware structure

$$W = \{W^\tau\}_{\tau=0}^{p},$$

where $W^0 \in \mathbb{R}^{d\times d}$ represents instantaneous causal effects among variables at time t, and $W^\tau \in \mathbb{R}^{d\times d}$ for $\tau = 1, \ldots, p$ represents cross-lag causal effects from time $t-\tau$ to time t.

A common temporal structural equation model writes each observation as

$$x_t = x_t W^0 + \sum_{\tau=1}^{p} x_{t-\tau} W^\tau + z_t, \tag{1}$$

where $z_t \in \mathbb{R}^d$ denotes a noise term. This model relates each observation to instantaneous and cross-lag effects, enabling the estimation of W from observational time series data.

In differentiable temporal causal structure learning, W is typically estimated by minimizing a data-fitting objective with sparsity regularization:

$$\min_{W} \; L_{\mathrm{data}}(W) + \lambda_0 \|W^0\|_1 + \sum_{\tau=1}^{p} \lambda_\tau \|W^\tau\|_1 \quad \text{s.t.} \quad h(W^0) = 0, \tag{2}$$

where $L_{\mathrm{data}}(W)$ denotes the data-fitting loss, and λ_0 and λ_τ are regularization coefficients [Pamfil et al., 2020, Sun et al., 2021]. The constraint $h(W^0) = 0$ enforces acyclicity on the instantaneous structure, following the NOTEARS-style [Zheng et al., 2018] smooth characterization

$$h(W^0) = \mathrm{tr}\big(\exp(W^0 \circ W^0)\big) - d, \tag{3}$$

where $\circ$ denotes the Hadamard product. Cross-lag structures do not require an acyclicity constraint, since their edges are temporally ordered by construction.

2.2 Partial Order Priors in Static Settings

Let $V = \{v_1, \ldots, v_d\}$ denote the set of variables. In the static setting, let $G_{\mathrm{stat}}(A)$ be a DAG over V with weighted adjacency matrix $A \in \mathbb{R}^{d\times d}$.

Definition 1 (Partial Order)**.** *For a set of variables V, a partial order is a binary relation $\prec$ on V that satisfies reflexivity, antisymmetry, and transitivity. For variables $v_i, v_j \in V$, the relation $v_i \prec v_j$ indicates that v_i precedes v_j in the topological ordering.*

In causal structure learning, a partial order prior is usually represented by a subset of causal precedence relations $O \subseteq V \times V$. Since partial orders are transitive, the explicitly specified relations may imply additional precedence relations. We denote by O^+ the transitive closure of O, which contains both the relations directly specified in O and those inferred through transitivity.

According to the causal precedence encoded by partial order priors, if $v_i \prec v_j$, then the reverse causal reachability from v_j to v_i should be forbidden. Therefore, this is equivalent to requiring that v_j is not an ancestor of v_i, or that no directed path exists from v_j to v_i [Ban et al., 2024, 2025]. Therefore, $G_{\mathrm{stat}}(A)$ satisfies the partial order prior set O if and only if

$$v_j \not\rightsquigarrow_{G_{\mathrm{stat}}(A)} v_i, \qquad \forall (v_i, v_j) \in O^+,$$

where $u \rightsquigarrow_G v$ denotes the existence of a directed path from u to v in graph G.

To characterize this condition in a differentiable form, a standard reachability representation is constructed by a matrix power-sum of the adjacency matrix:

$$P_{\mathrm{stat}}(A) = \sum_{k=1}^{d-1} |A|^k, \tag{4}$$

where $|A|$ denotes the element-wise absolute value of A. The entry $\big[P_{\mathrm{stat}}(A)\big]_{ji}$ continuously characterizes whether v_j can reach v_i. Then, the partial order penalty is written as

$$L_{\mathrm{stat}}(A; O) = \sum_{(v_i, v_j) \in O^+} \big[P_{\mathrm{stat}}(A)\big]_{ji}. \tag{5}$$

This provides the basis for extending partial order priors to temporal causal structure learning.

3 Method

This section presents the proposed method. We first define partial order constraints for temporal causal structure learning on a variable-level summary graph. We then show that directly representing the resulting path prohibition constraints with a matrix power-sum becomes unstable on cyclic summary graphs. Finally, we introduce a stable decoupled partial order constraint for integrating lag-agnostic partial order knowledge into lag-aware temporal structure learning.

3.1 Partial Order Constraints in Temporal Causal Structure Learning

Partial order priors are usually summary and lag-agnostic, specifying variable-level precedence relations without exact temporal lags, while temporal causal structure learning aims to recover lag-aware structures. To apply such priors to lag-aware structures, we first construct a variable-level summary representation that collapses instantaneous and cross-lag effects into whether a causal influence exists between two variables at any lag, thereby bridging the mismatch between lag-agnostic partial order priors and lag-aware temporal structures.

Definition 2 (Summary Graph)**.** *Given a lag-aware temporal causal structure $W = \{W^\tau\}_{\tau=0}^{p}$, define the summary matrix*

$$S = \sum_{\tau=0}^{p} |W^\tau|, \tag{6}$$

where $|W^\tau|$ denotes the element-wise absolute value of W^τ. The summary graph $G_{\mathrm{sum}}(S)$ is the directed graph over $V = \{v_1, \dots, v_d\}$ such that

$$v_j \to v_i \in G_{\mathrm{sum}}(S) \quad \Longleftrightarrow \quad S_{ji} > 0.$$

The summary graph records whether v_j directly influences v_i at any lag while discarding the exact lag, but still preserves the variable-level reachability needed by lag-agnostic partial order priors.

Proposition 1 (Summary Reachability Equivalence)**.** *For any pair of variables (v_i, v_j), v_j can reach v_i through a sequence of instantaneous or cross-lag causal effects in the lag-aware temporal structure if and only if there exists a directed path from v_j to v_i in $G_{\mathrm{sum}}(S)$.*

The proof is deferred to Appendix C.1. This proposition shows that variable-level partial order constraints can be characterized on the summary graph.

However, discarding exact temporal lags removes the temporal directionality that makes cross-lag effects acyclic by construction. For example, opposite causal directions may occur at different lags but collapse into a cycle in $G_{\mathrm{sum}}(S)$. Thus, the summary graph can be non-DAG, and a causal ordering is no longer naturally defined. Such cycles arise from the summary graph, whereas the expert-provided prior is a valid precedence relation over distinct variables, as discussed in Appendix D.

To define partial order constraints in temporal setting, we follow the causal meaning of a precedence relation. If $v_i \prec v_j$, then possible causality between v_i and v_j should only be allowed in the direction from v_i to v_j. Equivalently, all reversed causality from v_j to v_i, whether direct or indirect, should be forbidden. This leads to the following definition.

Definition 3 (Temporal Partial Order Constraint)**.** *Let $O \subseteq V \times V$ be a partial order prior set represented as strict causal precedence relations over distinct variables, and let O^+ denote its transitive closure. A lag-aware temporal causal structure W satisfies the partial order prior set O if its summary graph $G_{\mathrm{sum}}(S)$ contains no directed path from v_j to v_i for every $(v_i, v_j) \in O^+$, namely*

$$v_j \not\rightsquigarrow_{G_{\mathrm{sum}}(S)} v_i, \qquad \forall (v_i, v_j) \in O^+.$$

This definition converts temporal partial order constraints into path prohibition constraints on the summary graph. It avoids relying on a topological ordering and is therefore extendable to non-DAG summary structures. The remaining challenge is how to represent these path-prohibition constraints in a differentiable and stable way, especially when $G_{\mathrm{sum}}(S)$ contains cycles.

3.2 Cycle-Induced Instability of Power-Sum Path Prohibition

The previous subsection converts temporal partial order constraints into path prohibition constraints on the summary graph. A natural implementation is to follow the static formulation and use a matrix power-sum to characterize reachability on $G_{\mathrm{sum}}(S)$.

Definition 4 (Matrix Power-Sum Path Prohibition Constraint)**.** *Given the summary matrix S, define*

$$P_{\mathrm{sum}}(S) = \sum_{k=1}^{d-1} S^k. \tag{7}$$

The corresponding path-prohibition penalty for partial order constraints is

$$L_{\mathrm{sum}}(S; O) = \sum_{(v_i, v_j) \in O^+} \left[P_{\mathrm{sum}}(S)\right]_{ji}. \tag{8}$$

The basis of this construction is that matrix powers characterize directed walks.

Proposition 2 (Matrix Powers Count Directed Walks)**.** *For any nonnegative matrix S and any integer $k \geq 1$, the entry $\left[S^k\right]_{ji}$ equals the total weight of all directed walks of length k from v_j to v_i in $G_{\mathrm{sum}}(S)$. In particular,*

$$\left[S^k\right]_{ji} > 0 \quad \Longleftrightarrow \quad \textit{there exists a directed walk of length } k \textit{ from } v_j \textit{ to } v_i.$$

On acyclic graphs, directed walks cannot repeatedly traverse cycles, so the matrix power-sum provides a clean characterization of reachability.

Corollary 1 (Exactness on Acyclic Graphs)**.** *If $G_{\mathrm{sum}}(S)$ is acyclic, then for any pair (v_i, v_j),*

$$\left[P_{\mathrm{sum}}(S)\right]_{ji} > 0 \quad \Longleftrightarrow \quad v_j \rightsquigarrow_{G_{\mathrm{sum}}(S)} v_i.$$

Consequently,

$$L_{\mathrm{sum}}(S; O) = 0 \quad \Longleftrightarrow \quad v_j \not\rightsquigarrow_{G_{\mathrm{sum}}(S)} v_i, \qquad \forall (v_i, v_j) \in O^+.$$

However, temporal summary graphs can contain cycles after effects at different lags are collapsed. In this case, the power-sum does not only reflect whether a forbidden reachability relation exists, but also accumulates repeated walks that traverse the same cycle multiple times. Therefore, the penalty magnitude may be dominated by repeated-walk accumulation rather than by simple reachability.

Proposition 3 (Cycle-Induced Penalty Inflation)**.** *Suppose that $G_{\mathrm{sum}}(S)$ contains a directed cycle lying on a directed walk from v_j to v_i. Then $\left[P_{\mathrm{sum}}(S)\right]_{ji}$ depends not only on the reachability from v_j to v_i, but also on repeated traversals of that cycle. From a matrix perspective, repeated cyclic propagation contributes through higher-order powers of S, whose magnitude can grow when the propagation scale of S, reflected by its spectral radius, is not controlled. Thus, on cyclic summary graphs, the matrix power-sum can inflate the path-prohibition penalty and destabilize optimization.*

The proof is deferred to Appendix C.2. The resulting issue is illustrated by the following example.

Example 1 (Cycle-Induced Penalty Inflation)**.** *Consider the partial order prior $v_3 \prec v_1$, which forbids $v_1 \rightsquigarrow_{G_{\mathrm{sum}}(S)} v_3$. Only four active variables are shown, while the full summary graph may contain many more variables. First consider the acyclic graph with edges $v_1 \to v_2$, $v_2 \to v_3$, and $v_2 \to v_4$, and assume that all edge weights are equal to a common value $w > 0$. Then the only directed path from v_1 to v_3 is $v_1 \to v_2 \to v_3$, and thus*

$$\left[P_{\mathrm{sum}}(S)\right]_{13} = w^2.$$

Now add one more edge $v_4 \to v_2$, which creates the directed cycle $v_2 \leftrightarrow v_4$ while keeping the same forbidden reachability relation $v_1 \rightsquigarrow v_3$. Besides the original walk $v_1 \to v_2 \to v_3$, the matrix power-sum also counts repeated walks through the cycle, such as $v_1 \to v_2 \to v_4 \to v_2 \to v_3$ and $v_1 \to v_2 \to v_4 \to v_2 \to v_4 \to v_2 \to v_3$. More generally, for every integer $m \geq 0$, there exists a directed walk

$$v_1 \to v_2 \to (v_4 \to v_2)^m \to v_3$$

of length $2 + 2m$, whose contribution to $\left[P_{\mathrm{sum}}(S)\right]_{13}$ is w^{2+2m}. Hence, when

$$P_{\mathrm{sum}}(S) = \sum_{k=1}^{d-1} S^k,$$

the contribution to the same partial order penalty becomes

$$\left[P_{\mathrm{sum}}(S)\right]_{13} = \sum_{m=0}^{\lfloor (d-3)/2 \rfloor} w^{2+2m}.$$

This value is strictly larger than the acyclic case whenever $w > 0$ and $d \geq 5$.

This example shows that, under the same partial order prior and the same forbidden reachability relation, introducing a cycle can substantially inflate the matrix power-sum penalty. The inflation is caused not by a stronger path-level violation, but by repeated-walk accumulation through the cycle.

Therefore, matrix power-sum path prohibition is a natural extension of the static formulation, but it can become unstable on cyclic summary graphs because repeated-walk accumulation disproportionately amplifies the prior penalty. This motivates a more stable implementation in the next subsection.

3.3 Decoupled Stable Partial Order Constraint

The above analysis shows that directly using a matrix power-sum to represent path prohibition constraints can be unstable on cyclic summary graphs. The key issue is that cyclic propagation may repeatedly amplify higher-order walk contributions when the propagation scale of the summary matrix is not controlled. Therefore, instead of applying partial order constraints on the original summary matrix, we develop a decoupled stable partial order constraint for temporal causal structure learning.

The key idea is to decouple lag-aware data fitting from summary-level partial order constraints. Specifically, the original temporal structure $W = \{W^\tau\}_{\tau=0}^{p}$ is used for temporal structure learning, while partial order priors are imposed through a normalized summary representation derived from W. This design reflects the different requirements of the two objectives: data fitting depends on the magnitudes of instantaneous and cross-lag effects, whereas partial order priors only require variable-level reachability, which can be stably captured by the normalized summary graph.

To stabilize summary-level propagation, we first normalize the summary matrix.

Definition 5 (Normalized Summary Representation)**.** *Given the summary matrix S, define*

$$\tilde{S} = \alpha \frac{S}{\|S\|_F + \varepsilon}, \tag{9}$$

where $\alpha \in (0, 1)$ is a scaling coefficient and $\varepsilon > 0$ is a small constant.

This normalization preserves the variable-level support pattern of S while controlling the propagation scale of the summary representation. The following proposition gives the key stability property.

Proposition 4 (Bounded Propagation)**.** *For any nonzero summary matrix S, the normalized matrix $\tilde{S}$ satisfies*

$$\|\tilde{S}\|_F < 1.$$

Consequently, its spectral radius satisfies

$$\rho(\tilde{S}) < 1.$$

The proof is deferred to Appendix C.3. This result shows that higher-order propagation decays geometrically under the normalized representation. Therefore, even when the summary graph contains cycles, repeated cyclic walks cannot accumulate without control. Meanwhile, positive rescaling preserves the support pattern of S and keeps the summary-level reachability semantics unchanged. Based on the normalized representation, we define a closed-form connectivity characterization.

Definition 6 (Closed-form Connectivity Characterization)**.** *Given $\tilde{S}$, define*

$$P(\tilde{S}) = \sum_{k=1}^{\infty} \tilde{S}^k.$$

Since $\rho(\tilde{S}) < 1$, the series converges and admits the closed-form expression

$$P(\tilde{S}) = (I - \tilde{S})^{-1} - I. \tag{10}$$

Here, normalization provides stability by bounding the propagation scale, while the closed-form connectivity characterization efficiently computes all-order reachability without explicit matrix power-sum accumulation. It also reduces the computational complexity from $O(d^4)$ to $O(d^3)$.

Let $M \in \{0,1\}^{d\times d}$ be the mask induced by the transitive closure O^+, where

$$M_{ij} = 1 \iff (v_i, v_j) \in O^+ \text{and } i \neq j.$$

The decoupling-based partial order constraint is then defined as

$$L_{\text{ord}}(W; O) = \langle M^\top, P(\tilde{S}) \rangle. \tag{11}$$

Finally, the overall learning objective becomes

$$\min_W \; L_{\text{data}}(W) + \lambda_0 \|W^0\|_1 + \sum_{\tau=1}^{p} \lambda_\tau \|W^\tau\|_1 + \beta L_{\text{ord}}(W; O) \quad \text{s.t.} \quad h(W^0) = 0, \tag{12}$$

where $\beta > 0$ controls the strength of the partial order constraint. In this way, lag-agnostic partial order priors can be integrated into temporal causal structure learning through a stable and efficient constraint mechanism.

4 Experiments

This section evaluates the proposed method from multiple perspectives, including its effectiveness in linear, nonlinear, and nonstationary settings, as well as its practical value on real-world data. Additional experimental results are provided in Appendix F.

4.1 Experimental Setup

Synthetic Data and Backbones. We evaluate the proposed method under three synthetic regimes: linear, nonlinear, and nonstationary temporal structure learning. In the linear setting, we use DYNOTEARS [Pamfil et al., 2020] as the backbone. Temporal graphs are generated by Erdős–Rényi (ER) and Scale-Free (SF) models, and observations are simulated from linear temporal structural equations with both instantaneous and cross-lag dependencies under Gaussian and exponential noise. In the nonlinear setting, we use NTS-NOTEARS [Sun et al., 2021] as the backbone and generate nonlinear temporal data under two causal mechanism families, the Additive Noise Model (ANM) and Additive Index Model (AIM). In the nonstationary setting, we consider two distinct forms of nonstationarity using LIN [Liu and Kuang, 2023] and RHINO [Gong et al., 2022]. LIN models time-varying causal mechanisms induced by latent interventions, whereas RHINO models history-dependent noise whose distribution depends on past observations. In both settings, the underlying causal structure is kept fixed. The nonstationarity arises from changes in the causal mechanisms or noise process. For each parameter setting, we generate six independent datasets.

Real-World Data. For real-world evaluation, we use the DREAM4 gene regulatory network dataset, a standard benchmark for time-series structure learning [Marbach et al., 2009, 2010]. It provides gene expression trajectories under different perturbations together with known regulatory interactions. Since the reference network specifies regulatory relations but not their exact temporal lags, DREAM4 is well suited for evaluating lag-agnostic structural priors.

Prior Construction. For controlled evaluation, partial-order priors are sampled from the ground-truth or reference structure so that their coverage can be varied systematically. The prior ratio denotes the fraction of relations sampled from the corresponding transitive closure. This construction is used only as an evaluation protocol and does not assume access to the ground-truth structure in practical applications. Possible real-world sources of partial-order knowledge are discussed in Appendix B.3.

Metrics. For synthetic experiments, we report Scaled SHD, precision, recall, F1 score, and prior recovery rate. Unless otherwise specified, these structural metrics are computed on the complete lag-aware causal structure rather than only on the variable-level summary graph. Therefore, assigning a causal relation to an incorrect temporal lag is counted as a structural error. Specifically, $\text{Scaled SHD} = \text{SHD}/|E|$, where $|E|$ is the number of ground-truth lag-aware causal edges; $\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$, $\text{Recall} = \text{TP}/(\text{TP} + \text{FN})$, and $\text{F1} = 2 \cdot \text{Precision} \cdot$

$\mathrm{Recall}/(\mathrm{Precision} + \mathrm{Recall})$. The prior recovery rate is defined as the number of recovered prior relations divided by the total number of provided prior relations. For real-world experiments, we additionally report AUROC and data loss, which measure summary-level structural recovery and predictive fitting on the observed data, respectively.

Implementation Details. Unless otherwise specified, we set $\beta = 0.1$, $\alpha = 0.999$, and $\varepsilon = 10^{-8}$, and otherwise follow the default hyperparameters of the corresponding backbone learner. Methods using the proposed partial-order constraint are denoted by "-O-r", where r denotes the prior ratio. The data-only backbone is equivalently denoted by $r = 0$ when included in plots.

4.2 Effectiveness in Linear Temporal Structure Learning

This experiment evaluates the effectiveness of the proposed method in linear temporal structure learning using DYNOTEARS as the backbone. We generate synthetic multivariate time series from linear temporal structural equations with both instantaneous and cross-lag causal dependencies. We consider ER and SF graphs with edge-to-node ratios of $\{1, 2\}$ under Gaussian and exponential noise. The number of variables is varied over $d \in \{10, 20, 30, 40, 50\}$, the sample size is fixed at $T = 200$, and the prior ratio is varied over $r \in \{0.25, 0.50, 0.75, 1.00\}$. We compare the data-only DYNOTEARS baseline with its partial-order-guided variants, denoted by DYNOTEARS-O-r, and evaluate their performance using Scaled SHD, F1 score, and prior recovery rate.

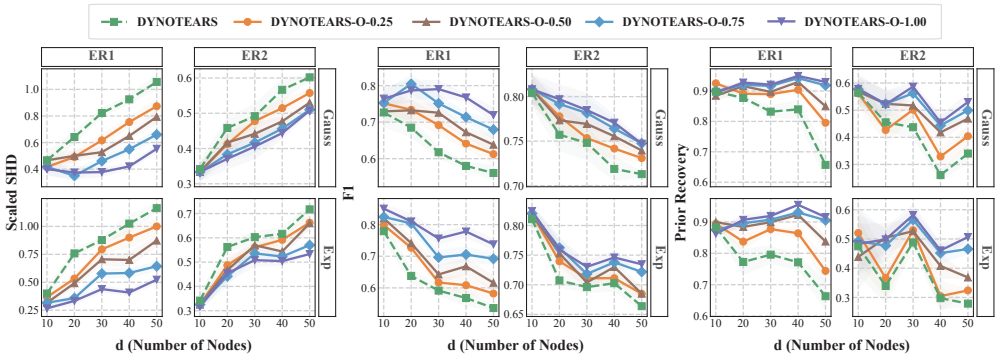

Figure 2: Performance in linear temporal structure learning under different settings.

Results are shown in Fig. 2. Across different graph sizes, graph densities, and noise distributions, incorporating partial-order priors generally improves temporal structure recovery over the data-only DYNOTEARS baseline, yielding lower Scaled SHD, higher F1 scores, and better prior recovery. The improvements become more pronounced as the prior ratio increases and are particularly clear on larger graphs, where structure learning from data alone becomes more challenging. These results demonstrate that the proposed partial-order constraint can effectively exploit lag-agnostic prior information to improve linear temporal structure learning.

4.3 Extension to Nonlinear Temporal Structure Learning

This experiment evaluates whether the proposed constraint method remains effective in nonlinear temporal structure learning using NTS-NOTEARS as the backbone. We generate nonlinear synthetic time series on 20-node ER graphs with edge-to-node ratios of $\{1, 2\}$, denoted by ER1 and ER2, respectively. Following the nonlinear setting of NTS-NOTEARS, the causal mechanisms are generated under the ANM and AIM, and the sample size is set to $T \in \{200, 500\}$. The prior ratio is varied over $r \in \{0, 0.25, 0.50, 0.75, 1.00\}$, where $r = 0$ corresponds to the data-only baseline. We compare NTS-NOTEARS with its partial-order-guided variants using Scaled SHD and F1 score.

Results are shown in Fig. 3. Across both AIM and ANM mechanisms and both sample sizes, incorporating partial-order priors generally improves nonlinear temporal structure recovery over the data-only baseline. Scaled SHD tends to decrease and F1 tends to increase as the prior ratio grows. The improvements are more consistent under AIM, while ANM also exhibits clear gains despite mild non-monotonicity at some intermediate prior ratios. These results show that the proposed constraint remains effective across different nonlinear causal mechanisms and sample sizes.

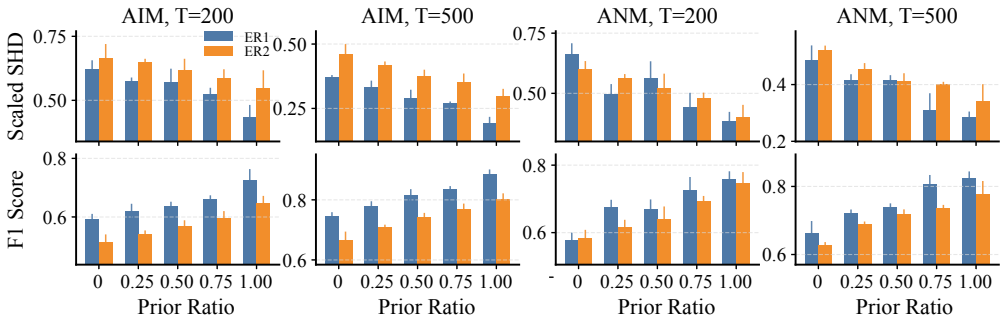

Figure 3: Performance in nonlinear temporal structure learning under different settings.

4.4 Extension to Nonstationary Temporal Structure Learning

This experiment evaluates whether the proposed constraint method remains effective in nonstationary temporal structure learning using LIN and RHINO as two representative backbones. The two methods consider distinct forms of nonstationarity while keeping the underlying causal structure fixed. LIN considers time-varying causal mechanisms induced by latent interventions. We use MLP-based nonlinear data on 5-node ER graphs with one lag, two intervention nodes, and latent intervention trajectories generated by one- or two-dimensional Brownian motion. RHINO considers nonstationarity induced by history-dependent noise, where the noise process depends on past observations. We use ER graphs with $d \in \{5, 10, 20\}$, two lags, spline-product noise, and an intervention history of length 3. For both backbones, the prior ratio is varied over $r \in \{0, 0.25, 0.50, 0.75, 1.00\}$, where $r = 0$ corresponds to the data-only baseline. All experiments use 5000 samples.

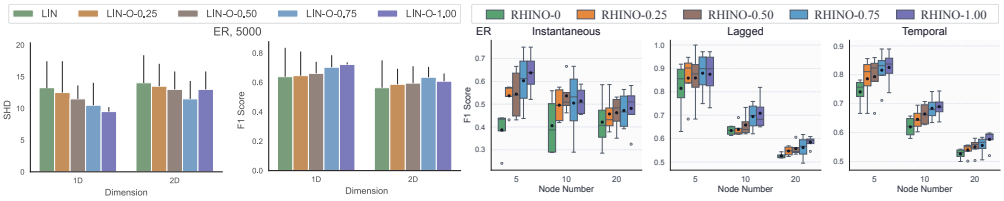

Figure 4: Performance in nonstationary temporal structure learning. Left: LIN. Right: RHINO.

Results are shown in Fig. 4. For LIN, incorporating partial-order priors generally reduces SHD and improves F1 as the prior ratio increases, indicating that the proposed constraint remains effective under time-varying causal mechanisms induced by latent interventions. For RHINO, the prior-guided variants achieve higher F1 scores on instantaneous, lagged, and complete temporal structures across different graph sizes, showing that the constraint also remains beneficial under history-dependent noise. These results demonstrate that the proposed method can be combined with different nonstationary temporal learners and remains effective across complementary forms of mechanism and noise nonstationarity.

4.5 Real-World Validation

This experiment evaluates the practical value of the proposed method on the DREAM4 gene regulatory network dataset. We use DYNOTEARS as the temporal backbone and include NOTEARS as a static reference method. We compare their data-only and partial-order-guided variants together with the Mean, Max, and Prob variants described in Appendix F.1. We evaluate DREAM4-63, DREAM4-126, and DREAM4-189, where 63, 126, and 189 denote different sample sizes, under prior ratios of 25% and 75%. Since the DREAM4 reference network specifies regulatory relations without exact temporal lags, we report AUROC to evaluate summary-level structural recovery and data loss to measure predictive fitting.

Results in Table 1 show that incorporating partial-order priors consistently improves summary-level structural recovery on DREAM4. DYNOTEARS-O achieves the highest AUROC across all three

Table 1: DREAM4 real-data results for partial order prior structure learning.

Setting		DREAM4-63		DREAM4-126		DREAM4-189	
Prior Rate	**Method**	**DataLoss**	**AUROC**	**DataLoss**	**AUROC**	**DataLoss**	**AUROC**
25%	NOTEARS	8.20 ± 0.73	0.54 ± 0.03	8.02 ± 0.92	0.55 ± 0.03	7.89 ± 0.73	0.55 ± 0.03
	NOTEARS-O	8.63 ± 0.58	0.58 ± 0.03	7.93 ± 0.76	0.56 ± 0.02	7.08 ± 0.85	0.56 ± 0.03
	DYNOTEARS	7.81 ± 1.14	0.58 ± 0.03	$\mathbf{5.61 \pm 1.04}$	0.62 ± 0.04	$\mathbf{4.08 \pm 0.60}$	0.65 ± 0.05
	DYNOTEARS-O-Mean	7.89 ± 1.06	0.64 ± 0.05	7.73 ± 0.89	0.64 ± 0.08	6.93 ± 1.34	0.66 ± 0.10
	DYNOTEARS-O-Max	8.40 ± 0.85	0.60 ± 0.08	7.35 ± 0.95	0.62 ± 0.11	7.61 ± 1.62	0.63 ± 0.15
	DYNOTEARS-O-Prob	9.91 ± 1.28	0.57 ± 0.04	8.88 ± 1.39	0.62 ± 0.04	7.23 ± 0.71	0.65 ± 0.06
	DYNOTEARS-O	$\mathbf{7.59 \pm 0.61}$	$\mathbf{0.66 \pm 0.02}$	6.31 ± 1.15	$\mathbf{0.68 \pm 0.09}$	6.14 ± 2.02	$\mathbf{0.69 \pm 0.07}$
75%	NOTEARS	8.20 ± 0.73	0.54 ± 0.03	8.02 ± 0.92	0.55 ± 0.03	7.89 ± 0.73	0.55 ± 0.03
	NOTEARS-O	8.25 ± 0.66	0.61 ± 0.05	7.57 ± 0.85	0.61 ± 0.05	7.53 ± 0.32	0.62 ± 0.04
	DYNOTEARS	7.81 ± 1.14	0.58 ± 0.03	5.61 ± 1.04	0.62 ± 0.04	$\mathbf{4.08 \pm 0.60}$	0.65 ± 0.05
	DYNOTEARS-O-Mean	7.42 ± 0.72	0.71 ± 0.03	7.19 ± 0.97	0.70 ± 0.05	6.20 ± 1.24	0.75 ± 0.09
	DYNOTEARS-O-Max	7.78 ± 0.54	0.67 ± 0.02	6.71 ± 0.86	0.69 ± 0.09	6.79 ± 1.57	0.69 ± 0.13
	DYNOTEARS-O-Prob	8.16 ± 2.16	0.56 ± 0.04	7.25 ± 3.04	0.64 ± 0.05	6.99 ± 0.84	0.67 ± 0.07
	DYNOTEARS-O	$\mathbf{7.41 \pm 0.65}$	$\mathbf{0.72 \pm 0.02}$	$\mathbf{5.36 \pm 0.91}$	$\mathbf{0.73 \pm 0.04}$	4.71 ± 0.69	$\mathbf{0.76 \pm 0.06}$

sample sizes under both prior ratios, with larger improvements generally observed when more prior information is available. The effect on data loss is less uniform because the partial-order constraint primarily regularizes structural consistency rather than directly optimizing predictive fitting. Consequently, improved structural recovery does not necessarily lead to a lower data loss in every setting. Overall, the results show that DYNOTEARS-O substantially improves structural recovery while maintaining comparable predictive fitting performance.

4.6 Additional Experimental Analyses

In addition to the main experiments, we conduct a series of supplementary analyses in Appendix F. These include ablations on summary construction and stabilization, stability diagnostics of different constraint formulations, robustness evaluation under incorrect partial-order priors, comparison with matched edge-prior constraints, and computational cost analysis. We further study sensitivity to parameter initialization, the prior weight β, the scaling coefficient α, the amount of available prior information, and the maximum temporal lag. Additional experiments on larger temporal graphs, qualitative visualizations, and complete results for the main experimental settings are also provided.

5 Conclusion

This paper studies how to integrate summary and lag-agnostic partial order priors into temporal causal structure learning. We first define temporal partial order constraints on the summary graph and convert them into path prohibition constraints that can be extended to non-DAG summary structures. We then show that a direct matrix power-sum formulation can become unstable on cyclic summary graphs, where repeated-walk accumulation inflates the path-prohibition penalty and destabilizes optimization. To address this issue, we propose a stable decoupled partial order constraint, which preserves the lag-aware temporal structure for data fitting while imposing partial order priors through a normalized summary representation. Experiments further extend the evaluation to nonstationary settings and demonstrate the effectiveness of the proposed method.

Acknowledgements

This research is supported in part by the National Key R&D Program of China (No. 2026YFE0104800), in part by the National Natural Science Foundation of China (No. 62406302, 62137002), in part by the Natural Science Foundation of Anhui Province (No. 2408085QF195), in part by the Fundamental Research Funds for the Central Universities under Grant WK2150110035, in part by the USTC Research Funds of the Double First-Class Initiative (Grant No. YD9110002085 and YD2155202608), and in part by the Research Funds of Centre for Leading Medicine and Advanced Technologies of IHM (Grant No. 2025IHM01030).

References

Ankur Ankan and Johannes Textor. Expert-in-the-loop causal discovery: Iterative model refinement using expert knowledge. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

Charles K Assaad, Emilie Devijver, and Eric Gaussier. Discovery of extended summary graphs in time series. In *Uncertainty in Artificial Intelligence*, pages 96–106, 2022.

Carles Balsells-Rodas, Xavier Sumba, Tanmayee Narendra, Ruibo Tu, Gabriele Schweikert, Hedvig Kjellstrom, and Yingzhen Li. Causal discovery from conditionally stationary time series. *arXiv preprint arXiv:2110.06257*, 2021.

Taiyu Ban, Lyuzhou Chen, Xiangyu Wang, Xin Wang, Derui Lyu, and Huanhuan Chen. Differentiable structure learning with partial orders. *Advances in Neural Information Processing Systems*, 37:117426–117455, 2024.

Taiyu Ban, Changxin Rong, Xiangyu Wang, Lyuzhou Chen, Xin Wang, Derui Lyu, Qinrui Zhu, and Huanhuan Chen. Differentiable structure learning with ancestral constraints. In *Forty-second International Conference on Machine Learning*, 2025.

Taiyu Ban, Changxin Rong, Xiangyu Wang, Lyuzhou Chen, Yanze Gao, Xin Wang, and Huanhuan Chen. Structure learning from time-series data with lag-agnostic structural prior. In *International Conference on Learning Representations*, 2026.

Christine W Bang and Vanessa Didelez. Do we become wiser with time? on causal equivalence with tiered background knowledge. In *Uncertainty in Artificial Intelligence*, pages 119–129, 2023.

Christine W Bang and Vanessa Didelez. Constraint-based causal discovery with tiered background knowledge and latent variables in single or overlapping datasets. *arXiv preprint arXiv:2503.21526*, 2025.

Alexander Capstick, Rahul G Krishnan, and Payam Barnaghi. Autoelicit: Using large language models for expert prior elicitation in predictive modelling. *arXiv preprint arXiv:2411.17284*, 2024.

Lyuzhou Chen, Taiyu Ban, Derui Lyu, Yijia Sun, Kangtao Hu, Xiangyu Wang, and Huanhuan Chen. Continuous structure constraint integration for robust causal discovery. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Tianjiao Chu, Clark Glymour, and Greg Ridgeway. Search for additive nonlinear time series causal models. *Journal of Machine Learning Research*, 9(5), 2008.

Anthony C Constantinou, Zhigao Guo, and Neville K Kitson. The impact of prior knowledge on causal structure learning. *Knowledge and Information Systems*, 65(8):3385–3434, 2023.

Simon Ferreira and Charles K Assaad. Identifiability of direct effects from summary causal graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20387–20394, 2024.

Tian Gao, Debarun Bhattacharjya, Elliot Nelson, Miao Liu, and Yue Yu. IDYNO: Learning nonparametric DAGs from interventional dynamic data. In *International Conference on Machine Learning*, pages 6988–7001, 2022.

Wenbo Gong, Joel Jennings, Cheng Zhang, and Nick Pawlowski. Rhino: Deep causal temporal relationship learning with history-dependent noise. *arXiv preprint arXiv:2210.14706*, 2022.

Wiebke Günther, Urmi Ninad, and Jakob Runge. Causal discovery for time series from multiple datasets with latent contexts. In *Uncertainty in Artificial Intelligence*, pages 766–776, 2023.

Stefan Haufe, Klaus-Robert Müller, Guido Nolte, and Nicole Krämer. Sparse causal discovery in multivariate time series. In *Causality: Objectives and Assessment*, pages 97–106, 2010.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

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

Xiaofei Li, Zaneta Andrusivova, Paulo Czarnewski, Christoffer Mattsson Langseth, Alma Andersson, Yang Liu, Daniel Gyllborg, Emelie Braun, Ludvig Larsson, Lijuan Hu, et al. Profiling spatiotemporal gene expression of the developing human spinal cord and implications for ependymoma origin. *Nature Neuroscience*, 26(5):891–901, 2023.

Chenxi Liu and Kun Kuang. Causal structure learning for latent intervened non-stationary data. In *International Conference on Machine Learning*, pages 21756–21777, 2023.

Sindy Löwe, David Madras, Richard Zemel, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Conference on Causal Learning and Reasoning*, pages 509–525, 2022.

Daniel Marbach, Thomas Schaffter, Claudio Mattiussi, and Dario Floreano. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of Computational Biology*, 16(2):229–239, 2009.

Daniel Marbach, Robert J Prill, Thomas Schaffter, Claudio Mattiussi, Dario Floreano, and Gustavo Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences*, 107(14):6286–6291, 2010.

Søren Wengel Mogensen, Karin Rathsman, and Per Nilsson. Causal discovery in a complex industrial system: A time series benchmark. In *Causal Learning and Reasoning*, pages 1218–1236, 2024.

Teppo Niinimäki, Pekka Parviainen, and Mikko Koivisto. Structure discovery in bayesian networks by sampling partial orders. *Journal of Machine Learning Research*, 17(57):1–47, 2016.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. DYNOTEARS: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pages 1595–1605, 2020.

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.

Xiangyu Sun, Oliver Schulte, Guiliang Liu, and Pascal Poupart. NTS-NOTEARS: Learning nonparametric DBNs with prior knowledge. *arXiv preprint arXiv:2109.04286*, 2021.

Sumanth Varambally, Yi-An Ma, and Rose Yu. Discovering mixtures of structural causal models from time series data. *arXiv preprint arXiv:2310.06312*, 2023.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425, 2020.

A Appendix Overview

This appendix provides additional material to support the main paper. Appendix B reviews related work on temporal causal structure learning, prior-guided causal discovery, and practical sources of partial order priors. Appendix C gives proofs for the main theoretical results, including the summary reachability equivalence, the instability of matrix power-sum path prohibition on cyclic summary graphs, and the bounded propagation property of the proposed normalized formulation. Appendix D discusses the mechanism, inductive bias, validity requirements, normalization trade-offs, and practical implications of the proposed method. Appendix E summarizes the limitations of the current formulation, including its dependence on valid partial order priors, the trade-off introduced by global normalization, and scalability considerations for very large temporal graphs. Appendix F reports complete experimental results, including ablation studies, stability diagnosis, robustness and sensitivity analyses, efficiency evaluation, and qualitative visualizations. Together, these materials complement the main text by clarifying the theoretical foundation, methodological design, and empirical behavior of stable partial order constraints for temporal causal structure learning.

B Related Work

This section reviews three lines of related research. We first summarize temporal causal structure learning, with an emphasis on differentiable, nonlinear, and nonstationary settings. We then discuss causal structure learning with prior knowledge, including prior integration in the static setting and existing temporal edge-prior methods, and clarify how our setting differs from them. Finally, we review practical sources of partial order priors to show why summary and lag-agnostic precedence knowledge is realistic and important in temporal applications.

B.1 Temporal Causal Structure Learning

Temporal causal structure learning extends causal discovery from static data to multivariate time series, where the goal is to recover both instantaneous and cross-lag causal relations. Early studies often relied on sparse autoregressive or Granger-style formulations. For example, sparse causal discovery in multivariate time series has been studied through structured regularization over vector autoregressive coefficients [Haufe et al., 2010], while structural vector autoregressive models have also been analyzed from the perspective of non-Gaussian identifiability [Hyvärinen et al., 2010]. In parallel, constraint-based and functional approaches were extended to temporal settings to handle lagged dependencies, hidden confounding, and nonlinear relations, as illustrated by additive nonlinear time-series causal models [Chu et al., 2008] and the PCMCI framework for large nonlinear time-series datasets [Runge et al., 2019].

More recent work has emphasized continuous optimization and differentiable structure learning for time series. DYNOTEARS [Pamfil et al., 2020] extended the NOTEARS framework to dynamic Bayesian networks by jointly estimating instantaneous and cross-lag edges under a smooth acyclicity constraint on the instantaneous graph. This line has been further generalized to nonlinear and nonparametric settings, for example through IDYNO [Gao et al., 2022], which learns nonlinear temporal causal structures from interventional and observational temporal data, NTS-NOTEARS [Sun et al., 2021], which incorporates neural temporal models into a differentiable DBN learning framework, and amortized neural approaches for inferring causal graphs from time-series data [Löwe et al., 2022]. At the same time, more specialized temporal settings have received increasing attention, including summary-graph learning [Assaad et al., 2022], multiple time-series datasets with latent contexts [Günther et al., 2023], mixtures of temporal causal models [Varambally et al., 2023], and nonstationary time series, where representative examples include latent-domain shifts [Liu and Kuang, 2023], history-dependent noise [Gong et al., 2022], and more general heterogeneous or conditionally stationary settings [Huang et al., 2020, Balsells-Rodas et al., 2021].

Despite this rich literature, most temporal causal discovery methods remain predominantly data-driven. Their focus has been on improving identifiability, expressiveness, and robustness to nonlinearities, latent contexts, and nonstationarity, rather than on systematically incorporating high-level prior knowledge into lag-aware temporal structure learning. In particular, while these methods estimate lag-aware structures or summary graphs from data, much less attention has been paid to

how summary and lag-agnostic prior information defined over variables should be integrated into temporal settings.

B.2 Causal Structure Learning with Prior Knowledge

The incorporation of prior knowledge has long been recognized as an effective way to improve causal structure learning, especially when data are limited, noisy, or weakly informative [Constantinou et al., 2023]. In Bayesian-network structure learning, partial orders and related ordering information have been studied for many years, with representative examples including structure discovery methods based on sampling partial orders [Niinimäki et al., 2016]. More recently, differentiable causal discovery has made it easier to encode prior knowledge through continuous objectives. NOTEARS [Zheng et al., 2018] provided a foundational optimization framework for DAG learning, and subsequent studies extended this paradigm to nonlinear settings [Zheng et al., 2020] and temporal settings [Pamfil et al., 2020].

Within this line, several studies have investigated how structural prior knowledge can be incorporated more explicitly. Representative examples include partial order constraints in differentiable structure learning [Ban et al., 2024], ancestral constraints for encoding higher-level causal knowledge [Ban et al., 2025], and soft edge-constraint integration for improving robustness to erroneous edge priors [Chen et al., 2025]. Related ideas have also been explored in the constraint-based literature, where tiered background knowledge has been shown to improve both informativeness and efficiency in settings with latent variables and overlapping datasets [Bang and Didelez, 2025, 2023]. These methods demonstrate the value of high-level prior knowledge, but most of them are developed under static DAG semantics, where reachability and ancestor relations are evaluated in an acyclic graph space.

Prior integration has also begun to appear in temporal causal structure learning, but existing efforts mainly focus on edge-level knowledge. For example, NTS-NOTEARS incorporates prior knowledge in the form of dependency constraints on temporal edges [Sun et al., 2021]. More recently, lag-agnostic structural priors have been studied in temporal structure learning, where the prior specifies whether a causal edge should exist without requiring its exact lag [Ban et al., 2026]. Such priors still describe local edge existence. By contrast, this work focuses on partial order priors, which are defined at the variable level and encode global precedence relations rather than individual dependencies. Compared with edge priors, partial order priors are often easier to obtain when experts know coarse causal precedence but cannot specify exact temporal lags, while their integration into temporal structure learning is also more challenging.

The main difficulty is that static path-based prior constraints cannot be directly reused in temporal settings. In static causal discovery, partial order and ancestral constraints are usually imposed on a DAG, and path-prohibition penalties can rely on the acyclic graph structure to characterize forbidden reachability. In temporal causal structure learning, however, a lag-aware structure is often compressed into a summary graph that records variable-level causal influence while discarding exact temporal lags. Such a summary graph can contain directed cycles, because opposite causal directions may occur at different lags and become collapsed at the variable level. As a result, directly applying static matrix power-sum penalties to temporal summary graphs may not only penalize forbidden reachability, but also repeatedly accumulate cyclic walks, leading to unstable optimization. Therefore, our work differs from existing static path-based prior methods by reformulating partial order priors under temporal summary semantics and introducing a decoupled normalized propagation mechanism. It also differs from temporal edge-prior integration methods, which provide local dependency supervision, whereas the proposed partial order constraints impose global precedence restrictions through direct and indirect reachability.

B.3 Practical Sources of Partial Order Priors

Prior knowledge in real applications can arise from multiple sources. A common source is domain expertise, where experts may know that certain variables tend to causally precede or influence others even when the exact temporal lag is unclear. Prior knowledge may also come from established scientific theories, system diagrams, intervention protocols, or process rules that encode coarse structural relations among variables. These forms of knowledge are typically easier to express as variable-level precedence, tiering, or ancestral relations than as exact temporal edges.

Recent work has also highlighted how such priors can be formalized and elicited in practice. Tiered background knowledge provides a natural formalization of coarse ordering information [Bang and Didelez, 2025, 2023]. The practical importance of expert-derived structural information for temporal discovery has also been illustrated by industrial time-series benchmarks whose reference causal graphs are constructed from expert knowledge [Mogensen et al., 2024]. More interactive forms of prior elicitation have been explored through expert-in-the-loop causal discovery frameworks that iteratively refine learned graphs using domain knowledge, including proof-of-concept settings in which a large language model acts as a domain expert [Ankan and Textor, 2025]. More broadly, recent work suggests that large language models can help elicit informative priors for predictive models from natural language and domain text [Capstick et al., 2024].

Taken together, these studies suggest that useful prior knowledge is often available in practice, but usually in a coarse, partial, and sometimes imperfect form. This motivates the study of prior representations that are easier to obtain than exact lag-specific edge supervision. In this sense, partial order priors are especially appealing: they capture high-level structural knowledge, are often more realistic for experts to provide, and align naturally with emerging LLM-assisted prior elicitation pipelines. At the same time, their coarse and potentially noisy nature means that temporal causal discovery methods must integrate them in a way that is both semantically principled and numerically stable, which is precisely the focus of our work.

C Proof of Statements

This section provides proofs for the main theoretical results. We first show that the summary graph preserves the variable-level reachability needed by lag-agnostic partial order constraints. We then analyze why the matrix power-sum characterization is exact on acyclic graphs but can become unstable on cyclic summary graphs. Finally, we prove that the proposed decoupled formulation yields bounded propagation and a well-defined closed-form connectivity characterization.

C.1 Summary Reachability Equivalence

Proof of Proposition 1. *Proposition 1.* For any pair of variables (v_i, v_j), v_j can reach v_i through a sequence of instantaneous or cross-lag causal effects in the lag-aware temporal structure if and only if there exists a directed path from v_j to v_i in $G_{\mathrm{sum}}(S)$.

Proof. We use an unfolded temporal graph only for the proof. Given $W = \{W^\tau\}_{\tau=0}^{p}$, define an unfolded graph with nodes $v_{i,t}$, where $v_{i,t}$ denotes variable v_i at time t. For any $\tau \in \{0, \dots, p\}$, the graph contains an edge

$$v_{j,t-\tau} \to v_{i,t}$$

if and only if $(W^\tau)_{ji} \neq 0$. The summary matrix is defined as

$$S = \sum_{\tau=0}^{p} |W^\tau|,$$

and the summary graph contains an edge $v_j \to v_i$ if and only if $S_{ji} > 0$. We prove the two directions separately.

Forward direction. Assume that v_j can reach v_i through a sequence of instantaneous or cross-lag causal effects in the lag-aware temporal structure. Equivalently, in the unfolded temporal graph, there exist times $t' \leq t$ and a directed path

$$v_{a_0,t_0} \to v_{a_1,t_1} \to \cdots \to v_{a_m,t_m},$$

where $a_0 = j$, $a_m = i$, $t_0 = t'$, and $t_m = t$. For each edge

$$v_{a_{r-1},t_{r-1}} \to v_{a_r,t_r},$$

there exists some $\tau_r \in \{0, \dots, p\}$ such that

$$t_r = t_{r-1} + \tau_r \qquad \text{and} \qquad (W^{\tau_r})_{a_{r-1}a_r} \neq 0.$$

Hence,

$$S_{a_{r-1}a_r} \geq |(W^{\tau_r})_{a_{r-1}a_r}| > 0,$$

so the summary graph contains the edge

$$v_{a_{r-1}} \to v_{a_r}.$$

Removing the time indices yields a directed walk

$$v_j = v_{a_0} \to v_{a_1} \to \cdots \to v_{a_m} = v_i$$

in $G_{\mathrm{sum}}(S)$. Every directed walk contains a directed path after removing repeated cycles, so there exists a directed path from v_j to v_i in $G_{\mathrm{sum}}(S)$.

Reverse direction. Assume that there is a directed path

$$v_{a_0} \to v_{a_1} \to \cdots \to v_{a_m}$$

in $G_{\mathrm{sum}}(S)$, where $a_0 = j$ and $a_m = i$. Since $S_{a_{r-1}a_r} > 0$ for every r, there exists at least one lag $\tau_r \in \{0, \ldots, p\}$ such that

$$(W^{\tau_r})_{a_{r-1}a_r} \neq 0.$$

Choose one such τ_r for each edge and define time indices recursively by

$$t_0 = 0, \qquad t_r = t_{r-1} + \tau_r \quad (r = 1, \ldots, m).$$

By construction, each edge

$$v_{a_{r-1},t_{r-1}} \to v_{a_r,t_r}$$

belongs to the unfolded temporal graph. Consequently,

$$v_{j,t_0} \to v_{a_1,t_1} \to \cdots \to v_{i,t_m}$$

is a directed path formed by instantaneous or cross-lag causal effects, and $t_0 \leq t_m$ since every $\tau_r \geq 0$. Thus, v_j can reach v_i through a sequence of instantaneous or cross-lag causal effects in the lag-aware temporal structure.

The two directions together prove the equivalence.

C.2 Instability of Matrix Power-Sum Path Prohibition

Proof of Proposition 2. *Proposition 2.* For any nonnegative matrix S and any integer $k \geq 1$, the entry $(S^k)_{ji}$ equals the total weight of all directed walks of length k from v_j to v_i in $G_{\mathrm{sum}}(S)$. In particular,

$$(S^k)_{ji} > 0 \quad \Longleftrightarrow \quad \text{there exists a directed walk of length } k \text{ from } v_j \text{ to } v_i.$$

Proof. By matrix multiplication,

$$(S^k)_{ji} = \sum_{a_1,\ldots,a_{k-1}} S_{ja_1} S_{a_1a_2} \cdots S_{a_{k-1}i},$$

where the sum runs over all sequences $(a_1, \ldots, a_{k-1})$ of intermediate vertices. Each monomial

$$S_{ja_1} S_{a_1a_2} \cdots S_{a_{k-1}i}$$

corresponds to the directed walk

$$v_j \to v_{a_1} \to v_{a_2} \to \cdots \to v_{a_{k-1}} \to v_i.$$

Since S is nonnegative, all contributions are nonnegative, and their sum equals the total weight of all directed walks of length k from v_j to v_i. Therefore,

$$(S^k)_{ji} > 0 \quad \Longleftrightarrow \quad \text{there exists a directed walk of length } k \text{ from } v_j \text{ to } v_i.$$

This proves the proposition.

Proof of Corollary 1. *Corollary 1.* If $G_{\mathrm{sum}}(S)$ is acyclic, then for any pair (v_i, v_j),

$$[P_{\mathrm{sum}}(S)]_{ji} > 0 \quad \Longleftrightarrow \quad v_j \rightsquigarrow_{G_{\mathrm{sum}}(S)} v_i.$$

Consequently,

$$L_{\mathrm{sum}}(S; O) = 0 \quad \Longleftrightarrow \quad v_j \not\rightsquigarrow_{G_{\mathrm{sum}}(S)} v_i, \qquad \forall (v_i, v_j) \in O^+.$$

Proof. Assume that $G_{\mathrm{sum}}(S)$ is acyclic. Then every directed walk is automatically a directed path. Indeed, if a directed walk repeated a vertex, the segment between its first and last occurrence would form a directed cycle, contradicting acyclicity.

Hence, in an acyclic graph, a directed walk from v_j to v_i exists if and only if a directed path from v_j to v_i exists. Moreover, any directed path in a d-node graph has length at most $d - 1$. Therefore,

$$[P_{\mathrm{sum}}(S)]_{ji} = \sum_{k=1}^{d-1} (S^k)_{ji} > 0$$

if and only if there exists a directed path from v_j to v_i in $G_{\mathrm{sum}}(S)$. This proves

$$[P_{\mathrm{sum}}(S)]_{ji} > 0 \quad \Longleftrightarrow \quad v_j \rightsquigarrow_{G_{\mathrm{sum}}(S)} v_i.$$

The second claim follows immediately: $L_{\mathrm{sum}}(S; O) = 0$ holds if and only if every forbidden pair $(v_i, v_j) \in O^+$ satisfies $[P_{\mathrm{sum}}(S)]_{ji} = 0$, which is exactly the path-prohibition condition in the summary graph.

Proof of Proposition 3. *Proposition 3.* Suppose that $G_{\mathrm{sum}}(S)$ contains a directed cycle lying on a directed walk from v_j to v_i. Then $\left[P_{\mathrm{sum}}(S)\right]_{ji}$ depends not only on the reachability from v_j to v_i, but also on repeated traversals of that cycle. From a matrix perspective, repeated cyclic propagation contributes through higher-order powers of S, whose magnitude is governed by the propagation scale of S, as reflected by its spectral radius. Thus, on cyclic summary graphs, the matrix power-sum can inflate the path-prohibition penalty and destabilize optimization.

Proof. Assume that $G_{\mathrm{sum}}(S)$ contains a directed cycle lying on a directed walk from v_j to v_i. Then there exist directed walks

$$\pi_{ju} : v_j \rightsquigarrow u, \qquad \pi_C : u \rightsquigarrow u, \qquad \pi_{ui} : u \rightsquigarrow v_i,$$

where π_C is a nontrivial directed cycle of length $\ell_C \geq 1$. For every integer $m \geq 0$, the concatenated walk

$$\pi_{ju} \circ \pi_C^m \circ \pi_{ui}$$

is again a directed walk from v_j to v_i, with length

$$\ell_{ju} + m\ell_C + \ell_{ui}.$$

Whenever this length is at most $d - 1$, the walk contributes a nonnegative term to

$$\left[P_{\mathrm{sum}}(S)\right]_{ji} = \sum_{k=1}^{d-1} (S^k)_{ji}.$$

Therefore, $\left[P_{\mathrm{sum}}(S)\right]_{ji}$ is not determined only by whether v_j can reach v_i. It also depends on how many repeated traversals of π_C fall within the truncation range and on the corresponding edge weights. In particular, two summary graphs can have the same reachability relation $v_j \rightsquigarrow v_i$ but produce different values of $\left[P_{\mathrm{sum}}(S)\right]_{ji}$ if one graph permits additional cyclic repetitions along the walk.

This walk-level accumulation also has a matrix-level interpretation. For a nonnegative matrix S, the growth of the matrix powers $\{S^k\}$ is governed by the spectral radius $\rho(S)$. By the Gelfand formula,

$$\rho(S) = \lim_{k \to \infty} \|S^k\|^{1/k},$$

for any matrix norm. Moreover, a directed cycle with edge-weight product γ and length ℓ_C contributes a closed walk whose average multiplicative weight is γ^{1/ℓ_C}, giving a cycle-dependent contribution to the propagation scale of S. Therefore, when cyclic propagation makes $\rho(S)$ close to or

larger than one, higher-order powers of S do not decay rapidly and may contribute substantially to the matrix power-sum. Consequently, the path-prohibition penalty

$$L_{\mathrm{sum}}(S;O) = \sum_{(v_i,v_j)\in O^+} \big[P_{\mathrm{sum}}(S)\big]_{ji}$$

can be inflated by repeated-walk accumulation on cyclic summary graphs, which may destabilize optimization.

Proof of Example 1. *Example 1.* Under the same partial order prior and the same forbidden reachability relation, introducing a directed cycle can strictly increase the matrix power-sum penalty because of repeated-walk accumulation.

Proof. In the acyclic graph with edges $v_1 \to v_2$, $v_2 \to v_3$, and $v_2 \to v_4$, the only directed walk from v_1 to v_3 is

$$v_1 \to v_2 \to v_3,$$

so

$$[P_{\mathrm{sum}}(S)]_{13} = w^2.$$

After adding the extra edge $v_4 \to v_2$, the graph contains the directed cycle

$$v_2 \to v_4 \to v_2.$$

Therefore, for every integer $m \geq 0$ such that

$$2 + 2m \leq d - 1,$$

there is a directed walk

$$v_1 \to v_2 \to (v_4 \to v_2)^m \to v_3$$

from v_1 to v_3 of length $2 + 2m$, contributing

$$w^{2+2m}$$

to $[P_{\mathrm{sum}}(S)]_{13}$. Hence

$$[P_{\mathrm{sum}}(S)]_{13} = \sum_{m:\, 2+2m\leq d-1} w^{2+2m},$$

which is strictly larger than w^2 whenever at least one repeated traversal is allowed. Thus, the matrix power-sum penalty increases even though the forbidden reachability relation itself is unchanged.

C.3 Stability of the Decoupled Formulation

Proof of Proposition 4. *Proposition 4.* For any nonzero summary matrix S, the normalized summary representation

$$\widetilde{S} = \frac{\alpha}{\|S\|_F + \varepsilon}\, S, \qquad \alpha \in (0,1),\ \varepsilon > 0,$$

satisfies

$$\|\widetilde{S}\|_F < 1 \qquad \text{and} \qquad \rho(\widetilde{S}) < 1.$$

Proof. For any nonzero S,

$$\|\widetilde{S}\|_F = \frac{\alpha}{\|S\|_F + \varepsilon}\, \|S\|_F = \alpha \frac{\|S\|_F}{\|S\|_F + \varepsilon} < \alpha < 1.$$

This proves the first claim.

For the second claim, recall that the spectral radius of a matrix is upper bounded by any consistent matrix norm. In particular,

$$\rho(\widetilde{S}) \leq \|\widetilde{S}\|_2 \leq \|\widetilde{S}\|_F.$$

Since $\|\widetilde{S}\|_F < 1$, it follows that

$$\rho(\widetilde{S}) < 1.$$

This proves the proposition.

Proof of Definition 6. *Definition 6.* The closed-form connectivity characterization

$$P(\widetilde{S}) := \sum_{k=1}^{\infty} \widetilde{S}^k$$

is well-defined, converges absolutely, admits the closed form

$$P(\widetilde{S}) = (I - \widetilde{S})^{-1} - I,$$

and remains bounded even when the summary graph contains cycles.

Proof. Let

$$q := \|\widetilde{S}\|_F.$$

By Proposition 4, we have $0 \leq q < 1$. Therefore,

$$\sum_{k=1}^{\infty} \|\widetilde{S}^k\|_F \leq \sum_{k=1}^{\infty} \|\widetilde{S}\|_F^k = \sum_{k=1}^{\infty} q^k = \frac{q}{1-q} < \infty.$$

Hence the matrix series

$$\sum_{k=1}^{\infty} \widetilde{S}^k$$

converges absolutely in Frobenius norm, so

$$P(\widetilde{S}) := \sum_{k=1}^{\infty} \widetilde{S}^k$$

is well-defined.

Since $\rho(\widetilde{S}) < 1$, the Neumann series applies:

$$(I - \widetilde{S})^{-1} = \sum_{k=0}^{\infty} \widetilde{S}^k.$$

Subtracting the identity gives

$$P(\widetilde{S}) = (I - \widetilde{S})^{-1} - I.$$

Thus the closed-form connectivity characterization is well-defined.

Finally, for every entry (j, i),

$$|[P(\widetilde{S})]_{ji}| \leq \|P(\widetilde{S})\|_F \leq \sum_{k=1}^{\infty} \|\widetilde{S}^k\|_F \leq \frac{q}{1-q},$$

which is finite. Therefore, even when the summary graph contains cycles, the resulting all-order propagation remains bounded and avoids unbounded amplification from repeated walks. This proves the proposition.

Remark. Because $\widetilde{S}$ is obtained from S by multiplying by a positive scalar, it preserves the zero–nonzero pattern of S:

$$S_{ji} = 0 \quad \Longleftrightarrow \quad \widetilde{S}_{ji} = 0.$$

Hence the proposed normalization changes only the propagation magnitude, not the support of the summary graph itself.

D Discussion

This section discusses why the proposed method works, what kind of inductive bias it introduces into temporal causal structure learning, what validity requirements are needed for partial order priors, and what practical lessons can be drawn from the empirical results.

Why the method works. The core role of the proposed method is to inject summary and lag-agnostic partial order knowledge into temporal causal structure learning in a way that remains compatible with lag-aware structure estimation. Its effect does not come from specifying the exact temporal lag of each causal relation. Instead, it constrains reversed reachability on the summary graph. Once a variable is known to causally precede another, the method discourages any reversed direct or indirect causal influence on the summary graph, thereby reducing implausible temporal structures. This is particularly useful when data alone cannot reliably distinguish multiple lag-aware explanations, such as in large graphs, dense graphs, or low-sample regimes. In these cases, the partial order prior acts as a global directional bias that removes implausible reversed reachability before the learner determines the detailed instantaneous and cross-lag effects.

The essence of the formulation. The essence of the proposed method is to convert summary and lag-agnostic partial order priors into path-prohibition constraints on the summary graph, and then impose these constraints in a stable way. Since the summary graph can be non-DAG, a topological ordering is no longer naturally defined. Therefore, the method follows the causal meaning of a precedence relation and forbids reversed reachability rather than relying on an explicit ordering of variables. The main difficulty is that a direct matrix power-sum representation can be unstable on cyclic summary graphs, because repeated-walk accumulation may inflate the path-prohibition penalty. The proposed normalized summary representation addresses this issue by controlling the propagation scale, while the closed-form connectivity characterization computes all-order reachability without explicit matrix power-sum accumulation. Thus, the method works not merely because it adds prior knowledge, but because it imposes partial order knowledge in a form that remains stable under summary-level propagation.

What the method is really learning. The proposed method does not tell the model where a temporal edge must appear, nor does it determine the exact lag of a causal effect. Instead, it tells the model which reversed causal explanations should be suppressed at the variable level. This distinction is important. Edge priors are local and usually specify whether a particular dependency should exist. By contrast, partial order priors are global: they exclude a family of reversed reachability relations on the summary graph, and this exclusion indirectly shapes the learned lag-aware structure. This is why the method can improve not only prior recovery but also overall structural recovery. It does not simply force several entries in the adjacency matrices; it changes the admissible solution space by ruling out structures that violate the provided causal precedence relations.

Validity of partial order priors. It is important to distinguish cycles in the temporal summary graph from conflicts in the partial order prior. In temporal causal systems, variables may mutually influence each other at different lags, and each variable may also depend on its own past states. After lag-specific relations are collapsed into a variable-level summary graph, such effects may naturally lead to bidirectional relations, self-dependencies, and directed cycles. These cycles are allowed in the summary graph and motivate our non-DAG summary-level formulation.

By contrast, an expert-provided partial order prior is not a full temporal dependency graph. It encodes valid high-confidence precedence relations between distinct variables. By definition, such a prior contains no self-pairs or cyclic precedence conflicts. If two variables are believed to mutually influence each other, or if their precedence is uncertain, then no partial order prior should be specified for that pair. Therefore, the proposed constraint does not rule out temporal feedback in general. It only prohibits reverse reachability for variable pairs where a valid precedence relation is explicitly provided. Mutually contradictory precedence information is not a valid partial order prior. Accordingly, when constructing the transitive closure and the prior mask, self-pairs are excluded, namely $(v_i, v_i) \notin O^+$ and $M_{ii} = 0$ for all variables. This prevents valid autoregressive effects from being penalized as partial-order violations while still allowing the prior to prohibit reverse reachability between distinct variables.

Trade-offs of global normalization. The proposed Frobenius normalization controls the global propagation scale of the summary representation, but it also introduces two trade-offs. First, it couples gradients among entries of the summary matrix, since all entries share the same normalization denominator. In principle, increasing an admissible edge may enlarge $\|S\|_F$ and reduce the normalized magnitude of other entries, which could indirectly weaken the partial-order penalty. However, this does not provide a free optimization shortcut, because the original lag-aware parameters are still

optimized for data fitting and penalized by sparsity regularization. Second, in large or dense graphs, the normalized magnitude of each entry may become smaller, making the partial-order gradient weaker relative to the data-fitting objective. This reflects a trade-off between stable propagation and prior strength. In practice, it can be mitigated by tuning the prior weight β or using scale-adaptive weighting, while designing more adaptive normalization schemes remains an interesting direction for future work.

Practical recommendations. The empirical results suggest several practical recommendations. First, the prior weight should remain balanced with the data objective. If it is too small, the partial order prior contributes little; if it is too large, the model may over-adapt to the prior and sacrifice data fit. Second, the proposed method is most useful when the available prior is coarse but broadly reliable. It is not designed to replace the backbone learner or to infer exact temporal lags by itself, but to provide stable global guidance when exact lag-aware supervision is unavailable. Third, data preprocessing can affect performance. For example, centering the observed time series can be helpful under asymmetric noise, because it reduces the mismatch between the data-fitting term and the prior-induced constraint. When the residual variation is not well captured by the data loss, the optimizer may compensate by introducing other admissible but incorrect edges even if the overall directional guidance is beneficial.

Implications for temporal prior integration. More broadly, the results suggest that summary and lag-agnostic partial order knowledge can be a meaningful form of supervision for temporal causal structure learning. Exact lag-aware priors are often difficult to obtain, while variable-level precedence relations are more realistic in many applications. The proposed method shows that such coarse knowledge can still guide lag-aware structure recovery, provided that it is represented on a summary graph and imposed through a stable connectivity characterization. This suggests a general principle for temporal prior integration: high-level priors should regularize the global directional organization of the learned structure, rather than directly forcing a particular instantaneous or cross-lag realization.

E Limitations and Broader Impact

Limitations. Although the proposed method provides a principled way to incorporate summary and lag-agnostic partial order priors into temporal causal structure learning, it still has several limitations. First, its effectiveness depends on the quality and coverage of the provided priors. Highly inaccurate, conflicting, or systematically biased priors may mislead the learned structure, while overly sparse priors may provide limited guidance. Second, the proposed formulation imposes partial order priors through a summary representation. This improves tractability and stability, but also discards exact temporal lag information. As a result, the method can guide variable-level directional reachability, but cannot directly identify which instantaneous or cross-lag effects realize that reachability. Third, the method still depends on the suitability of the underlying temporal structure learner. If the backbone assumptions are substantially mismatched to the true data-generating process, prior integration alone may not be sufficient to recover the correct structure. Finally, the normalized summary propagation and closed-form connectivity characterization introduce additional computation, which may become more pronounced for larger variable sets or longer lag horizons.

Broader Impact. This work may broaden the applicability of temporal causal structure learning in domains where coarse but informative prior knowledge is available. In practical settings such as healthcare, industrial monitoring, economics, and scientific system analysis, experts may know that one variable should causally precede another even when the exact temporal lag is difficult to specify. By incorporating such knowledge in a stable and differentiable way, the proposed method may improve the reliability and interpretability of temporal structure learning, especially in low-sample or noisy regimes. It may also provide a useful interface for future systems that elicit weak structural priors from documents, expert reports, or human–AI interaction.

At the same time, broader use of prior-guided causal discovery requires caution. If the supplied priors are incorrect, incomplete, or biased, the learned temporal structure may inherit or amplify these issues. This concern is especially relevant when priors are obtained from unreliable automatic extraction systems or from historical knowledge that reflects pre-existing biases. Therefore, the proposed

method should be viewed as a tool for assisting scientific analysis and hypothesis generation, rather than as a substitute for expert judgment or a fully automated basis for high-stakes decision-making.

F Complete Experiment Results

This appendix provides a complete experimental evaluation of the proposed method. In addition to the representative results reported in the main text, we present detailed results for the linear, nonlinear, nonstationary, and real-world settings, together with ablations, robustness analyses, sensitivity studies, efficiency evaluation, and qualitative visualization. These supplementary experiments offer a more comprehensive view of the effectiveness, stability, robustness, and transferability of the proposed partial-prior formulation across different temporal learning scenarios.

Implementation details. Unless otherwise specified, we set $\beta = 0.1$, $\alpha = 0.999$, and $\epsilon = 10^{-8}$, and follow the default hyperparameters of each backbone. Methods using order priors are denoted by "-O-r", where $r \in \{0.25, 0.50, 0.75, 1.00\}$ denotes the sampling ratio from the transitive closure. Continuous adjacency matrices are binarized with threshold 0.3 for all compared methods within each backbone. All reported results are averaged over six independent datasets, and error bars or shaded regions denote standard deviations across runs. For finite power-sum variants, the truncation order is set to $d - 1$, consistent with the standard reachability characterization on a d-node graph. All experiments were conducted on an AMD Ryzen 9 7950X CPU with 32 GB RAM.

F.1 Ablation Variants and Design Rationale

To better understand where the gain of the proposed method comes from, we compare several variants that differ in summary construction and stabilization strategy. The goal of this ablation is to determine whether the improvement can already be obtained by changing the summary representation alone, or whether an explicit stabilization mechanism is necessary.

Data-Only. This baseline uses the original temporal learner without any partial order prior constraint. It serves as the reference point for evaluating whether prior integration improves structure recovery at all.

Power-Sum. This variant directly extends the static power-based formulation to temporal summary graphs. It constructs the summary representation by aggregating all lag-specific structures through element-wise summation,

$$S_{\mathrm{sum}} = \sum_{\tau=0}^{p} |W^{\tau}|,$$

which preserves the total strength of lag-aware influences at the variable level. The corresponding partial order prior term is then built from the naive power-based propagation on this summary graph. This variant is used to test whether the direct extension itself is sufficient.

Decouple-Max. This variant replaces the sum-based summary with an element-wise max summary,

$$S_{\max} = \max_{\tau=0,\ldots,p} |W^{\tau}|,$$

while keeping the same power-based propagation. The motivation is that the sum-based summary may enlarge summary values by accumulating contributions from multiple lags, whereas the max-based summary keeps only the strongest lag-specific influence for each variable pair. This design tests whether a more conservative summary construction alone can alleviate instability.

Decouple-Mean. This variant uses an element-wise mean summary,

$$S_{\mathrm{mean}} = \frac{1}{p+1} \sum_{\tau=0}^{p} |W^{\tau}|,$$

again with the same power-based propagation. Compared with the max-based summary, the mean-based summary still reflects contributions from all lags, but rescales their overall magnitude through averaging. This variant is designed to test whether a smoother and less aggressive aggregation rule is sufficient to improve stability.

Decouple-ProbNorm. This variant returns to the sum-based summary, but instead of using the proposed normalization, it applies an element-wise probabilistic normalization,

$$\widehat{S}_{ij} = \frac{\exp\big((S_{\mathrm{sum}})_{ij}\big)}{\sum_v \exp\big((S_{\mathrm{sum}})_{iv}\big)}.$$

The motivation is that softmax normalization is a natural way to compress matrix values into a bounded range. This variant is therefore used to test whether a generic probabilistic normalization can stabilize propagation as effectively as the proposed formulation.

Decouple-Norm (Ours). This is the proposed method. It first constructs the sum-based summary

$$S_{\mathrm{sum}} = \sum_{\tau=0}^{p} |W^{\tau}|,$$

because this form preserves the most complete variable-level information from all lags. It then introduces the normalized propagation

$$\widetilde{S} = \alpha \frac{S_{\mathrm{sum}}}{\|S_{\mathrm{sum}}\|_F + \varepsilon},$$

and builds the partial prior term from the resulting stabilized connectivity characterization. Unlike the previous variants, this design does not rely only on changing the summary operator or heuristically shrinking matrix values. Instead, it explicitly combines informative summary construction with bounded all-partial order propagation. This variant is used to test whether the advantage of the proposed method comes from the coordinated design of summary construction and stabilization.

Taken together, these variants allow us to separately evaluate three factors: how the summary graph is constructed, how propagation is performed on it, and whether an explicit stabilization mechanism is introduced. This makes it possible to identify which design choices are truly responsible for the observed performance gain.

F.2 Ablation on Summary Design and Stabilization

Based on the variants introduced in Section F.1, we examine whether the gain of the proposed method comes from summary construction alone or from the combination of summary design and stabilization. We conduct this ablation on the linear DYNOTEARS backbone at $T = 200$, using ER1 and ER2 graphs under both Gaussian and exponential noise, and varying the prior ratio from 25% to 100%. We compare DATA-ONLY, DECOUPLE-MAX, DECOUPLE-MEAN, DECOUPLE-PROBNORM, and the proposed DECOUPLE-NORM. The results in Table 2 report Scaled SHD and F1, which together reflect structural error and recovery quality.

Table 2 shows that changing the summary operator alone is not sufficient to resolve the instability caused by cyclic propagation. DECOUPLE-MAX and DECOUPLE-MEAN can improve over the data-only baseline in many settings, indicating that summary-level prior integration is beneficial. However, their gains remain limited and inconsistent across priors, graph densities, and noise types, because both variants still rely on the same power-based propagation and are therefore still affected by repeated-walk accumulation on cyclic summary graphs. DECOUPLE-PROBNORM further introduces bounded normalization, but its softmax-style transformation changes the relative connectivity pattern row-wise and is therefore less compatible with summary-level reachability than the proposed design. By contrast, DECOUPLE-NORM is consistently the strongest method in almost all settings. Its advantage comes from combining two effects: the sum-based summary preserves the most complete variable-level information from all lags, while the proposed normalization explicitly controls all-order propagation and prevents the instability caused by repeated walks. As a result, it achieves lower Scaled SHD and higher F1 more reliably than the other variants, especially at higher prior ratios where stable use of prior information becomes more important. These results suggest that the gain of the proposed method comes from the coordinated design of informative summary construction and explicit stabilization, rather than from a single heuristic modification.

F.3 Stability Diagnosis of Constraint Variants

To further diagnose the optimization behavior of different partial-order constraint formulations, we conduct a stability statistics experiment on a representative temporal structure learning setting.

Table 2: Ablation results on summary design and stabilization. Bold indicates the best result within each prior-ratio column.

Metric	Noise	Algorithm	ER1				ER2			
			0.25	0.50	0.75	1.00	0.25	0.50	0.75	1.00
Scaled SHD	Gauss	DYNOTEARS	0.642	0.642	0.642	0.642	0.458	0.458	0.458	0.458
		DYNOTEARS-O-mean	0.550	0.567	0.442	0.408	0.433	0.446	0.404	0.400
		DYNOTEARS-O-max	0.600	0.625	0.608	0.558	0.446	0.471	0.421	0.417
		DYNOTEARS-O-prob	0.617	0.633	0.642	0.667	0.450	0.454	0.471	0.483
		DYNOTEARS-O	**0.492**	**0.500**	**0.350**	**0.375**	**0.413**	**0.417**	**0.383**	**0.371**
	Exp	DYNOTEARS	0.758	0.758	0.758	0.758	0.563	0.563	0.563	0.563
		DYNOTEARS-O-mean	0.608	0.517	0.450	0.400	0.500	0.483	0.454	0.458
		DYNOTEARS-O-max	0.592	0.617	0.633	0.542	**0.475**	0.546	0.529	0.517
		DYNOTEARS-O-prob	0.783	0.758	0.767	0.750	0.525	0.554	0.529	0.512
		DYNOTEARS-O	**0.533**	**0.492**	**0.358**	**0.333**	0.488	**0.467**	**0.442**	**0.454**
F1	Gauss	DYNOTEARS	0.685	0.685	0.685	0.685	0.757	0.757	0.757	0.757
		DYNOTEARS-O-mean	0.714	0.707	0.761	0.772	0.769	0.761	0.781	0.783
		DYNOTEARS-O-max	0.656	0.633	0.653	0.655	0.750	0.741	0.763	0.764
		DYNOTEARS-O-prob	0.694	0.691	0.696	0.690	0.762	0.763	0.758	0.755
		DYNOTEARS-O	**0.734**	**0.732**	**0.805**	**0.787**	**0.778**	**0.773**	**0.792**	**0.797**
	Exp	DYNOTEARS	0.638	0.638	0.638	0.638	0.707	0.707	0.707	0.707
		DYNOTEARS-O-mean	0.700	0.731	0.761	0.782	0.736	0.746	0.759	0.756
		DYNOTEARS-O-max	0.668	0.630	0.634	0.672	**0.743**	0.687	0.706	0.721
		DYNOTEARS-O-prob	0.638	0.648	0.651	0.666	0.726	0.718	0.729	0.741
		DYNOTEARS-O	**0.727**	**0.741**	**0.804**	**0.810**	0.740	**0.753**	**0.762**	**0.759**

Specifically, we use 20-node ER graphs with an edge-to-node ratio of 2, generate 200 samples under Gaussian noise, and provide 100% partial order priors. Each method is repeated over 12 independent runs under the same backbone and experimental setting. We compare POWER-SUM, DECOUPLE-MAX, DECOUPLE-MEAN, DECOUPLE-PROBNORM, and the proposed DECOUPLE-NORM. This experiment is not intended to show that other variants cannot be optimized, but to examine whether their optimization trajectories are more prone to unstable penalty amplification. We record invalid runs, exploding-penalty runs, the maximum partial-order penalty, and the standard deviation of the partial-order penalty across runs. An invalid run refers to a run that produces NaN or Inf values in the objective, parameters, or constraint values. An exploding-penalty run refers to a run in which the partial-order penalty exceeds the threshold $\mathcal{L}_{\mathrm{ord}} > 50$ at any optimization step.

Table 3: Stability statistics of different partial-order constraint variants. Invalid runs indicate NaN or Inf values. Exploding-penalty runs indicate cases where the partial-order penalty exceeds $\mathcal{L}_{\mathrm{ord}} > 50$ during optimization.

Method	**Invalid Runs**	**Exploding Penalty Runs**	**Global Max Penalty**	**Std. of Peak Penalty**
POWER-SUM	4/12	8/12	2453.82	842.15
DECOUPLE-MAX	0/12	6/12	116.44	32.31
DECOUPLE-MEAN	0/12	5/12	92.18	21.05
DECOUPLE-PROBNORM	0/12	1/12	54.72	22.18
DECOUPLE-NORM (Ours)	**0/12**	**0/12**	**1.25**	**0.14**

Table 3 reports the optimization stability statistics of different variants. These results should be interpreted together with the structural recovery ablation: most variants can complete optimization in standard settings, but their optimization trajectories can differ substantially in penalty scale and numerical stability. The direct POWER-SUM extension produces the largest and most variable partial-order penalties, because it applies power-based reachability propagation to the raw summary matrix without controlling the propagation scale. DECOUPLE-MAX and DECOUPLE-MEAN reduce the summary magnitude through alternative aggregation operators, but they do not provide a spectral control guarantee. DECOUPLE-PROBNORM bounds individual entries, but entry-wise boundedness does not necessarily stabilize all-order reachability. In contrast, DECOUPLE-NORM explicitly controls the propagation scale and shows no exploding-penalty cases under this setting. This diagnostic experiment complements the structural recovery ablation by showing that the advantage of the proposed method lies not only in final structure recovery, but also in more stable summary-level propagation.

The stability difference among these variants can be explained by whether the propagation scale of the summary representation is explicitly controlled. POWER-SUM applies the reachability penalty to the raw summary matrix S, so the magnitude of higher-order terms depends directly on the scale of S and the cyclic structure of the summary graph. Even when optimization can finish, repeated walks on cycles may still amplify the path-prohibition penalty and lead to larger penalty fluctuations. DECOUPLE-MAX and DECOUPLE-MEAN reduce the summary magnitude through alternative aggregation operators, but this reduction is heuristic and does not provide a spectral control guarantee. DECOUPLE-PROBNORM bounds individual entries, but bounded entries alone do not ensure that the spectral radius is smaller than one. By contrast, DECOUPLE-NORM constructs $\widetilde{S} = \alpha S/(\|S\|_F + \epsilon)$ with $\alpha \in (0, 1)$, which gives $\|\widetilde{S}\|_F < 1$ and therefore $\rho(\widetilde{S}) < 1$. As a result, the infinite connectivity series $\sum_{k=1}^{\infty} \widetilde{S}^k$ is well-defined and admits the closed form $(I - \widetilde{S})^{-1} - I$. This explains why DECOUPLE-NORM can stabilize summary-level propagation and reduce penalty amplification, rather than merely improving optimization success rates.

F.4 Robustness to Incorrect Partial Order Priors

This experiment evaluates whether the proposed method remains reliable when the provided partial order priors are partially incorrect. We use linear temporal synthetic data on 20-node ER1 and ER2 graphs under both Gaussian and exponential noise, with sample size fixed at $T = 200$. The prior ratio varies from 25% to 100%, and the prior error ratio varies from 10% to 60%. For each setting, incorrect priors are generated by randomly replacing the same proportion of true partial order relations with wrong ones. We compare the proposed method with the data-only baseline using Scaled SHD, F1 score, and prior recovery rate.

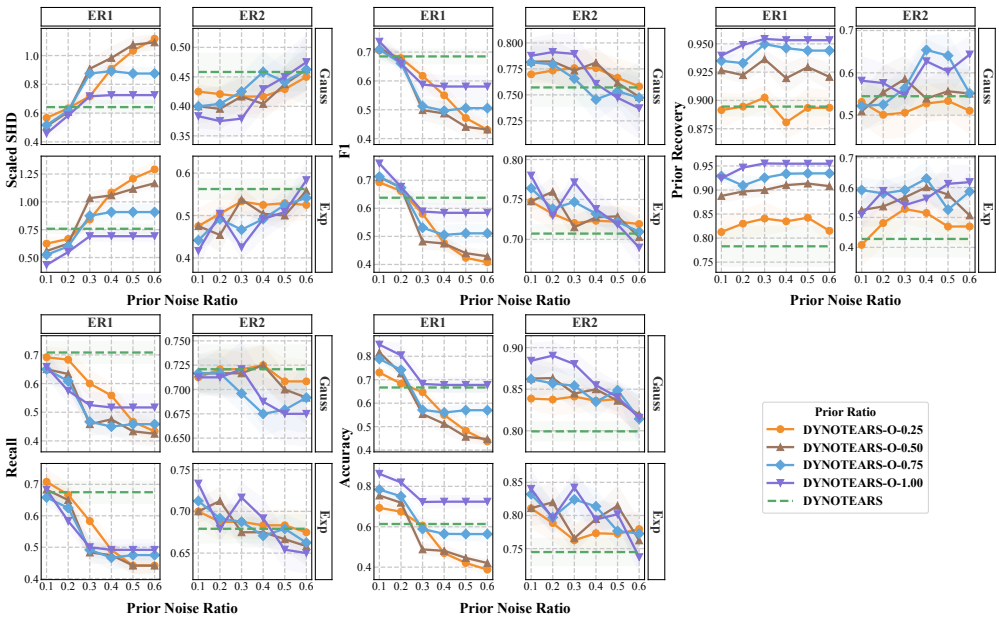

Figure 5: Robustness to incorrect partial order priors.

Results are shown in Fig. 5. The proposed method remains effective under moderate prior error, and still outperforms the data-only baseline in many settings. As the prior error ratio increases, performance degrades gradually rather than collapsing abruptly, and this trend is consistent across graph densities, noise types, and prior ratios. In particular, the method is more robust when the prior ratio is moderate and the error level is low to medium, showing that it can still extract useful guidance from imperfect partial order knowledge. These results suggest that the proposed formulation is not overly sensitive to moderate prior inaccuracies and remains practically useful when available priors are only partially reliable.

F.5 Comparison with Edge Priors

This experiment compares partial order priors and edge priors under a matched information setting. We use linear temporal synthetic data on 20-node ER1 and ER2 graphs under both Gaussian and exponential noise, with sample size fixed at $T = 200$. The prior ratio varies from 25% to 100%. We first sample variable-level partial order relations from the ground-truth structure and compute their transitive closure O^+. For the proposed method, these priors are imposed as partial order constraints on the summary graph. For the edge-prior baseline, the same O^+ is converted into summary-level direct edge prohibitions, so that only reversed direct edges are penalized. We compare the two methods using Scaled SHD, F1 score, and prior recovery rate.

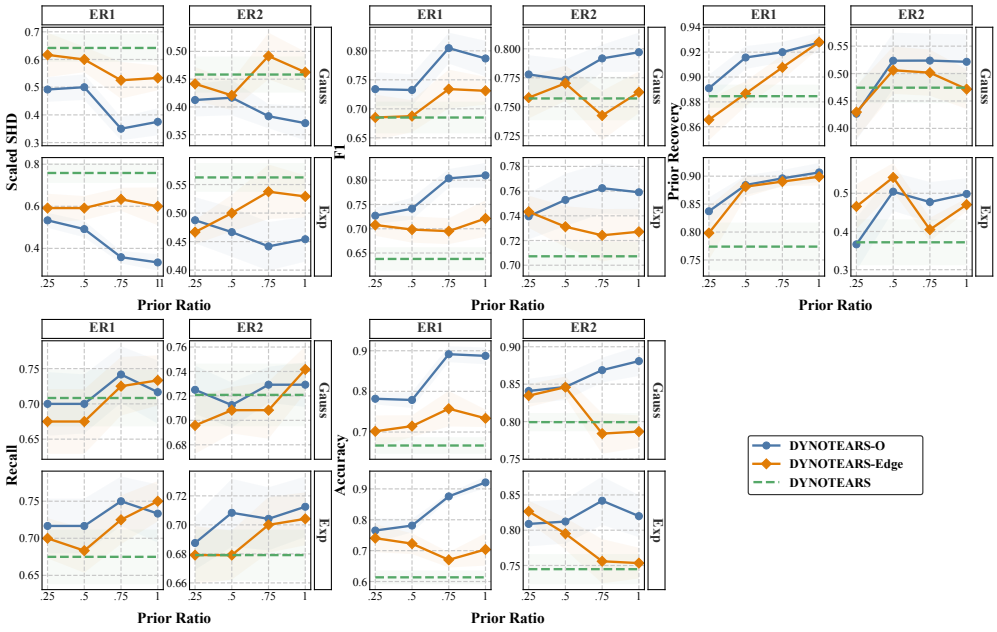

Figure 6: Comparison between partial order priors and edge priors.

Results are shown in Fig. 6. The proposed partial order prior formulation consistently outperforms the matched edge-prior baseline across most settings. Although both methods use the same variable-level prior information, the edge-prior baseline only constrains reversed direct edges and cannot capture violations arising through indirect paths. In contrast, the proposed method enforces the same priors at the path level, which better reflects their transitive and global structural meaning. As a result, it achieves lower Scaled SHD, higher F1, and better prior recovery more consistently across graph densities, noise types, and prior ratios. These results suggest that the advantage of partial order priors comes from their ability to encode global reachability rather than from receiving more prior information.

F.6 Computational Cost and Scalability

This experiment evaluates the additional runtime introduced by the proposed partial order prior formulation and whether it remains manageable as the problem size grows. We use linear temporal synthetic data on ER1 and ER2 graphs with Gaussian noise, sample size fixed at $T = 200$, prior ratio fixed at 100%, and vary the number of variables as $d \in \{10, 30, 50\}$ and the maximum lag as $p \in \{3, 5, 7\}$. We compare three variants on the same DYNOTEARS backbone: the data-only baseline, the proposed method with closed-form propagation, and a finite-series variant that replaces the closed-form connectivity characterization with explicit truncated power accumulation.

Table 4 shows that the proposed method introduces additional runtime over the data-only baseline, as expected, because prior integration requires summary-level propagation beyond the original temporal learner. However, the closed-form implementation is consistently much more efficient than the finite-series variant, especially in larger graphs, where explicit power accumulation becomes

Table 4: Runtime comparison. Entries report mean runtime (s) over random seeds. Bold indicates the fastest method under each setting.

Graph	Algorithm	$p = 3$			$p = 5$			$p = 7$		
		$d = 10$	$d = 30$	$d = 50$	$d = 10$	$d = 30$	$d = 50$	$d = 10$	$d = 30$	$d = 50$
ER1	DYNOTEARS	**2.59**	**17.36**	**84.45**	**2.56**	**29.50**	**82.79**	**1.98**	**29.93**	**84.84**
	DYNOTEARS-O	8.53	19.89	498.54	9.27	30.79	580.44	5.13	32.65	477.04
	DYNOTEARS-O-Finite	9.95	30.68	4478.75	7.35	39.65	2907.73	5.71	41.23	2428.34
ER2	DYNOTEARS	**2.21**	**24.19**	**119.40**	**2.22**	**34.46**	**110.38**	**2.08**	**36.69**	**86.42**
	DYNOTEARS-O	8.33	31.11	525.50	7.92	49.84	462.95	8.01	47.20	554.44
	DYNOTEARS-O-Finite	8.91	56.49	3774.94	8.83	54.95	3401.96	9.15	65.57	3016.81

prohibitively expensive. This difference is particularly clear when $d = 50$, where the finite-series variant often incurs several times higher runtime than DYNOTEARS-O. The proposed method therefore benefits not only from a more stable formulation, but also from a more practical computational design. Although runtime increases with graph size and lag order for all prior-based methods, the closed-form variant remains substantially more scalable than the finite alternative, showing that the additional computational cost mainly comes from prior-guided propagation itself rather than from the specific implementation used in our method.

F.7 Qualitative Visualization of Learned Temporal Structures

This experiment provides a qualitative case study of the learned temporal structures. We use a linear synthetic example with 5 variables, maximum lag 3, Gaussian noise, an ER4 graph, sample size $T = 200$, and prior ratio fixed at 100%. Figure 7 visualizes the prior mask, the summary graph, and the lag-specific matrices W^0, W^1, W^2, and W^3 for three cases: the data-only baseline DYNOTEARS, the ground truth, and the proposed method. This setting is intended to provide an intuitive view of how the partial order prior changes the learned temporal structure under the same data condition.

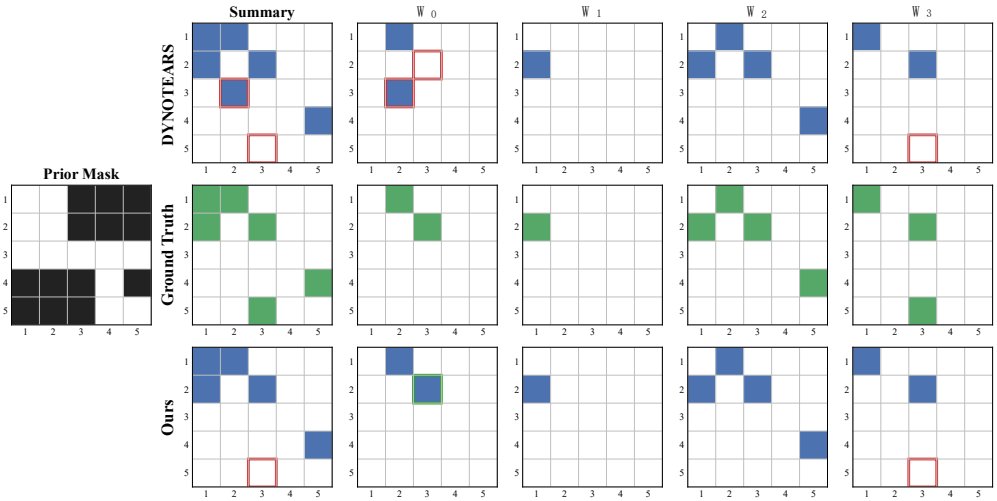

Figure 7: Qualitative visualization of learned temporal structures. From left to right are the prior mask, the summary graph, and the lag-specific matrices W^0, W^1, W^2, and W^3. From top to bottom are DYNOTEARS, the ground truth, and the proposed method.

As shown in Fig. 7, the learned structure of DYNOTEARS contains several spurious entries, including reversed effects that are inconsistent with the prior mask, and these errors further appear in the summary graph. By contrast, the proposed method produces a summary structure that is visibly closer to the ground truth and yields cleaner lag-specific matrices. In particular, it suppresses order-violating reversed dependencies while preserving the true causal pattern more faithfully across

different lags. This behavior is consistent with the mechanism of the proposed method: the partial order prior is enforced at the summary level, so it constrains global reversed reachability rather than only isolated edges, and this global guidance is then reflected back into the lag-specific matrices. Although this experiment is qualitative, it gives an intuitive illustration that the proposed partial order constraint improves both structural plausibility and interpretability.

F.8 Sensitivity to Initialization Scale

This experiment examines whether the proposed method is sensitive to the scale of parameter initialization. We use linear temporal synthetic data on 20-node ER1 and ER2 graphs under both Gaussian and exponential noise, with sample size fixed at $T = 200$ and prior ratios from 25% to 100%. Model parameters are initialized from uniform distributions with ranges $[0, 0.1]$, $[0, 0.5]$, $[0, 1.0]$, $[0, 1.5]$, $[0, 2.0]$, $[0, 2.5]$, and $[0, 3.0]$. We compare the proposed method with the data-only baseline using Scaled SHD, F1 score, and prior recovery rate.

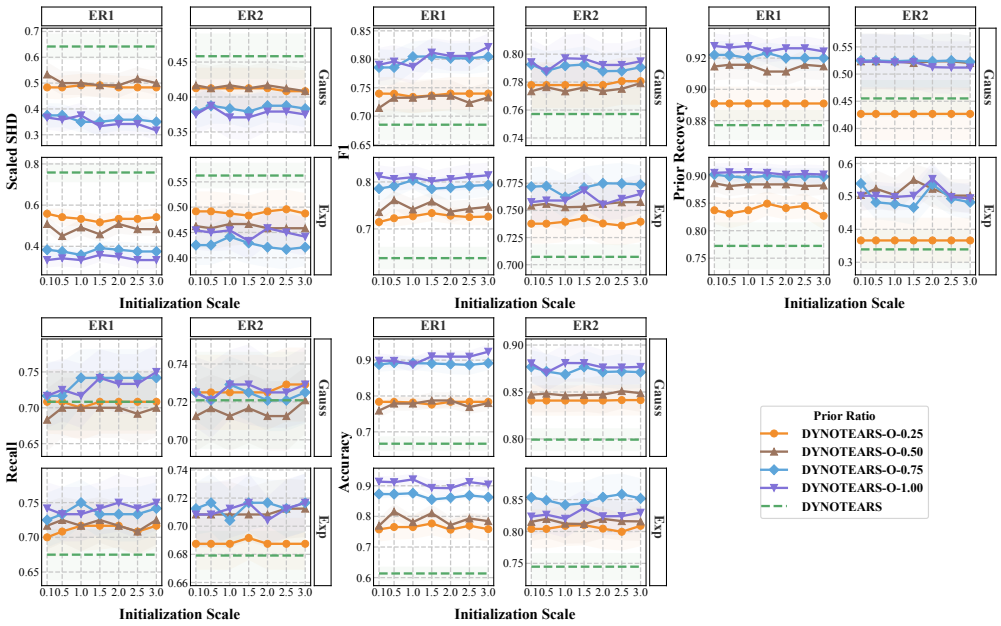

Figure 8: Sensitivity to initialization scale.

Results are shown in Fig. 8. The proposed method remains effective across a broad range of initialization scales. Although larger initialization ranges generally make optimization harder, the proposed method consistently maintains better structure recovery than the data-only baseline and degrades more gradually as the scale increases. This suggests that the proposed formulation is not overly sensitive to initialization and retains its advantage under substantially different optimization starting points.

F.9 Sensitivity to the partial order prior Weight

This experiment studies the sensitivity of the proposed method to the weight of the partial order prior penalty. We use linear temporal synthetic data on 20-node ER1 and ER2 graphs under both Gaussian and exponential noise, with sample size fixed at $T = 200$ and prior ratios from 25% to 100%. The prior weight is varied over $\beta \in \{0.001, 0.01, 0.1, 1, 10\}$. We evaluate performance using Scaled SHD, F1 score, and prior recovery rate.

Results are shown in Fig. 9. The proposed method performs well over a broad intermediate range of prior weights, especially around $\beta = 0.1$, where it consistently improves structure recovery over the data-only baseline. When β is too small, the prior contributes only weakly and the gain is limited. When β becomes too large, the prior term starts to dominate data fitting, and performance deteriorates; in particular, $\beta = 10$ is often worse than the data-only baseline. These results show

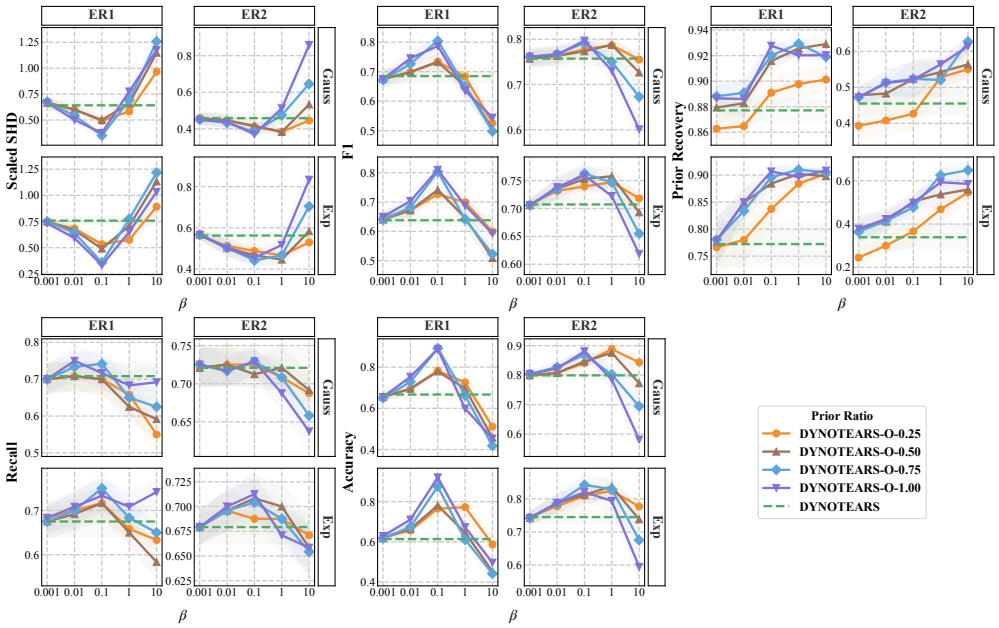

Figure 9: Sensitivity to the partial order prior weight.

that the proposed method is reasonably robust to the choice of prior weight, while still requiring the prior term to remain balanced with the data objective.

F.10 Sensitivity to the Scaling Coefficient

This experiment evaluates the effect of the scaling coefficient in the normalization step. We use linear temporal synthetic data on 20-node ER1 and ER2 graphs under both Gaussian and exponential noise, with sample size fixed at $T = 200$ and prior ratios from 25% to 100%. The scaling coefficient is varied over $\alpha \in \{0.3, 0.5, 0.7, 0.9, 0.999\}$. We evaluate performance using Scaled SHD, F1 score, and prior recovery rate.

Results are shown in Fig. 10. The proposed method remains effective across a broad range of scaling coefficients. Different values lead to different levels of contraction in the normalized summary propagation, but the overall structure recovery remains stable in most settings. In particular, intermediate and relatively large values generally yield similar performance, indicating that the normalization is not effective only under a narrowly tuned parameter choice. These results suggest that the proposed method is reasonably robust to the scaling coefficient in practical use.

F.11 Sensitivity to Prior Ratio

This experiment examines how the performance of the proposed method changes with the amount of available partial order prior knowledge. We use linear temporal synthetic data on 20-node ER1 and ER2 graphs under both Gaussian and exponential noise, with sample size fixed at $T = 200$. The prior ratio is varied over $\{0.1, 0.2, \ldots, 1.0\}$. We compare the proposed method with the data-only baseline using Scaled SHD, F1 score, and prior recovery rate.

Results are shown in Fig. 11. The proposed method benefits consistently from increasing prior ratios. Even at relatively low prior coverage, it already improves over the data-only baseline, indicating that a small amount of partial order information can provide useful guidance. As the prior ratio increases, Scaled SHD generally decreases and F1 increases, with the gain becoming less pronounced at high prior ratios. These results show that the proposed method can effectively use sparse priors while continuing to benefit from richer partial order information.

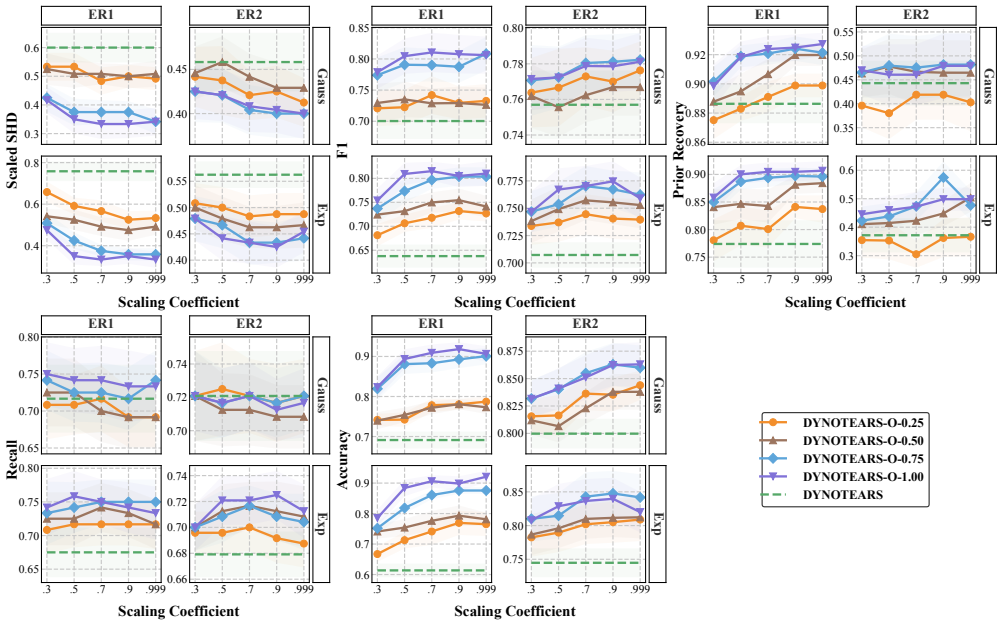

Figure 10: Sensitivity to the scaling coefficient.

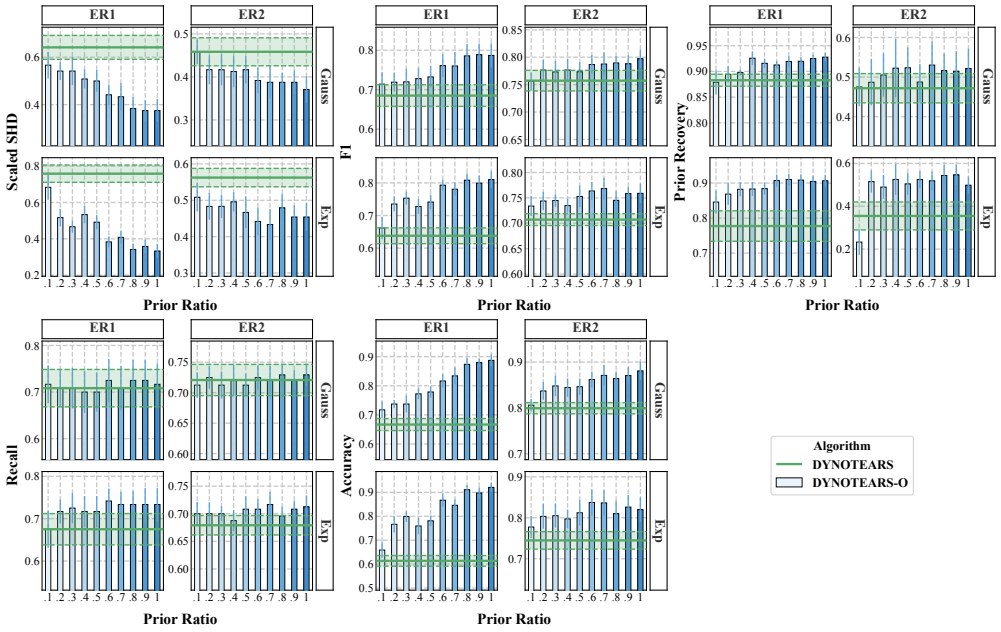

Figure 11: Sensitivity to prior ratio.

F.12 Sensitivity to the Maximum Lag

This experiment evaluates whether the proposed method remains effective under different choices of the maximum lag. We use linear temporal synthetic data on 20-node ER1 and ER2 graphs under both Gaussian and exponential noise, with sample size fixed at $T = 200$ and prior ratios from 25% to 100%. The maximum lag is varied from 1 to 10. We compare the proposed method with the data-only baseline using Scaled SHD, F1 score, and prior recovery rate.

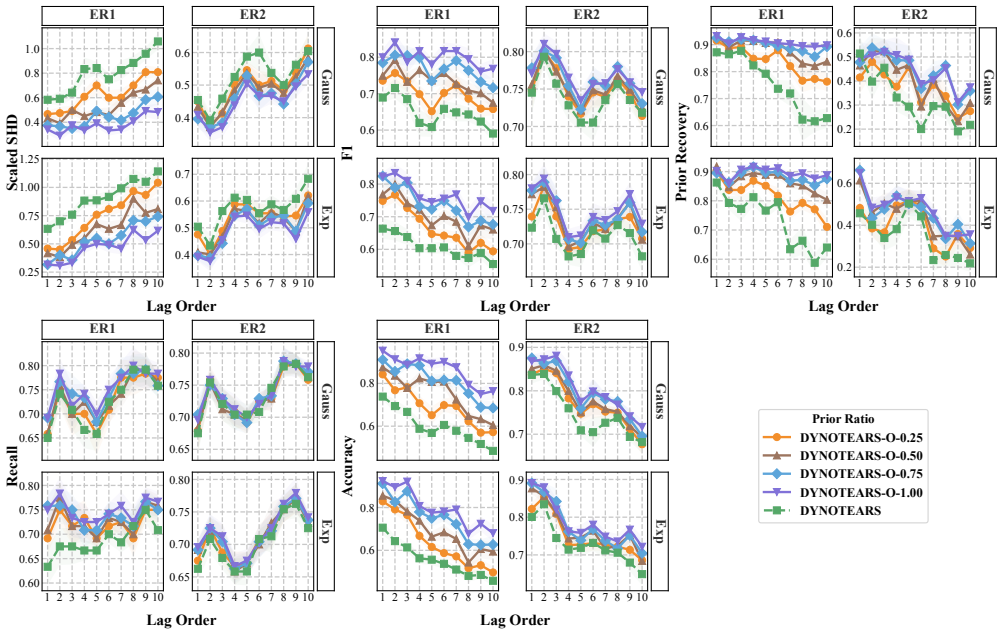

Figure 12: Sensitivity to the maximum lag.

Results are shown in Fig. 12. The proposed method remains effective across different lag choices. Although the absolute performance changes with the maximum lag, the prior-augmented model maintains clear advantages over the data-only baseline in most settings. This indicates that the benefit of the proposed formulation is not tied to a narrowly selected lag value, but persists under a broad range of temporal modeling choices.

F.13 Performance on Larger Graphs

To further evaluate whether the proposed method remains effective on larger temporal systems, we conduct an additional experiment on Gaussian synthetic data generated from ER1 graphs. The number of variables is varied over $\{100, 150, 200\}$, and the prior ratio is set to $\{0.0, 0.25, 0.5, 0.75, 1.0\}$. The remaining settings follow the default linear temporal structure learning configuration. We compare the proposed method with the corresponding data-only baseline using Scaled SHD, F1 score.

Table 5: Results on larger graphs under different prior ratios. Bold indicates the best value within each node setting.

Metric	Method	Number of Nodes		
		100	150	200
Scaled SHD	DYNOTEARS	1.485 ± 0.095	1.593 ± 0.217	1.632 ± 0.025
	DYNOTEARS-O-0.25	1.415 ± 0.015	1.530 ± 0.153	1.610 ± 0.090
	DYNOTEARS-O-0.50	1.275 ± 0.025	1.510 ± 0.190	1.578 ± 0.068
	DYNOTEARS-O-0.75	1.020 ± 0.070	1.447 ± 0.213	1.545 ± 0.047
	DYNOTEARS-O-1.00	$\mathbf{0.880 \pm 0.060}$	$\mathbf{1.273 \pm 0.113}$	$\mathbf{1.522 \pm 0.032}$
F1	DYNOTEARS	0.473 ± 0.007	0.465 ± 0.047	0.444 ± 0.003
	DYNOTEARS-O-0.25	0.492 ± 0.006	0.459 ± 0.031	0.454 ± 0.019
	DYNOTEARS-O-0.50	0.528 ± 0.017	0.479 ± 0.039	0.462 ± 0.014
	DYNOTEARS-O-0.75	0.591 ± 0.005	0.504 ± 0.044	0.477 ± 0.011
	DYNOTEARS-O-1.00	$\mathbf{0.637 \pm 0.013}$	$\mathbf{0.539 \pm 0.032}$	$\mathbf{0.486 \pm 0.014}$

Results in Table 5 show that the proposed method remains effective on substantially larger graphs. Across 100 to 200 nodes, increasing the prior ratio generally leads to lower Scaled SHD and higher F1, indicating that partial order priors continue to provide useful guidance as graph size grows.

Although temporal structure learning becomes more difficult on larger graphs for all methods, the proposed method maintains clear advantages over the data-only baseline, especially at higher prior ratios. These results support the effectiveness of the proposed formulation in large-node temporal systems.

F.14 Additional Analyses for the Main Experimental Settings

To complement the main experiments in the paper, we provide more detailed results for all core settings, including linear, nonlinear, nonstationary, and real-world temporal structure learning. These supplementary experiments are intended to refine the conclusions in the main text by showing how the proposed method behaves under different graph structures, noise types, sample sizes, and prior ratios.

Linear setting: complete results. For the linear temporal structure learning setting, we report the full results on ER1 and ER2 graphs under Gaussian and exponential noise, with sample size fixed at $T = 200$ and the number of variables ranging from 10 to 50. The prior ratio varies over $\{0.25, 0.5, 0.75, 1.0\}$, and we compare the proposed method with the data-only baseline using Scaled SHD, F1 score, prior recovery rate, accuracy, and recall. As shown in Fig. 13, the proposed method consistently improves structure recovery across different graph sizes, graph densities, and noise types. The improvement is especially clear when the graph becomes larger or denser, where purely data-driven learning is more difficult. These detailed results further support the conclusion that the proposed partial order constraint can effectively exploit partial order priors in standard linear temporal discovery.

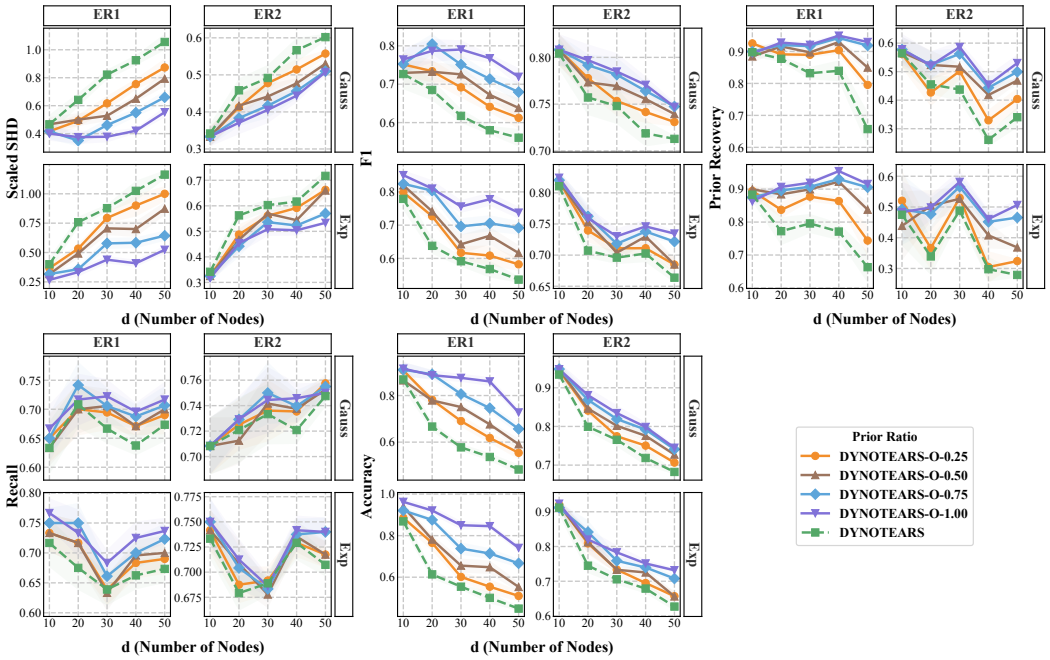

Figure 13: Complete results for linear temporal structure learning under different graph sizes, graph densities, noise types, and prior ratios.

Nonlinear setting: complete results. For the nonlinear temporal structure learning setting, we report complete results for NTS-NOTEARS under both Additive Noise Model (ANM) and Additive Index Model (AIM) mechanisms. The experiments are conducted on 20-node ER1 and ER2 graphs with sample sizes $T \in \{200, 500\}$, and the prior ratio varies over $\{0.0, 0.25, 0.5, 0.75, 1.0\}$. We evaluate Scaled SHD, F1 score, prior recovery rate, precision, and recall. As shown in Fig. 14, the proposed method remains beneficial across both nonlinear mechanisms and both sample sizes. In particular, as the prior ratio increases, Scaled SHD generally decreases while F1 and prior recovery

improve, indicating that the gain of partial order prior integration is not restricted to linear models. These results strengthen the conclusion that the proposed formulation can serve as a general mechanism for incorporating partial order priors into nonlinear temporal structure learning.

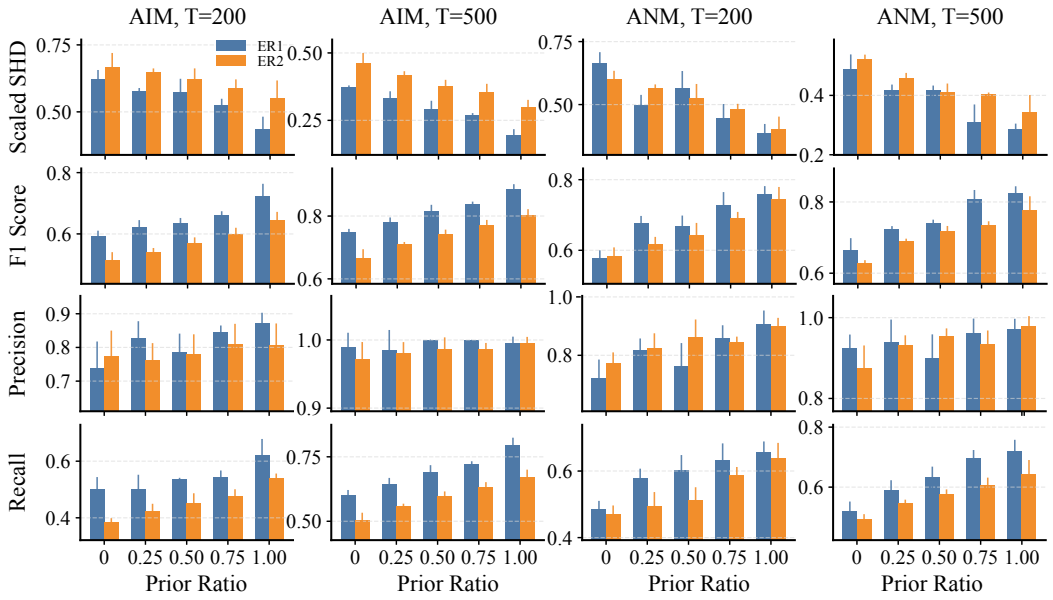

Figure 14: Complete results for nonlinear temporal structure learning with NTS-NOTEARS under ANM and AIM mechanisms, different graph densities, sample sizes, and prior ratios.

Nonstationary setting: complete results. For the nonstationary setting, we provide complete results separately for LIN and RHINO. For LIN, we report experiments on BA and ER graphs with 5 nodes and sample sizes 1000 and 5000, under prior ratios $\{0.0, 0.25, 0.5, 0.75, 1.0\}$. The evaluation metrics are SHD, F1 score, accuracy, and recall. For RHINO, we report experiments on ER and SF graphs with 5, 10, and 20 nodes, again under prior ratios $\{0.0, 0.25, 0.5, 0.75, 1.0\}$. Following the RHINO setting, we evaluate F1 score and Scaled SHD on three structural views: Instantaneous, Lagged, and Temporal. As shown in Fig. 15 and Fig. 16, the proposed method remains effective under different forms of nonstationarity. For LIN, larger prior ratios generally improve SHD and F1 across both graph families and both sample sizes. For RHINO, the prior-augmented variants consistently improve recovery quality across different node sizes and structural views. These results further confirm that the proposed partial order constraint can transfer to different nonstationary temporal learners rather than being tied to a particular backbone.

Real-world setting: complete results. For the real-world setting, we provide full results on the DREAM4 dataset under different prior ratios and sample sizes. We report AUROC and data loss to evaluate summary-level structural recovery and predictive fitting. Results in Table 6 show that incorporating partial order priors consistently improves summary-level structural recovery on DREAM4, as reflected by higher AUROC values. However, the effect on data loss is mixed. This is expected because the order constraint mainly guides structural recovery rather than directly optimizing predictive fit. In some cases, enforcing prior-consistent structures may sacrifice part of the regression fit, especially when the real-world prior information is coarse or only summary-level. Overall, DYNOTEARS-O achieves stronger structural recovery while maintaining reasonable data-fitting performance.

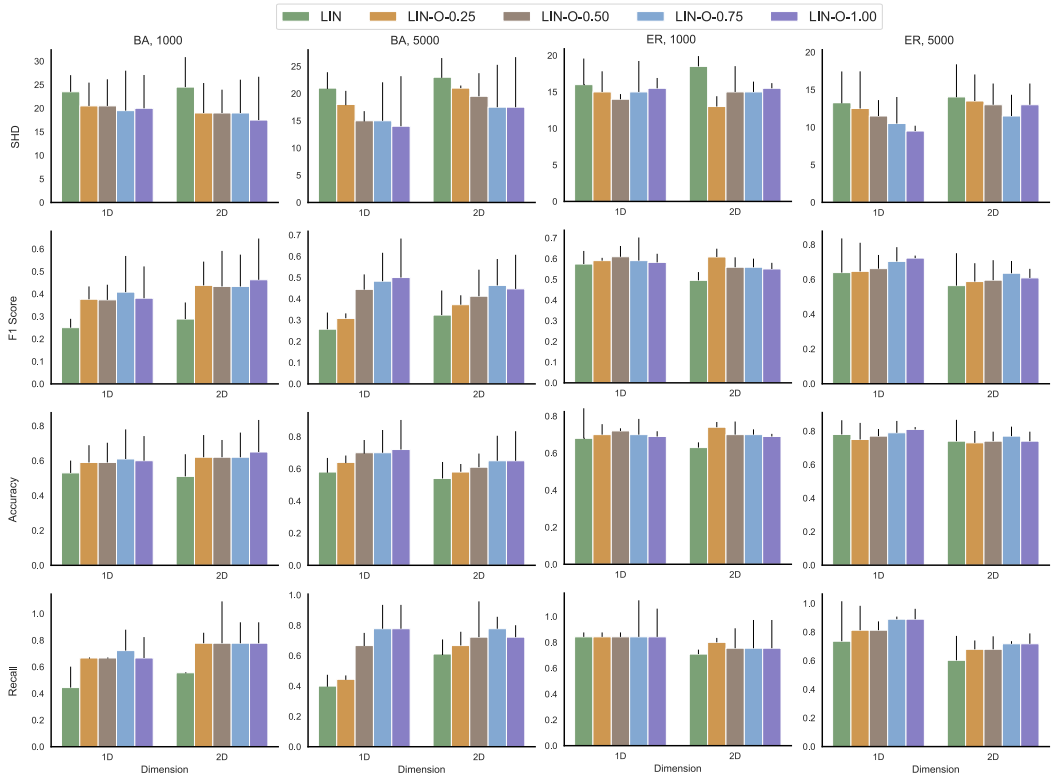

Figure 15: Complete results for nonstationary temporal structure learning with LIN under different graph types, sample sizes, and prior ratios.

Table 6: DREAM4 real-data results for partial order prior structure learning.

Setting		DREAM4-63		DREAM4-126		DREAM4-189	
Prior Rate	**Method**	**DataLoss**	**AUROC**	**DataLoss**	**AUROC**	**DataLoss**	**AUROC**
25%	NOTEARS	8.20 ± 0.73	0.54 ± 0.03	8.02 ± 0.92	0.55 ± 0.03	7.89 ± 0.73	0.55 ± 0.03
	NOTEARS-O	8.63 ± 0.58	0.58 ± 0.03	7.93 ± 0.76	0.56 ± 0.02	7.08 ± 0.85	0.56 ± 0.03
	DYNOTEARS	7.81 ± 1.14	0.58 ± 0.03	**5.61 ± 1.04**	0.62 ± 0.04	**4.08 ± 0.60**	0.65 ± 0.05
	DYNOTEARS-O-Mean	7.89 ± 1.06	0.64 ± 0.05	7.73 ± 0.89	0.64 ± 0.08	6.93 ± 1.34	0.66 ± 0.10
	DYNOTEARS-O-Max	8.40 ± 0.85	0.60 ± 0.08	7.35 ± 0.95	0.62 ± 0.11	7.61 ± 1.62	0.63 ± 0.15
	DYNOTEARS-O-Prob	9.91 ± 1.28	0.57 ± 0.04	8.88 ± 1.39	0.62 ± 0.04	7.23 ± 0.71	0.65 ± 0.06
	DYNOTEARS-O	**7.59 ± 0.61**	**0.66 ± 0.02**	6.31 ± 1.15	**0.68 ± 0.09**	6.14 ± 2.02	**0.69 ± 0.07**
50%	NOTEARS	8.20 ± 0.73	0.54 ± 0.03	8.02 ± 0.92	0.55 ± 0.03	7.89 ± 0.73	0.55 ± 0.03
	NOTEARS-O	8.62 ± 0.58	0.58 ± 0.05	7.61 ± 0.81	0.58 ± 0.05	8.08 ± 0.85	0.60 ± 0.02
	DYNOTEARS	7.81 ± 1.14	0.58 ± 0.03	**5.61 ± 1.04**	0.62 ± 0.04	**4.08 ± 0.60**	0.65 ± 0.05
	DYNOTEARS-O-Mean	7.63 ± 1.05	0.67 ± 0.03	7.73 ± 0.84	0.68 ± 0.06	6.35 ± 1.22	0.71 ± 0.08
	DYNOTEARS-O-Max	8.35 ± 0.90	0.63 ± 0.09	7.21 ± 1.01	0.66 ± 0.11	6.83 ± 1.45	0.65 ± 0.13
	DYNOTEARS-O-Prob	8.22 ± 1.88	0.58 ± 0.04	7.95 ± 1.23	0.62 ± 0.04	7.22 ± 1.04	0.66 ± 0.07
	DYNOTEARS-O	**7.46 ± 0.54**	**0.70 ± 0.04**	5.87 ± 0.89	**0.71 ± 0.03**	5.26 ± 1.14	**0.73 ± 0.06**
75%	NOTEARS	8.20 ± 0.73	0.54 ± 0.03	8.02 ± 0.92	0.55 ± 0.03	7.89 ± 0.73	0.55 ± 0.03
	NOTEARS-O	8.25 ± 0.66	0.61 ± 0.05	7.57 ± 0.85	0.61 ± 0.05	7.53 ± 0.32	0.62 ± 0.04
	DYNOTEARS	7.81 ± 1.14	0.58 ± 0.03	5.61 ± 1.04	0.62 ± 0.04	**4.08 ± 0.60**	0.65 ± 0.05
	DYNOTEARS-O-Mean	7.42 ± 0.72	0.71 ± 0.03	7.19 ± 0.97	0.70 ± 0.05	6.20 ± 1.24	0.75 ± 0.09
	DYNOTEARS-O-Max	7.78 ± 0.54	0.67 ± 0.02	6.71 ± 0.86	0.69 ± 0.09	6.79 ± 1.57	0.69 ± 0.13
	DYNOTEARS-O-Prob	8.16 ± 2.16	0.56 ± 0.04	7.25 ± 3.04	0.64 ± 0.05	6.99 ± 0.84	0.67 ± 0.07
	DYNOTEARS-O	**7.41 ± 0.65**	**0.72 ± 0.02**	**5.36 ± 0.91**	**0.73 ± 0.04**	4.71 ± 0.69	**0.76 ± 0.06**
100%	NOTEARS	8.20 ± 0.73	0.54 ± 0.03	8.02 ± 0.92	0.55 ± 0.03	7.89 ± 0.73	0.55 ± 0.03
	NOTEARS-O	8.13 ± 0.62	0.63 ± 0.03	7.14 ± 0.87	0.62 ± 0.03	7.22 ± 0.86	0.67 ± 0.05
	DYNOTEARS	7.81 ± 1.14	0.58 ± 0.03	5.61 ± 1.04	0.62 ± 0.04	**4.08 ± 0.60**	0.65 ± 0.05
	DYNOTEARS-O-Mean	7.27 ± 0.61	0.73 ± 0.04	6.58 ± 0.71	0.73 ± 0.04	5.09 ± 1.01	0.76 ± 0.05
	DYNOTEARS-O-Max	7.65 ± 0.49	0.70 ± 0.01	6.23 ± 0.85	0.69 ± 0.03	5.64 ± 0.79	0.73 ± 0.05
	DYNOTEARS-O-Prob	8.06 ± 3.09	0.60 ± 0.04	6.61 ± 1.81	0.65 ± 0.05	6.44 ± 2.15	0.68 ± 0.07
	DYNOTEARS-O	**7.24 ± 0.62**	**0.74 ± 0.03**	**5.17 ± 0.66**	**0.76 ± 0.04**	4.39 ± 0.65	**0.80 ± 0.07**

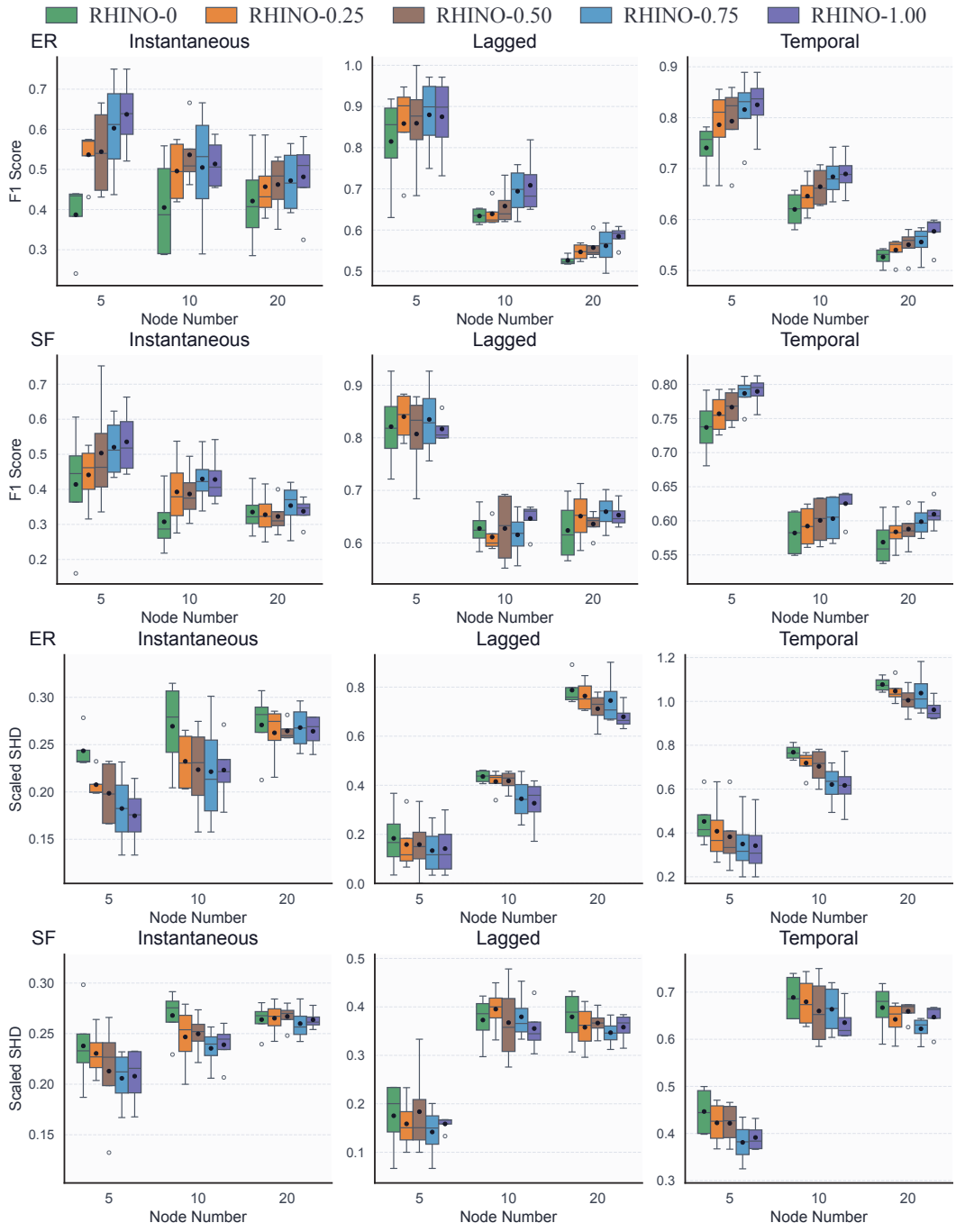

Figure 16: Complete results for nonstationary temporal structure learning with RHINO under different graph types, node sizes, structural views, and prior ratios.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made match theoretical and experimental results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are explicitly discussed in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Full set of assumptions and a complete and correct proof are provided for all theoretical results, statements, and examples.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All necessary settings and codes are available for the experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Data and codes are available in the supplementary material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All training and test details necessary to understand results are specified.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Statsitical significance are considered and reported in the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Sufficient information on the computer resources is provided in the experimental setup.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Both potential positive societal impacts and potentially negative societal impacts are discussed in Appendix E

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the creators or original owners of assets (e.g., code, data, models), used in the paper, are properly credited and the license and terms of use are explicitly mentioned and properly respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve human subjects, crowdsourcing, user studies, or personal data collection, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.