

950	<i>Supplement to</i>	
951	“Online Time Series Forecasting with Theoretical Guarantees”	
952		
953	Appendix organization:	
954	<hr/>	
955	A Related Works	23
956	A.1 Time Series Forecasting	23
957	A.2 Online Time Series Forecasting	23
958	A.3 Continual Learning	23
959	A.4 Causal Representation Learning	23
960	B Notations	24
961	C Proof	25
962	C.1 Proof of Theorem 1	25
963	C.2 Proof of Theorem 2.	26
964	C.3 More Discussion of injective linear operators	30
965	C.4 Monotonicity and Normalization Assumption	30
966	C.5 Definition of Neighborhood	30
967	C.6 More Discussion of Uniqueness of Spectral Decomposition	30
968	C.7 Proof of Theorem 3.	31
969	C.8 More Discussion on the Sparse Mixing Procedure	35
970	D Experiment Details	35
971	D.1 Synthetic Experiment	35
972	D.1.1 Data Generation Process	35
973	D.1.2 Evaluation Metric	35
974	D.1.3 Prior Likelihood Derivation	36
975	D.1.4 Evident Lower Bound	37
976	D.1.5 More Synthetic Experiment Results	37
977	D.2 Real-world Experiment	37
978	D.2.1 Dataset Description	37
979	D.2.2 Implementation Details	37
980	D.2.3 More Experiment Results	42
981	D.2.4 Experiment Results of Mean and Standard Deviation	43
982	E Broader Impacts	44
983	<hr/>	

A Related Works

A.1 Time Series Forecasting

Recent advancements in time series forecasting have been driven by the application of deep learning techniques, which have proven to be highly effective in this area. These methods can be broadly categorized into several groups. First, Recurrent Neural Networks (RNNs) are commonly used for capturing temporal dependencies by leveraging their recursive structure and memory to model hidden state transitions over time [Graves and Graves, 2012, Lai et al., 2018, Salinas et al., 2020]. Another popular approach is based on Temporal Convolutional Networks (TCNs), which employ a shared convolutional kernel to model hierarchical temporal patterns and extract relevant features [Bai et al., 2018, Wang et al., Wu et al., 2022]. Additionally, simpler yet highly effective methods, such as Multi-Layer Perceptrons (MLP) [Oreshkin et al., 2019, Zeng et al., 2023, Zhang et al., 2022, Li et al., 2024a] and state-space models [Gu et al., 2022, 2021b,a], have also been utilized in forecasting tasks. Among these, Transformer-based models have emerged as particularly noteworthy, demonstrating significant progress in the time series forecasting domain [Kitaev et al., 2020, Liu et al., 2021, Wu et al., 2021, Zhou et al., 2021]. Despite the success of these methods, they are generally designed for offline data processing, which limits their applicability to real-time, online training scenarios.

A.2 Online Time Series Forecasting

The rapid growth of training data and the need for real-time updates have made online time series forecasting more popular than offline methods [Anava et al., 2013, Liu et al., 2016, Gultekin and Paisley, 2018, Aydore et al., 2019]. Recent approaches include Pan et al. [2024], which uses structural consistency regularization and memory replay to retain temporal dependencies, and Luan et al. [2024], which applies tensor factorization for low-complexity online updates. Additionally, Mejri et al. [2024] addresses nonlinear forecasting by mapping low-dimensional series to high-dimensional spaces for better adaptation. Online forecasting is widely used in practice due to continuous data and frequent concept drift. Models are trained over multiple rounds, where they predict and incorporate new observations to refine performance. Recent work, such as [Pham et al., 2022, Cai et al., 2025, yee Ava Lau et al., 2025] and Wen et al. [2024], focuses on optimizing fast adaptation and information retention. However, simultaneously adapting to new data while retaining past knowledge can lead to suboptimal results, highlighting the need to decouple long- and short-term dependencies for improved predictions. However, most of these methods rarely explore the theoretical guarantees for online time series forecasting.

A.3 Continual Learning

Our work is also related to continual learning. Continual learning is an emerging field focused on developing intelligent systems that can sequentially learn tasks with limited access to prior experience [Lopez-Paz and Ranzato, 2017]. A key challenge in continual learning is balancing the retention of knowledge from current tasks with the flexibility to learn future tasks, known as the stability-plasticity dilemma [Lin, 1992, Grossberg, 2013]. Inspired by neuroscience, various continual learning algorithms have been developed. This approach aligns with the needs of online time series forecasting, where continuous learning allows models to update in real time as new data arrives, improving their ability to adapt to changing data dynamics and enhancing forecasting accuracy.

A.4 Causal Representation Learning

To ensure the identifiability of latent variables, Independent Component Analysis (ICA) has been widely used for causal representation learning [Yao et al., 2023, Schölkopf et al., 2021, Liu et al., 2023, Gresele et al., 2020]. Traditional ICA methods assume a linear mixing function between latent and observed variables [Comon, 1994, Hyvärinen, 2013, Lee and Lee, 1998, Zhang and Chan, 2007], but this is often impractical. To address this, studies have proposed assumptions for nonlinear ICA, such as sparse generation processes and auxiliary variables [Zheng et al., 2022, Hyvärinen and Pajunen, 1999, Hyvärinen et al., 2024, Khemakhem et al., 2020b, Li et al., 2023]. For example, Aapo et al. confirmed identifiability by assuming latent sources belong to the exponential family, with auxiliary variables like domain and time indices [Khemakhem et al., 2020a, Hyvarinen and Morioka, 2016, 2017, Hyvarinen et al., 2019]. In contrast, Zhang et al. showed that nonlinear ICA can achieve

1035 component-wise identifiability without the exponential family assumption [Kong et al., 2022, Xie
1036 et al., 2023, Kong et al., 2023, Yan et al., 2024].

1037 Other studies also use sparsity assumptions to achieve identifiability without supervised signals. For
1038 instance, Lachapelle et al. applied sparsity regularization to discover latent components [Lachapelle
1039 et al., 2023, Lachapelle and Lacoste-Julien, 2022], while Zhang et al. used sparse structures to
1040 maintain identifiability under distribution shifts [Zhang et al., 2024a]. Nonlinear ICA has been
1041 used for time series identifiability [Hyvarinen and Morioka, 2016, Yan et al., 2024, Huang et al.,
1042 2023, Hälvä and Hyvarinen, 2020, Lippe et al., 2022]. Aapo et al. used variance changes to detect
1043 nonstationary time series data identifiability, while permutation-based contrastive learning was applied
1044 for stationary time series [Hyvarinen and Morioka, 2016]. More recently, techniques like TDRL
1045 [Yao et al., 2022], LEAP [Yao et al., 2021] and IDOL [Li et al., 2025] incorporated independent
1046 noise and variability features. Additionally, Song et al. identified latent variables without domain-
1047 specific observations [Song et al., 2024], and Imant et al. used multimodal comparative learning for
1048 modality identifiability [Daunhawer et al., 2023]. Yao et al. showed that multi-perspective causal
1049 representations remain identifiable despite incomplete observations [Yao et al., 2023]. However, these
1050 methods typically assume invertibility in the mixing process. This paper relaxes that assumption and
1051 provides identifiability guarantees for online time series forecasting.

1052 B Notations

1053 This section collects the notations used in the theorem proofs for clarity and consistency.

Table A4: List of notations, explanations, and corresponding values.

Index	Explanation	Support
n	Number of variables	$n \in \mathbb{N}^+$
i, j, k, l	Index of latent variables	$i, j, k, l \in \{1, \dots, n\}$
t	Time index	$t \in \mathbb{N}^+$
Variable		
\mathcal{X}_t	Support of observed variables in time-index t	$\mathcal{X}_t \subseteq \mathbb{R}^n$
\mathcal{Z}_t	Support of latent variables	$\mathcal{Z}_t \subseteq \mathbb{R}^n$
\mathbf{x}_t	Observed variables in time-index t	$\mathbf{x}_t \in \mathbb{R}^n$
\mathbf{z}_t	Latent variables in time-index t	$\mathbf{z}_t \in \mathbb{R}^n$
\mathbf{u}_t	$\{\mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \mathbf{z}_t, \mathbf{x}_t\}$	$\mathbf{u}_t \in \mathbb{R}^{4 \times n}$
ϵ_t^o	Independent noise of mixing procedure	$\epsilon_t^o \sim p_{\epsilon_t^o}$
$\epsilon_{t,i}^z$	Independent noise of the latent transition of $\mathbf{z}_{t,i}$	$\epsilon_{t,i}^z \sim p_{\epsilon_{t,i}^z}$
Function		
$p_{a b}(\cdot b)$	Density function of a given b	/
$\mathbf{g}(\cdot)$	Nonlinear mixing function	$\mathbb{R}^{2 \times n+1} \rightarrow \mathbb{R}^n$
$f_i(\cdot)$	Transition function of $\mathbf{z}_{t,i}$	$\mathbb{R}^{n+1} \rightarrow \mathbb{R}$
$h(\cdot)$	Invertible transformation from \mathbf{z}_t to $\hat{\mathbf{z}}_t$	$\mathbb{R}^n \rightarrow \mathbb{R}^n$
$\pi(\cdot)$	Permutation function	$\mathbb{R}^n \rightarrow \mathbb{R}^n$
\mathcal{F}	Function space	/
$\mathcal{M}_{\mathbf{u}_t}$	Markov network over \mathbf{u}_t	/
ϕ_l	Encoder for $\hat{\mathbf{z}}_t^l$	/
ψ_l	Decoder	/
$r_{t,i}^z$	Noise estimator of $\hat{\epsilon}_{t,i}^z$.	/
$r_{t,i}^o$	Noise estimator of $\hat{\epsilon}_{t,i}^o$.	/
Symbol		
\mathcal{R}	Bayes Risk	/
\mathbf{J}_κ	Jacobian matrix of r_t^l	/

1054 C Proof

1055 C.1 Proof of Theorem 1

1056 **Theorem A1. (Predictive-Risk Reduction via Temporal Latent Variables)** Let $\mathbf{x}_t, \mathbf{z}_t$, and $\hat{\mathbf{z}}_t$ be
 1057 the observed variables, ground-truth latent variables, and the estimated latent variables, respec-
 1058 tively. We let $\mathbf{x}_{t-\tau:t} = \{\mathbf{x}_{t-\tau}, \dots, \mathbf{x}_t\}$ be the historical $(\tau + 1)$ -step observed variables. More-
 1059 over, we let $\mathcal{R}_o, \mathcal{R}_z$, and $\mathcal{R}_{\hat{z}}$ be the expected mean squared error for the models that consider
 1060 $\{\mathbf{x}_{t-\tau:t}\}, \{\mathbf{x}_{t-\tau:t}, \mathbf{z}_t\}$, and $\{\mathbf{x}_{t-\tau:t}, \hat{\mathbf{z}}_t\}$, respectively. Then, in general, we have $\mathcal{R}_o \geq \mathcal{R}_{\hat{z}} \geq \mathcal{R}_z$,
 1061 and if \mathbf{z}_t is identifiable we have $\mathcal{R}_o > \mathcal{R}_{\hat{z}} = \mathcal{R}_z$.

1062 *Proof.* Suppose that the observed variables \mathbf{x}_t , ground-truth latent variables \mathbf{z}_t , and estimated latent
 1063 variables $\hat{\mathbf{z}}_t$ follow the data generation process as shown in Figure 1. We let $\mathcal{F}_o := \sigma(\mathbf{x}_{t-\tau:t})$, $\mathcal{F}_z :=$
 1064 $\sigma(\mathbf{x}_{t-\tau:t}, \mathbf{z}_t)$, and $\mathcal{F}_{\hat{z}} := \sigma(\mathbf{x}_{t-\tau:t}, \hat{\mathbf{z}}_t)$ be the information σ -algebras generated by the variables
 1065 available to the forecaster in the three settings (only observed variables, observed and ground
 1066 truth latent variables, observed and the estimated latent variables). And the corresponding optimal
 1067 Bayes forecaster can be formalized as $\hat{\mathbf{x}}_{t+1}^{(o)} := \mathbb{E}[\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}]$, $\hat{\mathbf{x}}_{t+1}^{(z)} := \mathbb{E}[\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}, \mathbf{z}_t]$, and
 1068 $\hat{\mathbf{x}}_{t+1}^{(\hat{z})} := \mathbb{E}[\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}, \hat{\mathbf{z}}_t]$, respectively. Then we let $\mathcal{R}_o, \mathcal{R}_z$, and $\mathcal{R}_{\hat{z}}$ be the corresponding Bayes
 1069 risk. By using the law of total expectation, we have:

$$\begin{aligned} \mathcal{R}_o &= \mathbb{E}_{\mathbf{x}_{t+1}, \mathbf{x}_{t-\tau:t}} [(\mathbf{x}_{t+1} - \hat{\mathbf{x}}_{t+1}^{(o)})^2] = \mathbb{E}_{\mathbf{x}_{t-\tau:t}} [\mathbb{E}_{\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}} [(\mathbf{x}_{t+1} - \hat{\mathbf{x}}_{t+1}^{(o)})^2]] \\ &= \mathbb{E}_{\mathbf{x}_{t-\tau:t}} [\text{Var}(\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t})], \end{aligned} \quad (\text{A1})$$

$$\begin{aligned} \mathcal{R}_z &= \mathbb{E}_{\mathbf{x}_{t+1}, \mathbf{x}_{t-\tau:t}, \mathbf{z}_t} [(\mathbf{x}_{t+1} - \hat{\mathbf{x}}_{t+1}^{(z)})^2] = \mathbb{E}_{\mathbf{x}_{t-\tau:t}, \mathbf{z}_t} [\mathbb{E}_{\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}, \mathbf{z}_t} [(\mathbf{x}_{t+1} - \hat{\mathbf{x}}_{t+1}^{(z)})^2]] \\ &= \mathbb{E}_{\mathbf{x}_{t-\tau:t}, \mathbf{z}_t} [\text{Var}(\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}, \mathbf{z}_t)], \end{aligned} \quad (\text{A2})$$

$$\begin{aligned} \mathcal{R}_{\hat{z}} &= \mathbb{E}_{\mathbf{x}_{t+1}, \mathbf{x}_{t-\tau:t}, \hat{\mathbf{z}}_t} [(\mathbf{x}_{t+1} - \hat{\mathbf{x}}_{t+1}^{(\hat{z})})^2] = \mathbb{E}_{\mathbf{x}_{t-\tau:t}, \hat{\mathbf{z}}_t} [\mathbb{E}_{\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}, \hat{\mathbf{z}}_t} [(\mathbf{x}_{t+1} - \hat{\mathbf{x}}_{t+1}^{(\hat{z})})^2]] \\ &= \mathbb{E}_{\mathbf{x}_{t-\tau:t}, \hat{\mathbf{z}}_t} [\text{Var}(\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}, \hat{\mathbf{z}}_t)]. \end{aligned} \quad (\text{A3})$$

1072 By using the law of total variance, Equation (A1) can be written as:

$$\begin{aligned} \underbrace{\mathbb{E}_{\mathbf{x}_{t-\tau:t}} [\text{Var}(\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t})]}_{\mathcal{R}_o} &= \underbrace{\mathbb{E}_{\mathbf{x}_{t-\tau:t}, \mathbf{z}_t} [\text{Var}(\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}, \mathbf{z}_t)]}_{\mathcal{R}_z} \\ &\quad + \underbrace{\mathbb{E}_{\mathbf{x}_{t-\tau:t}} [\text{Var}_{\mathbf{z}_t | \mathbf{x}_{t-\tau:t}} (\mathbb{E}[\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}, \mathbf{z}_t])]}_c, \end{aligned} \quad (\text{A4})$$

1073 and

$$\begin{aligned} \underbrace{\mathbb{E}_{\mathbf{x}_{t-\tau:t}} [\text{Var}(\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t})]}_{\mathcal{R}_o} &= \underbrace{\mathbb{E}_{\mathbf{x}_{t-\tau:t}, \hat{\mathbf{z}}_t} [\text{Var}(\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}, \hat{\mathbf{z}}_t)]}_{\mathcal{R}_{\hat{z}}} \\ &\quad + \underbrace{\mathbb{E}_{\mathbf{x}_{t-\tau:t}} [\text{Var}_{\hat{\mathbf{z}}_t | \mathbf{x}_{t-\tau:t}} (\mathbb{E}[\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}, \hat{\mathbf{z}}_t])]}_e. \end{aligned} \quad (\text{A5})$$

1074 Suppose in Equation (A4), $c = 0$. That implies $\text{Var}_{\mathbf{z}_t | \mathbf{x}_{t-\tau:t}} (\mathbb{E}[\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}, \mathbf{z}_t]) = 0$, meaning that
 1075 \mathbf{z}_t does not have any influence on the mapping $\mathbf{x}_{t-\tau:t} \rightarrow \mathbf{x}_{t+1}$, which is false because $\mathbf{z}_t \not\perp \mathbf{x}_{t+1} |$
 1076 $\mathbf{x}_{t-\tau:t}$. This leads to a contradiction, which implies $c > 0$ and hence $\mathcal{R}_o > \mathcal{R}_z$.

1077 Consider Equation (A5). Here, in general, for any $\hat{\mathbf{z}}_t$ we have $e \geq 0$ and hence $\mathcal{R}_o \geq \mathcal{R}_{\hat{z}}$.

1078 Then we leverage the law of total variance again and have:

$$\begin{aligned} &\text{Var}_{\mathbf{z}_t | \mathbf{x}_{t-\tau:t}} (\mathbb{E}(\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}, \mathbf{z}_t)) \\ &= \mathbb{E}_{\hat{\mathbf{z}}_t | \mathbf{x}_{t-\tau:t}} [\text{Var}_{\mathbf{z}_t | \mathbf{x}_{t-\tau:t}, \hat{\mathbf{z}}_t} (\mathbb{E}[\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}, \mathbf{z}_t])] + \text{Var}_{\hat{\mathbf{z}}_t | \mathbf{x}_{t-\tau:t}} (\mathbb{E}_{\mathbf{z}_t | \mathbf{x}_{t-\tau:t}, \hat{\mathbf{z}}_t} [\mathbb{E}[\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}, \mathbf{z}_t]]) \\ &= \mathbb{E}_{\hat{\mathbf{z}}_t | \mathbf{x}_{t-\tau:t}} [\text{Var}_{\mathbf{z}_t | \mathbf{x}_{t-\tau:t}, \hat{\mathbf{z}}_t} (\mathbb{E}[\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}, \mathbf{z}_t])] + \text{Var}_{\hat{\mathbf{z}}_t | \mathbf{x}_{t-\tau:t}} (\mathbb{E}[\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}, \hat{\mathbf{z}}_t]) \end{aligned} \quad (\text{A6})$$

1079 Then we take the expectation on both sides of Equation (A6) and have :

$$\begin{aligned} \underbrace{\mathbb{E}_{\mathbf{x}_{t-\tau:t}} [\text{Var}_{\mathbf{z}_t | \mathbf{x}_{t-\tau:t}} (\mathbb{E}[\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}, \mathbf{z}_t])]}_c &= \underbrace{\mathbb{E}_{\mathbf{x}_{t-\tau:t}} [\text{Var}_{\hat{\mathbf{z}}_t | \mathbf{x}_{t-\tau:t}} (\mathbb{E}[\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}, \hat{\mathbf{z}}_t])]}_e \\ &\quad + \mathbb{E}_{\hat{\mathbf{z}}_t, \mathbf{x}_{t-\tau:t}} [\text{Var}_{\mathbf{z}_t | \mathbf{x}_{t-\tau:t}, \hat{\mathbf{z}}_t} (\mathbb{E}[\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}, \mathbf{z}_t])] \end{aligned} \quad (\text{A7})$$

1080 Similarly, $\mathbb{E}_{\hat{\mathbf{z}}_t, \mathbf{x}_{t-\tau:t}} [\text{Var}_{\mathbf{z}_t | \mathbf{x}_{t-\tau:t}, \hat{\mathbf{z}}_t} (\mathbb{E}[\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}, \mathbf{z}_t])] = 0$ means that
 1081 $\text{Var}_{\mathbf{z}_t | \mathbf{x}_{t-\tau:t}, \hat{\mathbf{z}}_t} (\mathbb{E}[\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}, \mathbf{z}_t]) = 0$, implying that $\mathbb{E}[\mathbf{x}_{t+1} | \mathbf{x}_{t-\tau:t}, \mathbf{z}_t]$ is a constant, i.e.,
 1082 \mathbf{z}_t and $\hat{\mathbf{z}}_t$ have a one-to-one correspondence. Therefore, in general, $c \geq e$, and $c = e$ iff \mathbf{z}_t is
 1083 identifiable.

1084 By combining Equation (A4) and (A5), in general, we have $\mathcal{R}_o \geq \mathcal{R}_{\hat{\mathbf{z}}} \geq \mathcal{R}_{\mathbf{z}}$, and if \mathbf{z}_t is identifiable
 1085 we have $\mathcal{R}_o > \mathcal{R}_{\hat{\mathbf{z}}} = \mathcal{R}_{\mathbf{z}}$.

1086 □

1087 C.2 Proof of Theorem 2.

1088 For a better understanding of our proof, we begin by introducing an additional operator to represent
 1089 the point-wise distributional transformation. For generality, we denote two variables as \mathbf{a} and \mathbf{b} , with
 1090 corresponding support sets \mathcal{A} and \mathcal{B} .

1091 **Definition 5. (Diagonal Operator)** Consider two random variable a and b , density functions p_a
 1092 and p_b are defined on some support \mathcal{A} and \mathcal{B} , respectively. The diagonal operator $D_{b|a}$ maps the
 1093 density function p_a to another density function $D_{b|a} \circ p_a$ defined by the pointwise multiplication of
 1094 the function $p_{b|a}$ at a fixed point b :

$$p_{b|a}(b | \cdot) p_a = D_{b|a} \circ p_a, \text{ where } D_{b|a} = p_{b|a}(b | \cdot). \quad (\text{A8})$$

1095 **Theorem A2. (Block-wise Identification under 4 Adjacent Observed Variables.)** Suppose that
 1096 the observed and latent variables follow the data generation process. By matching the true joint
 1097 distribution of 4 adjacent observed variables, i.e., $\{\mathbf{x}_{t-2}, \mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}\}$, we further consider the
 1098 following assumptions:

- 1099 • **A1 (Bound and Continuous Density):** The joint distribution of \mathbf{x}, \mathbf{z} and their marginal and condi-
 1100 tional densities are bounded and continuous.
- 1101 • **A2 (Injectivity):** There exists observed variables \mathbf{x}_t such that for any $\mathbf{x}_t \in \mathcal{X}_t$, there exist a
 1102 $\mathbf{x}_{t-1} \in \mathcal{X}_{t-1}$ and a neighborhood ² \mathcal{N}^r around $(\mathbf{x}_t, \mathbf{x}_{t-1})$ such that, for any $(\bar{\mathbf{x}}_t, \bar{\mathbf{x}}_{t-1}) \in \mathcal{N}^r$,
 1103 $L_{\bar{\mathbf{x}}_t, \mathbf{x}_{t+1} | \mathbf{x}_{t-2}, \bar{\mathbf{x}}_{t-1}}$ is injective; $L_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t}, L_{\mathbf{x}_t | \mathbf{x}_{t-2}, \mathbf{x}_{t-1}}$ is injective for any $\mathbf{x}_t \in \mathcal{X}_t$ and $\mathbf{x}_{t-1} \in$
 1104 \mathcal{X}_{t-1} , respectively.
- 1105 • **A3 (Uniqueness of Spectral Decomposition)** For any $\mathbf{x}_t \in \mathcal{X}_t$ and any $\bar{\mathbf{z}}_t \neq \tilde{\mathbf{z}}_t \in \mathcal{Z}_t$, there exists a
 1106 $\mathbf{x}_{t-1} \in \mathcal{X}_{t-1}$ and corresponding neighborhood \mathcal{N}^r satisfying Assumption A2 such that, for some
 1107 $(\bar{\mathbf{x}}_t, \bar{\mathbf{x}}_{t-1}) \in \mathcal{N}^r$ with $\bar{\mathbf{x}}_t \neq \mathbf{x}_t, \bar{\mathbf{x}}_{t-1} \neq \mathbf{x}_{t-1}$:

1108 i $k(\mathbf{x}_t, \bar{\mathbf{x}}_t, \mathbf{x}_{t-1}, \bar{\mathbf{x}}_{t-1}, \bar{\mathbf{z}}_t) < C < \infty$ for any $\mathbf{z}_t \in \mathcal{Z}_t$ and some constant C .

1109 ii $k(\mathbf{x}_t, \bar{\mathbf{x}}_t, \mathbf{x}_{t-1}, \bar{\mathbf{x}}_{t-1}, \bar{\mathbf{z}}_t) \neq k(\mathbf{x}_t, \bar{\mathbf{x}}_t, \mathbf{x}_{t-1}, \bar{\mathbf{x}}_{t-1}, \tilde{\mathbf{z}}_t)$, where

$$k(\mathbf{x}_t, \bar{\mathbf{x}}_t, \mathbf{x}_{t-1}, \bar{\mathbf{x}}_{t-1}, \mathbf{z}_t) = \frac{p_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t}(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t) p_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t}(\bar{\mathbf{x}}_t | \bar{\mathbf{x}}_{t-1}, \mathbf{z}_t)}{p_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t}(\bar{\mathbf{x}}_t | \mathbf{x}_{t-1}, \mathbf{z}_t) p_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t}(\mathbf{x}_t | \bar{\mathbf{x}}_{t-1}, \mathbf{z}_t)}. \quad (\text{A9})$$

1110 Suppose that we have learned $(\hat{\mathbf{g}}, \hat{\mathbf{f}}, p_{\hat{\epsilon}})$ to achieve Equations (1), then the combination of Markov
 1111 state $\mathbf{z}_t, \mathbf{x}_t$ is identifiable, i.e., $[\hat{\mathbf{z}}_t, \hat{\mathbf{x}}_t] = H(\mathbf{z}_t, \mathbf{x}_t)$, where H is invertible and differentiable.

²Please refer to Appendix C.5 for the definition of neighborhood.

1112 *Proof.* By the definition of data generation process, the observed density $p_{\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}}$ equals

$$\begin{aligned}
& p_{\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} \\
&= \iint p_{\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{z}_t, \mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} d\mathbf{z}_t d\mathbf{z}_{t-1} \\
&= \iint p_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \mathbf{z}_t, \mathbf{z}_{t-1}} p_{\mathbf{x}_t, \mathbf{z}_t | \mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \mathbf{z}_{t-1}} p_{\mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} d\mathbf{z}_t d\mathbf{z}_{t-1} \\
&= \iint p_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t} p_{\mathbf{x}_t, \mathbf{z}_t | \mathbf{x}_{t-1}, \mathbf{z}_{t-1}} p_{\mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} d\mathbf{z}_t d\mathbf{z}_{t-1} \\
&= \iint p_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t} p_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t, \mathbf{z}_{t-1}} p_{\mathbf{z}_t | \mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \mathbf{z}_{t-1}} p_{\mathbf{z}_{t-1}, \mathbf{x}_{t-2}, \mathbf{z}_{t-1}} d\mathbf{z}_t d\mathbf{z}_{t-1} \\
&= \iint p_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t} p_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t, \mathbf{z}_{t-1}} p_{\mathbf{z}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \mathbf{z}_{t-1}} d\mathbf{z}_t d\mathbf{z}_{t-1}.
\end{aligned}$$

1113 According to the property of Markov process,

$$\begin{aligned}
p_{\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} &= \int p_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t} p_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t} \left(\int p_{\mathbf{z}_t, \mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} d\mathbf{z}_{t-1} \right) d\mathbf{z}_t \\
&= \int p_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t} p_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t} p_{\mathbf{z}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} d\mathbf{z}_t.
\end{aligned} \tag{A10}$$

1114 In operator notation, given values of $(\mathbf{x}_t, \mathbf{x}_{t-1}) \in \mathcal{X}_t \times \mathcal{X}_{t-1}$, this is

$$L_{\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} = L_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t} D_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t} L_{\mathbf{z}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}}. \tag{A11}$$

1115 After obtaining the representation of observed density function, furthermore, the structure of Markov
1116 process implies the following two equalities:

$$\begin{aligned}
p_{\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} &= \int p_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t} p_{\mathbf{x}_t, \mathbf{z}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} d\mathbf{z}_t, \\
p_{\mathbf{x}_t, \mathbf{z}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} &= \int p_{\mathbf{x}_t, \mathbf{z}_t | \mathbf{x}_{t-1}, \mathbf{z}_{t-1}} p_{\mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} d\mathbf{z}_{t-1}.
\end{aligned} \tag{A12}$$

1117 In operator notation, for fixed $\mathbf{x}_t, \mathbf{x}_{t-1}$, the above equations are expressed:

$$\begin{aligned}
L_{\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} &= L_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t} L_{\mathbf{x}_t, \mathbf{z}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}}, \\
L_{\mathbf{x}_t, \mathbf{z}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} &= L_{\mathbf{x}_t, \mathbf{z}_t | \mathbf{x}_{t-1}, \mathbf{z}_{t-1}} L_{\mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}}.
\end{aligned} \tag{A13}$$

1118 Substituting the second line into the first, we get

$$\begin{aligned}
L_{\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} &= L_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t} L_{\mathbf{x}_t, \mathbf{z}_t | \mathbf{x}_{t-1}, \mathbf{z}_{t-1}} L_{\mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} \\
\Leftrightarrow L_{\mathbf{x}_t, \mathbf{z}_t | \mathbf{x}_{t-1}, \mathbf{z}_{t-1}} L_{\mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} &= L_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t}^{-1} L_{\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}}.
\end{aligned} \tag{A14}$$

1119 The second line uses Assumption A2. Next, we eliminate $L_{\mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}}$ from the above. Again,
1120 using the conditional independence of Markov process, we have:

$$p_{\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} = \int p_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_{t-1}} p_{\mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} d\mathbf{z}_{t-1}, \tag{A15}$$

1121 which can be represented in terms of operator (for fixed \mathbf{x}_{t-1}) as:

$$\begin{aligned}
L_{\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} &= L_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_{t-1}} L_{\mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}}, \\
\Rightarrow L_{\mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} &= L_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_{t-1}}^{-1} L_{\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}}.
\end{aligned} \tag{A16}$$

1122 The R.H.S. applies Assumption A2. Hence, substituting the above into Eq. A14, we obtain the desired
1123 representation:

$$\begin{aligned}
& L_{\mathbf{x}_t, \mathbf{z}_t | \mathbf{x}_{t-1}, \mathbf{z}_{t-1}} L_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_{t-1}}^{-1} L_{\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} = L_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t}^{-1} L_{\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} \\
\Rightarrow L_{\mathbf{x}_t, \mathbf{z}_t | \mathbf{x}_{t-1}, \mathbf{z}_{t-1}} &= L_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t}^{-1} L_{\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} L_{\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}}^{-1} L_{\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}}.
\end{aligned} \tag{A17}$$

1124 The second line applies Assumption A2 to post-multiply by $L_{\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}}^{-1}$, while in the third line, we
 1125 postmultiply both sides by $L_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_{t-1}}$.

1126 For each \mathbf{x}_t , choose a \mathbf{x}_{t-1} and a neighborhood \mathcal{N}^r around $(\mathbf{x}_t, \mathbf{x}_{t-1})$ to satisfy Assumption A2
 1127 and A2, and pick a $(\bar{\mathbf{x}}_t, \bar{\mathbf{x}}_{t-1})$ within the neighborhood \mathcal{N}^r to satisfy Assumption A2. Because
 1128 $(\bar{\mathbf{x}}_t, \bar{\mathbf{x}}_{t-1}) \in \mathcal{N}^r$, also $(\mathbf{x}_t, \bar{\mathbf{x}}_{t-1}), (\bar{\mathbf{x}}_t, \mathbf{x}_{t-1}) \in \mathcal{N}^r$. By the representation of observations in
 1129 Eq. A11, we have

$$L_{\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} = L_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t} D_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t} L_{\mathbf{z}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}}.$$

1130 The first term on the R.H.S., $L_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t}$, does not depend on \mathbf{x}_{t-1} , and the last term $L_{\mathbf{z}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}}$
 1131 does not depend on \mathbf{x}_t . This feature suggests that, by evaluating Eq. A1 at the four pairs of points
 1132 $(\mathbf{x}_t, \mathbf{x}_{t-1}), (\bar{\mathbf{x}}_t, \mathbf{x}_{t-1}), (\mathbf{x}_t, \bar{\mathbf{x}}_{t-1}), (\bar{\mathbf{x}}_t, \bar{\mathbf{x}}_{t-1})$, each pair of equations will share one operator in
 1133 common. Specifically:

$$L_{\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} = L_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t} D_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t} L_{\mathbf{z}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}}, \quad (\text{A18})$$

$$L_{\mathbf{x}_{t+1}, \bar{\mathbf{x}}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} = L_{\mathbf{x}_{t+1} | \bar{\mathbf{x}}_t, \mathbf{z}_t} D_{\bar{\mathbf{x}}_t | \mathbf{x}_{t-1}, \mathbf{z}_t} L_{\mathbf{z}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}}, \quad (\text{A19})$$

$$L_{\mathbf{x}_{t+1}, \mathbf{x}_t, \bar{\mathbf{x}}_{t-1}, \mathbf{x}_{t-2}} = L_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t} D_{\mathbf{x}_t | \bar{\mathbf{x}}_{t-1}, \mathbf{z}_t} L_{\mathbf{z}_t, \bar{\mathbf{x}}_{t-1}, \mathbf{x}_{t-2}}, \quad (\text{A20})$$

$$L_{\mathbf{x}_{t+1}, \bar{\mathbf{x}}_t, \bar{\mathbf{x}}_{t-1}, \mathbf{x}_{t-2}} = L_{\mathbf{x}_{t+1} | \bar{\mathbf{x}}_t, \mathbf{z}_t} D_{\bar{\mathbf{x}}_t | \bar{\mathbf{x}}_{t-1}, \mathbf{z}_t} L_{\mathbf{z}_t, \bar{\mathbf{x}}_{t-1}, \mathbf{x}_{t-2}}. \quad (\text{A21})$$

1134 Assumption A2 implies that $L_{\mathbf{x}_{t+1} | \bar{\mathbf{x}}_t, \mathbf{z}_t}$ is invertible. Moreover, Assumption A2 implies
 1135 $p_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t}(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t) > 0$ for all \mathbf{z}_t , so that $D_{\bar{\mathbf{x}}_t | \mathbf{x}_{t-1}, \mathbf{z}_t}$ is invertible. We can then solve
 1136 for $L_{\mathbf{z}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}}$ from Eq. A19 as

$$D_{\bar{\mathbf{x}}_t | \mathbf{x}_{t-1}, \mathbf{z}_t}^{-1} L_{\mathbf{x}_{t+1} | \bar{\mathbf{x}}_t, \mathbf{z}_t}^{-1} L_{\mathbf{x}_{t+1}, \bar{\mathbf{x}}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} = L_{\mathbf{z}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}}. \quad (\text{A22})$$

1137 Plugging this expression into Eq. A18 leads to

$$L_{\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} = L_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t} D_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t} D_{\bar{\mathbf{x}}_t | \mathbf{x}_{t-1}, \mathbf{z}_t}^{-1} L_{\mathbf{x}_{t+1} | \bar{\mathbf{x}}_t, \mathbf{z}_t}^{-1} L_{\mathbf{x}_{t+1}, \bar{\mathbf{x}}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}}. \quad (\text{A23})$$

1138 Lemma 1 of [Hu and Schennach \[2008b\]](#) shows that, given the injectivity of $L_{\mathbf{x}_{t-2}, \bar{\mathbf{x}}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}}$ as in
 1139 Assumption A2, we can postmultiply by $L_{\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}}^{-1}$ to obtain:

$$\mathbf{A} \equiv L_{\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}}^{-1} L_{\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} = L_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t} D_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t} D_{\bar{\mathbf{x}}_t | \mathbf{x}_{t-1}, \mathbf{z}_t}^{-1} L_{\mathbf{x}_{t+1} | \bar{\mathbf{x}}_t, \mathbf{z}_t}^{-1}. \quad (\text{A24})$$

1140 Similarly, manipulations of Eq. A20 and Eq. A21 lead to

$$\mathbf{B} \equiv L_{\mathbf{x}_{t+1}, \bar{\mathbf{x}}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}}^{-1} L_{\mathbf{x}_{t+1}, \bar{\mathbf{x}}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}} = L_{\mathbf{x}_{t+1} | \bar{\mathbf{x}}_t, \mathbf{z}_t} D_{\bar{\mathbf{x}}_t | \bar{\mathbf{x}}_{t-1}, \mathbf{z}_t} D_{\mathbf{x}_t | \bar{\mathbf{x}}_{t-1}, \mathbf{z}_t}^{-1} L_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t}^{-1}. \quad (\text{A25})$$

1141 Assumption A2 guarantees that, for any \mathbf{x}_t , $(\bar{\mathbf{x}}_t, \mathbf{x}_{t-1}, \bar{\mathbf{x}}_{t-1})$ exist so that Eq. A24 and Eq. A25 are
 1142 valid operations. Finally, we postmultiply Eq. A24 by Eq. A25 to obtain:

$$\begin{aligned} \mathbf{AB} &= L_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t} D_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t} D_{\bar{\mathbf{x}}_t | \mathbf{x}_{t-1}, \mathbf{z}_t}^{-1} (L_{\mathbf{x}_{t+1} | \bar{\mathbf{x}}_t, \mathbf{z}_t} L_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t}) \times D_{\bar{\mathbf{x}}_t | \bar{\mathbf{x}}_{t-1}, \mathbf{z}_t} D_{\mathbf{x}_t | \bar{\mathbf{x}}_{t-1}, \mathbf{z}_t}^{-1} L_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t}^{-1} \\ &= L_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t} \left(D_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t} D_{\bar{\mathbf{x}}_t | \mathbf{x}_{t-1}, \mathbf{z}_t}^{-1} D_{\bar{\mathbf{x}}_t | \bar{\mathbf{x}}_{t-1}, \mathbf{z}_t} D_{\mathbf{x}_t | \bar{\mathbf{x}}_{t-1}, \mathbf{z}_t}^{-1} \right) L_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t}^{-1} \\ &\equiv L_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t} D_{\mathbf{x}_t, \bar{\mathbf{x}}_t, \mathbf{x}_{t-1}, \bar{\mathbf{x}}_{t-1}, \mathbf{z}_t} L_{\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{z}_t}^{-1}, \end{aligned} \quad (\text{A26})$$

1143 where

$$\begin{aligned} (D_{\mathbf{x}_t, \bar{\mathbf{x}}_t, \mathbf{x}_{t-1}, \bar{\mathbf{x}}_{t-1}, \mathbf{z}_t} h)(\mathbf{z}_t) &= \left(D_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t} D_{\bar{\mathbf{x}}_t | \mathbf{x}_{t-1}, \mathbf{z}_t}^{-1} D_{\bar{\mathbf{x}}_t | \bar{\mathbf{x}}_{t-1}, \mathbf{z}_t} D_{\mathbf{x}_t | \bar{\mathbf{x}}_{t-1}, \mathbf{z}_t}^{-1} h \right)(\mathbf{z}_t) \\ &= \frac{p_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t}(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t) p_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t}(\bar{\mathbf{x}}_t | \bar{\mathbf{x}}_{t-1}, \mathbf{z}_t)}{p_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t}(\bar{\mathbf{x}}_t | \mathbf{x}_{t-1}, \mathbf{z}_t) p_{\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t}(\mathbf{x}_t | \bar{\mathbf{x}}_{t-1}, \mathbf{z}_t)} h(\mathbf{z}_t) \\ &\equiv k(\mathbf{x}_t, \bar{\mathbf{x}}_t, \mathbf{x}_{t-1}, \bar{\mathbf{x}}_{t-1}, \mathbf{z}_t) h(\mathbf{z}_t). \end{aligned} \quad (\text{A27})$$

1144 By matching the marginal distribution of observed variables, we can define the operator $\hat{\mathbf{A}}\hat{\mathbf{B}}$ as the
 1145 estimated counterpart of the operator \mathbf{AB} , constructed using the estimated densities of $\hat{\mathbf{x}}_{t-2}, \hat{\mathbf{x}}_{t-1},$
 1146 $\hat{\mathbf{x}}_t, \hat{\mathbf{x}}_{t+1}$, and $\hat{\mathbf{z}}_t$. Since both the marginal and conditional distributions of the observed variables

are matched, the true model and the estimated model yield the same distribution over the observed variables. Therefore, we also have:

$$\hat{\mathbf{A}}\hat{\mathbf{B}} = \mathbf{A}\mathbf{B}. \quad (\text{A28})$$

Eq. A26 implies that the observed operator $\mathbf{A}\mathbf{B}$ has an inherent eigenvalue–eigenfunction decomposition, with the eigenvalues corresponding to the function $k(\mathbf{x}_t, \bar{\mathbf{x}}_t, \mathbf{x}_{t-1}, \bar{\mathbf{x}}_{t-1}, \mathbf{z}_t)$ and the eigenfunctions corresponding to the density $p_{\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{z}_t}(\cdot | \mathbf{x}_t, \mathbf{z}_t)$. Furthermore, Eq. A28 implies that $\mathbf{A}\mathbf{B}$ and $\hat{\mathbf{A}}\hat{\mathbf{B}}$ admit the same eigendecompositions, which are similar to the decomposition in [Hu and Schennach \[2008b\]](#) or [Carroll et al. \[2010\]](#). Assumption A2 ensures that this decomposition is unique. Specifically, the operator $\mathbf{A}\mathbf{B}$ on the L.H.S. has the same spectrum as the diagonal operator $D_{\mathbf{x}_t, \bar{\mathbf{x}}_t, \mathbf{x}_{t-1}, \bar{\mathbf{x}}_{t-1}, \mathbf{z}_t}$. Assumption A2 guarantees that the spectrum of the diagonal operator is bounded. Since an operator is bounded by the largest element of its spectrum, Assumption A2 also implies that the operator $\mathbf{A}\mathbf{B}$ is bounded, whence we can apply Theorem XV.4.3.5 from [Dunford and Schwartz \[1971\]](#) to show the uniqueness of the spectral decomposition of bounded linear operators:

$$L_{\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{z}_t} = CL_{\hat{\mathbf{x}}_{t+1}|\hat{\mathbf{x}}_t, \hat{\mathbf{z}}_t}P^{-1}. \quad D_{\mathbf{x}_t, \bar{\mathbf{x}}_t, \mathbf{x}_{t-1}, \bar{\mathbf{x}}_{t-1}, \mathbf{z}_t} = PD_{\hat{\mathbf{x}}_t, \hat{\bar{\mathbf{x}}}_t, \hat{\mathbf{x}}_{t-1}, \hat{\bar{\mathbf{x}}}_{t-1}, \hat{\mathbf{z}}_t}P^{-1} \quad (\text{A29})$$

where C is a scalar accounting for scaling indeterminacy and P is a permutation on the order of elements in $L_{\hat{\mathbf{x}}_{t+1}|\hat{\mathbf{x}}_t, \hat{\mathbf{z}}_t}$, as discussed in [\[Dunford and Schwartz, 1971\]](#). These forms of indeterminacy are analogous to those in eigendecomposition, which can be viewed as a finite-dimensional special case. We will show how to resolve the indeterminacies in eigen(spectral) decomposition.

First, Eq. A29 itself does not imply that the eigenvalues $k(\mathbf{x}_t, \bar{\mathbf{x}}_t, \mathbf{x}_{t-1}, \bar{\mathbf{x}}_{t-1}, \mathbf{z}_t)$ are distinct for different values \mathbf{z}_t . When the eigenvalues are the same for multiple values of \mathbf{z}_t , the corresponding eigenfunctions are only determined up to an arbitrary linear combination, implying that they are not identified. Assumption A2 rules out this possibility, and implies that for each \mathbf{x}_t , we can find values $\bar{\mathbf{x}}_t, \mathbf{x}_{t-1}, \bar{\mathbf{x}}_{t-1}$ such that the eigenvalues are distinct across all \mathbf{z}_t .

Second, since the normalizing condition

$$\int_{\hat{\mathcal{X}}_{t+1}} p_{\hat{\mathbf{x}}_{t+1}|\hat{\mathbf{x}}_t, \hat{\mathbf{z}}_t} d\hat{\mathbf{x}}_{t+1} = 1 \quad (\text{A30})$$

must hold for every $\hat{\mathbf{z}}_t$, one only solution is to set $C = 1$, that is, the scaling indeterminacy is resolved.

Ultimately, the unorder of eigenvalues/eigenfunctions is left. We have match the observational distributions $\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t+1}$, hence, the operator, $L_{\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{z}_t}$, corresponding to the set $\{p_{\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{z}_t}(\cdot | \mathbf{x}_t, \mathbf{z}_t)\}$ for all $\mathbf{x}_t, \mathbf{z}_t$, admits a unique solution (ordering ambiguity of eigendecomposition only changes the entry position):

$$\{p_{\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{z}_t}(\cdot | \mathbf{x}_t, \mathbf{z}_t)\} = \{p_{\mathbf{x}_{t+1}|\hat{\mathbf{x}}_t, \hat{\mathbf{z}}_t}(\mathbf{x}_{t+1} | \hat{\mathbf{x}}_t, \hat{\mathbf{z}}_t)\}, \quad \text{for all } \mathbf{x}_t, \mathbf{z}_t, \hat{\mathbf{x}}_t, \hat{\mathbf{z}}_t \quad (\text{A31})$$

Due to the set is unorder, the only way to match the R.H.S. with the L.H.S. in a consistent order is to exchange the conditioning variables, that is,

$$\begin{aligned} & \{p_{\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{z}_t}(\cdot | \mathbf{x}_t^{(1)}, \mathbf{z}_t^{(1)}), p_{\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{z}_t}(\cdot | \mathbf{x}_t^{(2)}, \mathbf{z}_t^{(2)}), \dots\} \\ &= \{p_{\mathbf{x}_{t+1}|\hat{\mathbf{x}}_t, \hat{\mathbf{z}}_t}(\cdot | \hat{\mathbf{x}}_t^{(1)}, \hat{\mathbf{z}}_t^{(1)}), p_{\mathbf{x}_{t+1}|\hat{\mathbf{x}}_t, \hat{\mathbf{z}}_t}(\cdot | \hat{\mathbf{x}}_t^{(2)}, \hat{\mathbf{z}}_t^{(2)}), \dots\} \\ \Rightarrow & [p_{\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{z}_t}(\cdot | \mathbf{x}_t^{(\pi(1))}, \mathbf{z}_t^{(\pi(1))}), p_{\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{z}_t}(\cdot | \mathbf{x}_t^{(\pi(2))}, \mathbf{z}_t^{(\pi(2))}), \dots] \\ &= [p_{\mathbf{x}_{t+1}|\hat{\mathbf{x}}_t, \hat{\mathbf{z}}_t}(\cdot | \hat{\mathbf{x}}_t^{(\pi(1))}, \hat{\mathbf{z}}_t^{(\pi(1))}), p_{\mathbf{x}_{t+1}|\hat{\mathbf{x}}_t, \hat{\mathbf{z}}_t}(\cdot | \hat{\mathbf{x}}_t^{(\pi(2))}, \hat{\mathbf{z}}_t^{(\pi(2))}), \dots] \end{aligned}$$

where superscript (\cdot) denotes the index of the conditioning variables $[\mathbf{x}_t, \mathbf{z}_t]$, and π is reindexing the conditioning variables. We use a relabeling map H to represent its corresponding value mapping:

$$p_{\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{z}_t}(\cdot | H(\mathbf{x}_t, \mathbf{z}_t)) = p_{\mathbf{x}_{t+1}|\hat{\mathbf{x}}_t, \hat{\mathbf{z}}_t}(\cdot | \hat{\mathbf{x}}_t, \hat{\mathbf{z}}_t), \quad \text{for all } \mathbf{x}_t, \mathbf{z}_t, \hat{\mathbf{x}}_t, \hat{\mathbf{z}}_t \quad (\text{A32})$$

By Assumption A2, different x^* corresponds to different $p_{\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{z}_t}(\cdot | H(\mathbf{x}_t, \mathbf{z}_t))$, there is no repeated element in $\{p_{\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{z}_t}(\cdot | H(\mathbf{x}_t, \mathbf{z}_t))\}$ (and $\{p_{\mathbf{x}_{t+1}|\hat{\mathbf{x}}_t, \hat{\mathbf{z}}_t}(\cdot | \hat{\mathbf{x}}_t, \hat{\mathbf{z}}_t)\}$). Hence, the relabelling map H is one-to-one.

Furthermore, Assumption A2 implies that $p_{\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{z}_t}(\cdot | H(\mathbf{x}_t, \mathbf{z}_t))$ determines a unique $H(\mathbf{x}_t, \mathbf{z}_t)$. The same holds for the $p_{\mathbf{x}_{t+1}|\hat{\mathbf{x}}_t, \hat{\mathbf{z}}_t}(\cdot | \hat{\mathbf{x}}_t, \hat{\mathbf{z}}_t)$, implying that

$$p_{\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{z}_t}(\cdot | H(\mathbf{x}_t, \mathbf{z}_t)) = p_{\mathbf{x}_{t+1}|\hat{\mathbf{x}}_t, \hat{\mathbf{z}}_t}(\cdot | \hat{\mathbf{x}}_t, \hat{\mathbf{z}}_t) \implies \hat{\mathbf{x}}_t, \hat{\mathbf{z}}_t = H(\mathbf{x}_t, \mathbf{z}_t), \quad (\text{A33})$$

1183 implying that $\mathbf{x}_t, \mathbf{z}_t$ is block-wise identifiable.

1184 Next, suppose the implemented MLP used in the transition module is differentiable, then we can assert
 1185 that there exists a functional M such that $M[p_{\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{z}_t}(\cdot | \mathbf{x}_t, \mathbf{z}_t)] = H(\mathbf{x}_t, \mathbf{z}_t)$ for all $\mathbf{z}_t \in \mathcal{Z}_t$ and
 1186 $\mathbf{x}_t \in \mathcal{X}_t$, where H is differentiable, that is, we can learn a differentiable function H that

$$M[p_{\mathbf{x}_{t+1}|\hat{\mathbf{x}}_t, \hat{\mathbf{z}}_t}(\cdot | \mathbf{x}_t, \mathbf{z}_t)] = M[p_{\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{z}_t}(\cdot | H(\mathbf{x}_t, \mathbf{z}_t))] = H(\mathbf{x}_t, \mathbf{z}_t), \quad (\text{A34})$$

1187 which is equal to $\hat{\mathbf{x}}_t, \hat{\mathbf{z}}_t$ only if H is differentiable. \square

1188 C.3 More Discussion of injective linear operators

1189 A linear operator can be intuitively understood as a function that maps one distribution of random
 1190 variables to another. Specifically, when we assume the injectivity of a linear operator in the context
 1191 of nonparametric identification, we are asserting that distinct input distributions of the operator
 1192 correspond to distinct output distributions. This injectivity ensures that there is no ambiguity in the
 1193 transformation from the input space to the output space, making the operator’s behavior predictable
 1194 and identifiable. An example from a real-world scenario can be seen in weather forecasting. The
 1195 temperature on a given day can be influenced by several previous days’ temperatures. If we view the
 1196 relationship between past and future temperatures as a linear operator, injectivity would mean that
 1197 each unique pattern of past temperatures leads to a distinct forecast for the future temperature. The
 1198 injectivity of this operator ensures that the mapping from past weather data to future forecasts does
 1199 not result in ambiguity, allowing for more accurate and reliable predictions.

1200 For a better understanding of this assumption, we provide several examples that describe the mapping
 1201 from $p_{\mathbf{a}} \rightarrow p_{\mathbf{b}}$, where \mathbf{a} and \mathbf{b} are random variables.

1202 **Example 1 (Inverse Transformation).** $b = g(a)$, where g is an invertible function.

1203 **Example 2 (Additive Transformation).** $b = a + \epsilon$, where $p(\epsilon)$ must not vanish everywhere after the
 1204 Fourier transform (Theorem 2.1 in [Mattner \[1993\]](#)).

1205 **Example 3.** $b = g(a) + \epsilon$, where the same conditions from Examples 1 and 2 are required.

1206 **Example 4 (Post-linear Transformation).** $b = g_1(g_2(a) + \epsilon)$, a post-nonlinear model with invertible
 1207 nonlinear functions g_1, g_2 , combining the assumptions in **Examples 1-3**.

1208 **Example 5 (Nonlinear Transformation with Exponential Family).** $b = g(a, \epsilon)$, where the joint
 1209 distribution $p(a, b)$ follows an exponential family.

1210 **Example 6 (General Nonlinear Transformation).** $b = g(a, \epsilon)$, a general nonlinear formulation.
 1211 Certain deviations from the nonlinear additive model (**Example 3**), e.g., polynomial perturbations,
 1212 can still be tractable.

1213 C.4 Monotonicity and Normalization Assumption

1214 **Assumption 1** (Monotonicity and Normalization Assumption [[Hu and Shum, 2012](#)]). For any
 1215 $\mathbf{x}_t \in \mathcal{X}_t$, there exists a known functional G such that $G[p_{\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{z}_t}(\cdot | \mathbf{x}_t, \mathbf{z}_t)]$ is monotonic in \mathbf{z}_t . We
 1216 normalize $\mathbf{z}_t = G[p_{\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{z}_t}(\cdot | \mathbf{x}_t, \mathbf{z}_t)]$.

1217 C.5 Definition of Neighborhood

1218 **Definition 6 (Neighborhood).** Given a point x in a metric space, and a positive number r , the
 1219 neighborhood N^r of x is defined as:

$$N^r(x) = \{y : d(x, y) < r\}. \quad (\text{A35})$$

1220 C.6 More Discussion of Uniqueness of Spectral Decomposition

1221 This assumption essentially states that, in order to identify the latent variables of the system, it is
 1222 necessary to observe four different transitions of the observed variables that are governed by the
 1223 same latent variables. For a better understanding of this assumption, we provide an economic model.
 1224 Consider an economic model where \mathbf{x}_t represents the inflation rate at time t , and \mathbf{z}_t represents the
 1225 economic regime (such as a recession or a period of growth). To accurately identify the economic
 1226 regime, we would need to observe inflation under four distinct scenarios: transitions from a high-
 1227 inflation state to a low-inflation state, and from a low-inflation state to a high-inflation state, under

different historical conditions. These four observed inflation transitions allow us to identify whether the economy is in a recession or growth phase, based on the changes in inflation behavior.

This assumption is straightforward to satisfy in real-world economic modeling, especially when there is access to sufficient historical inflation data. In practice, there are often multiple transitions between inflation states over time, corresponding to shifts in the economic regime (e.g., moving from high inflation during an economic boom to low inflation during a recession). By collecting enough observations across different periods of economic change, this assumption can be easily fulfilled, ensuring that we can identify the underlying economic regime with confidence.

C.7 Proof of Theorem 3.

Lemma A1. (Component-wise Identification of \mathbf{z}_t with instantaneous dependencies under sparse causal influence on latent dynamics. *Li et al. [2025]* For a series of observed variables $\mathbf{x}_t \in \mathbb{R}^n$ and estimated latent variables $\hat{\mathbf{z}}_t \in \mathbb{R}^n$ with the corresponding process $\hat{f}_i, \hat{p}(\epsilon), \hat{\mathbf{g}}$, where $\hat{\mathbf{g}}$ is invertible, suppose the process subject to observational equivalence $\mathbf{x}_t = \hat{\mathbf{g}}(\hat{\mathbf{z}}_t)$. Let $\mathbf{c}_t \triangleq \{\mathbf{z}_{t-1}, \mathbf{z}_t\} \in \mathbb{R}^{2n}$ and $\mathcal{M}_{\mathbf{c}_t}$ be the variable set of two consecutive timestamps and the corresponding Markov network, respectively. Suppose the following assumptions hold:

- **A4 (Smooth and Positive Density):** The conditional probability function of the latent variables \mathbf{c}_t is smooth and positive, i.e., $p(\mathbf{c}_t | \mathbf{z}_{t-2})$ is third-order differentiable and $p(\mathbf{c}_t | \mathbf{z}_{t-2}) > 0$ over \mathbb{R}^{2n} ,
- **A5 (Sufficient Variability):** Denote $|\mathcal{M}_{\mathbf{c}_t}|$ as the number of edges in Markov network $\mathcal{M}_{\mathbf{c}_t}$. Let

$$w(m) = \left(\frac{\partial^3 \log p(\mathbf{c}_t | \mathbf{z}_{t-2})}{\partial c_{t,1}^2 \partial z_{t-2,m}}, \dots, \frac{\partial^3 \log p(\mathbf{c}_t | \mathbf{z}_{t-2})}{\partial c_{t,2n}^2 \partial z_{t-2,m}} \right) \oplus \left(\frac{\partial^2 \log p(\mathbf{c}_t | \mathbf{z}_{t-2})}{\partial c_{t,1} \partial z_{t-2,m}}, \dots, \frac{\partial^2 \log p(\mathbf{c}_t | \mathbf{z}_{t-2})}{\partial c_{t,2n} \partial z_{t-2,m}} \right) \oplus \left(\frac{\partial^3 \log p(\mathbf{c}_t | \mathbf{z}_{t-2})}{\partial c_{t,i} \partial c_{t,j} \partial z_{t-2,m}} \right)_{(i,j) \in \mathcal{E}(\mathcal{M}_{\mathbf{c}_t})}, \quad (\text{A36})$$

where \oplus denotes concatenation operation and $(i, j) \in \mathcal{E}(\mathcal{M}_{\mathbf{c}_t})$ denotes all pairwise indice such that $c_{t,i}, c_{t,j}$ are adjacent in $\mathcal{M}_{\mathbf{c}_t}$. For $m \in [1, \dots, n]$, there exist $4n + |\mathcal{M}_{\mathbf{c}_t}|$ different values of $\mathbf{z}_{t-2,m}$, such that the $4n + |\mathcal{M}_{\mathbf{c}_t}|$ values of vector functions $w(m)$ are linearly independent.

- **A6 (Latent Process Sparsity):** For any $z_{it} \in \mathbf{z}_t$, the intimate neighbor set of z_{it} is an empty set.

When the observational equivalence is achieved with the minimal number of edges of the estimated Markov network of $\mathcal{M}_{\hat{\mathbf{c}}_t}$, there exists a permutation π of the estimated latent variables, such that z_{it} and $\hat{z}_{\pi(i)t}$ is one-to-one corresponding, i.e., z_{it} is component-wise identifiable.

Proof. The proof can be summarized into three steps. First, we leverage the sparsity among latent variables to show the relationships between ground-truth and estimated latent variables. Sequentially, we show that the estimated Markov network $\mathcal{M}_{\hat{\mathbf{c}}_t}$ is isomorphic to the ground-truth Markov networks $\mathcal{M}_{\mathbf{c}_t}$. Finally, we show that the latent variables are component-wise identifiable under the sparse mixture procedure condition.

Step1: Relationships between Ground-truth and Estimated Latent Variables. We start from the matched marginal distribution to develop the relationship between \mathbf{z}_t and $\hat{\mathbf{z}}_t$ as follows:

$$p(\hat{\mathbf{x}}_t) = p(\mathbf{x}_t) \iff p(\hat{\mathbf{g}}(\hat{\mathbf{z}}_t)) = p(\mathbf{g}(\mathbf{z}_t)) \iff p((\mathbf{g}^{-1} \circ \hat{\mathbf{g}})(\hat{\mathbf{z}}_t)) = p(\mathbf{z}_t) \iff p(h_z(\hat{\mathbf{z}}_t)) = p(\mathbf{z}_t), \quad (\text{A37})$$

where $\hat{\mathbf{g}} : \mathcal{Z} \rightarrow \mathcal{X}$ denotes the estimated mixing function, and $h := \mathbf{g}^{-1} \circ \hat{\mathbf{g}}$ is the transformation between the ground-truth latent variables and the estimated ones. Since $\hat{\mathbf{g}}$ and \mathbf{g} are invertible, h is invertible as well. Since Equation (A37) holds true for all time steps, there must exist an invertible function h_c such that $p(h_c(\hat{\mathbf{c}}_t)) = p(\mathbf{c}_t)$, whose Jacobian matrix at time step t is

$$\mathbf{J}_{h_c,t} = \begin{bmatrix} \mathbf{J}_{h_z,t-1} & 0 \\ 0 & \mathbf{J}_{h_z,t} \end{bmatrix}. \quad (\text{A38})$$

Then for each value of \mathbf{x}_{t-2} , the Jacobian matrix of the mapping from $(\mathbf{x}_{t-2}, \hat{\mathbf{c}}_t)$ to $(\mathbf{x}_{t-2}, \mathbf{c}_t)$ can be written as follows:

$$\begin{bmatrix} \mathbf{I} & 0 \\ * & \mathbf{J}_{h_c,t} \end{bmatrix},$$

1266 where $*$ denotes any matrix. Since \mathbf{x}_{t-2} can be fully characterized by itself, the left top and right
 1267 top block are $\mathbf{1}$ and $\mathbf{0}$ respectively, and the determinant of this Jacobian matrix is the same as $|\mathbf{J}_{h_c,t}|$.
 1268 Therefore, we have:

$$p(\hat{\mathbf{c}}_t, \mathbf{x}_{t-2}) = p(\mathbf{c}_t, \mathbf{x}_{t-2}) |\mathbf{J}_{h_c,t}|. \quad (\text{A39})$$

1269 Dividing both sides of Equation (A39) by $p(\mathbf{x}_{t-2})$, we further have:

$$p(\hat{\mathbf{c}}_t | \mathbf{x}_{t-2}) = p(\mathbf{c}_t | \mathbf{x}_{t-2}) |\mathbf{J}_{h_c,t}|. \quad (\text{A40})$$

1270 Since $p(\mathbf{c}_t | \mathbf{x}_{t-2}) = p(\mathbf{c}_t | g(\mathbf{z}_{t-2})) = p(\mathbf{c}_t | \mathbf{z}_{t-2})$, and similarly $p(\hat{\mathbf{c}}_t | \mathbf{x}_{t-2}) = p(\hat{\mathbf{c}}_t | \hat{\mathbf{z}}_{t-2})$, we have:

$$\log p(\hat{\mathbf{c}}_t | \hat{\mathbf{z}}_{t-2}) = \log p(\mathbf{c}_t | \mathbf{z}_{t-2}) + \log |\mathbf{J}_{h_c,t}|. \quad (\text{A41})$$

1271 Let $\hat{c}_{t,k}, \hat{c}_{t,l}$ be two different variables that are not adjacent in the estimated Markov network $\mathcal{M}_{\hat{\mathbf{c}}_t}$
 1272 over $\hat{\mathbf{c}}_t = \{\hat{\mathbf{z}}_{t-1}, \hat{\mathbf{z}}_t\}$. We conduct the first-order derivative w.r.t. $\hat{c}_{t,k}$ and have

$$\frac{\partial \log p(\hat{\mathbf{c}}_t | \hat{\mathbf{z}}_{t-2})}{\partial \hat{c}_{t,k}} = \sum_{i=1}^{2n} \frac{\partial \log p(\mathbf{c}_t | \mathbf{z}_{t-2})}{\partial c_{t,i}} \cdot \frac{\partial c_{t,i}}{\partial \hat{c}_{t,k}} + \frac{\partial \log |\mathbf{J}_{h_c,t}|}{\partial \hat{c}_{t,k}}. \quad (\text{A42})$$

1273 We further conduct the second-order derivative w.r.t. $\hat{c}_{t,k}$ and $\hat{c}_{t,l}$, then we have:

$$\begin{aligned} \frac{\partial^2 \log p(\hat{\mathbf{c}}_t | \hat{\mathbf{z}}_{t-2})}{\partial \hat{c}_{t,k} \partial \hat{c}_{t,l}} &= \sum_{i=1}^{2n} \sum_{j=1}^{2n} \frac{\partial^2 \log p(\mathbf{c}_t | \mathbf{z}_{t-2})}{\partial c_{t,i} \partial c_{t,j}} \cdot \frac{\partial c_{t,i}}{\partial \hat{c}_{t,k}} \cdot \frac{\partial c_{t,j}}{\partial \hat{c}_{t,l}} \\ &\quad + \sum_{i=1}^{2n} \frac{\partial \log p(\mathbf{c}_t | \mathbf{z}_{t-2})}{\partial c_{t,i}} \cdot \frac{\partial^2 c_{t,i}}{\partial \hat{c}_{t,k} \partial \hat{c}_{t,l}} + \frac{\partial^2 \log |\mathbf{J}_{h_c,t}|}{\partial \hat{c}_{t,k} \partial \hat{c}_{t,l}}. \end{aligned} \quad (\text{A43})$$

1274 Since $\hat{c}_{t,k}, \hat{c}_{t,l}$ are not adjacent in $\mathcal{M}_{\hat{\mathbf{c}}_t}$, $\hat{c}_{t,k}$ and $\hat{c}_{t,l}$ are conditionally independent given
 1275 $\hat{\mathbf{c}}_t \setminus \{\hat{c}_{t,k}, \hat{c}_{t,l}\}$. Utilizing the fact that conditional independence can lead to zero cross derivative [Lin,
 1276 1997], for each value of $\hat{\mathbf{z}}_{t-2}$, we have

$$\begin{aligned} \frac{\partial^2 \log p(\hat{\mathbf{c}}_t | \hat{\mathbf{z}}_{t-2})}{\partial \hat{c}_{t,k} \partial \hat{c}_{t,l}} &= \frac{\partial^2 \log p(\hat{c}_{t,k} | \hat{\mathbf{c}}_t \setminus \{\hat{c}_{t,k}, \hat{c}_{t,l}\}, \hat{\mathbf{z}}_{t-2})}{\partial \hat{c}_{t,k} \partial \hat{c}_{t,l}} + \frac{\partial^2 \log p(\hat{c}_{t,l} | \hat{\mathbf{c}}_t \setminus \{\hat{c}_{t,k}, \hat{c}_{t,l}\}, \hat{\mathbf{z}}_{t-2})}{\partial \hat{c}_{t,k} \partial \hat{c}_{t,l}} \\ &\quad + \frac{\partial^2 \log p(\hat{\mathbf{c}}_t \setminus \{\hat{c}_{t,k}, \hat{c}_{t,l}\} | \hat{\mathbf{z}}_{t-2})}{\partial \hat{c}_{t,k} \partial \hat{c}_{t,l}} = 0. \end{aligned} \quad (\text{A44})$$

1277 Bring in Equation (A44), Equation (A43) can be further derived as

$$\begin{aligned} 0 &= \underbrace{\sum_{i=1}^{2n} \frac{\partial^2 \log p(\mathbf{c}_t | \mathbf{z}_{t-2})}{\partial c_{t,i}^2} \cdot \frac{\partial c_{t,i}}{\partial \hat{c}_{t,k}} \cdot \frac{\partial c_{t,i}}{\partial \hat{c}_{t,l}}}_{\text{(i) } i=j} + \underbrace{\sum_{i=1}^{2n} \sum_{j:(j,i) \in \mathcal{E}(\mathcal{M}_{\mathbf{c}_t})} \frac{\partial^2 \log p(\mathbf{c}_t | \mathbf{z}_{t-2})}{\partial c_{t,i} \partial c_{t,j}} \cdot \frac{\partial c_{t,i}}{\partial \hat{c}_{t,k}} \cdot \frac{\partial c_{t,j}}{\partial \hat{c}_{t,l}}}_{\text{(ii) } c_{t,i} \text{ and } c_{t,j} \text{ are adjacent in } \mathcal{M}_{\mathbf{c}_t}} \\ &\quad + \underbrace{\sum_{i=1}^{2n} \sum_{j:(j,i) \notin \mathcal{E}(\mathcal{M}_{\mathbf{c}_t})} \frac{\partial^2 \log p(\mathbf{c}_t | \mathbf{z}_{t-2})}{\partial c_{t,i} \partial c_{t,j}} \cdot \frac{\partial c_{t,i}}{\partial \hat{c}_{t,k}} \cdot \frac{\partial c_{t,j}}{\partial \hat{c}_{t,l}}}_{\text{(iii) } c_{t,i} \text{ and } c_{t,j} \text{ are not adjacent in } \mathcal{M}_{\mathbf{c}_t}} \\ &\quad + \sum_{i=1}^{2n} \frac{\partial \log p(\mathbf{c}_t | \mathbf{z}_{t-2})}{\partial c_{t,i}} \cdot \frac{\partial^2 c_{t,i}}{\partial \hat{c}_{t,k} \partial \hat{c}_{t,l}} + \frac{\partial \log |\mathbf{J}_{h_c,t}|}{\partial \hat{c}_{t,k} \partial \hat{c}_{t,l}}, \end{aligned} \quad (\text{A45})$$

1278 where $(j, i) \in \mathcal{E}(\mathcal{M}_{\mathbf{c}_t})$ denotes that $c_{t,i}$ and $c_{t,j}$ are adjacent in $\mathcal{M}_{\mathbf{c}_t}$. Similar to Equation (A44), we
 1279 have $\frac{\partial^2 p(\mathbf{c}_t | \mathbf{z}_{t-2})}{\partial c_{t,i} \partial c_{t,j}} = 0$ when $c_{t,i}, c_{t,j}$ are not adjacent in $\mathcal{M}_{\mathbf{c}_t}$. Thus, Equation (A45) can be rewritten
 1280 as

$$\begin{aligned} 0 &= \sum_{i=1}^{2n} \frac{\partial^2 \log p(\mathbf{c}_t | \mathbf{z}_{t-2})}{\partial c_{t,i}^2} \cdot \frac{\partial c_{t,i}}{\partial \hat{c}_{t,k}} \cdot \frac{\partial c_{t,i}}{\partial \hat{c}_{t,l}} + \sum_{i=1}^{2n} \sum_{j:(j,i) \in \mathcal{E}(\mathcal{M}_{\mathbf{c}_t})} \frac{\partial^2 \log p(\mathbf{c}_t | \mathbf{z}_{t-2})}{\partial c_{t,i} \partial c_{t,j}} \cdot \frac{\partial c_{t,i}}{\partial \hat{c}_{t,k}} \cdot \frac{\partial c_{t,j}}{\partial \hat{c}_{t,l}} \\ &\quad + \sum_{i=1}^{2n} \frac{\partial \log p(\mathbf{c}_t | \mathbf{z}_{t-2})}{\partial c_{t,i}} \cdot \frac{\partial^2 c_{t,i}}{\partial \hat{c}_{t,k} \partial \hat{c}_{t,l}} + \frac{\partial \log |\mathbf{J}_{h_c,t}|}{\partial \hat{c}_{t,k} \partial \hat{c}_{t,l}}. \end{aligned} \quad (\text{A46})$$

1281 Then for each $m = 1, 2, \dots, n$ and each value of $z_{t-2,m}$, we conduct partial derivative on both sides
 1282 of Equation (A46) and have:

$$0 = \sum_{i=1}^{2n} \frac{\partial^3 \log p(\mathbf{c}_t | \mathbf{z}_{t-2})}{\partial c_{t,i}^2 \partial z_{t-2,m}} \cdot \frac{\partial c_{t,i}}{\partial \hat{c}_{t,k}} \cdot \frac{\partial c_{t,i}}{\partial \hat{c}_{t,l}} + \sum_{i=1}^{2n} \sum_{j:(j,i) \in \mathcal{E}(\mathcal{M}_e)} \frac{\partial^3 \log p(\mathbf{c}_t | \mathbf{z}_{t-2})}{\partial c_{t,i} \partial c_{t,j} \partial z_{t-2,m}} \cdot \frac{\partial c_{t,i}}{\partial \hat{c}_{t,k}} \cdot \frac{\partial c_{t,j}}{\partial \hat{c}_{t,l}} + \sum_{i=1}^{2n} \frac{\partial^2 \log p(c_t | \mathbf{z}_{t-2})}{\partial c_{t,i} \partial z_{t-2,m}} \cdot \frac{\partial c_{t,i}^2}{\partial \hat{c}_{t,k} \partial \hat{c}_{t,l}}, \quad (\text{A47})$$

1283 Finally we have

$$0 = \sum_{i=1}^{2n} \frac{\partial^3 \log p(\mathbf{c}_t | \mathbf{z}_{t-2})}{\partial c_{t,i}^2 \partial z_{t-2,m}} \cdot \frac{\partial c_{t,i}}{\partial \hat{c}_{t,k}} \cdot \frac{\partial c_{t,i}}{\partial \hat{c}_{t,l}} + \sum_{i=1}^{2n} \frac{\partial^2 \log p(c_t | \mathbf{z}_{t-2})}{\partial c_{t,i} \partial z_{t-2,m}} \cdot \frac{\partial c_{t,i}^2}{\partial \hat{c}_{t,k} \partial \hat{c}_{t,l}} + \sum_{i,j:(j,i) \in \mathcal{E}(\mathcal{M}_e)} \frac{\partial^3 \log p(\mathbf{c}_t | \mathbf{z}_{t-2})}{\partial c_{t,i} \partial c_{t,j} \partial z_{t-2,m}} \cdot \left(\frac{\partial c_{t,i}}{\partial \hat{c}_{t,k}} \cdot \frac{\partial c_{t,j}}{\partial \hat{c}_{t,l}} + \frac{\partial c_{t,j}}{\partial \hat{c}_{t,k}} \cdot \frac{\partial c_{t,i}}{\partial \hat{c}_{t,l}} \right). \quad (\text{A48})$$

1284 According to Assumption A2, we can construct $4n + |\mathcal{M}_e|$ different equations with different values
 1285 of $z_{t-2,m}$, and the coefficients of the equation system they form are linearly independent. To ensure
 1286 that the right-hand side of the equations are always 0, the only solution is

$$\frac{\partial c_{t,i}}{\partial \hat{c}_{t,k}} \cdot \frac{\partial c_{t,i}}{\partial \hat{c}_{t,l}} = 0, \quad (\text{A49})$$

1287

$$\frac{\partial c_{t,i}}{\partial \hat{c}_{t,k}} \cdot \frac{\partial c_{t,j}}{\partial \hat{c}_{t,l}} + \frac{\partial c_{t,j}}{\partial \hat{c}_{t,k}} \cdot \frac{\partial c_{t,i}}{\partial \hat{c}_{t,l}} = 0, \quad (\text{A50})$$

1288

$$\frac{\partial c_{t,i}^2}{\partial \hat{c}_{t,k} \partial \hat{c}_{t,l}} = 0. \quad (\text{A51})$$

1289 Bringing Eq A49 into Eq A50, at least one product must be zero, thus the other must be zero as well.
 1290 That is,

$$\frac{\partial c_{t,i}}{\partial \hat{c}_{t,k}} \cdot \frac{\partial c_{t,j}}{\partial \hat{c}_{t,l}} = 0. \quad (\text{A52})$$

1291 According to the aforementioned results, for any two different entries $\hat{c}_{t,k}, \hat{c}_{t,l} \in \hat{\mathbf{c}}_t$ that are **not**
 1292 **adjacent** in the Markov network $\mathcal{M}_{\hat{\mathbf{c}}_t}$ over estimated $\hat{\mathbf{c}}_t$, we draw the following conclusions.

1293 (i) Equation (A49) implies that, each ground-truth latent variable $c_{t,i} \in \mathbf{c}_t$ is a function of at most
 1294 one of $\hat{c}_{t,k}$ and $\hat{c}_{t,l}$,

1295 (ii) Equation (A52) implies that, for each pair of ground-truth latent variables $c_{t,i}$ and $c_{t,j}$ that are
 1296 **adjacent** in $\mathcal{M}_{\mathbf{c}_t}$ over \mathbf{c}_t , they can not be a function of $\hat{c}_{t,k}$ and $\hat{c}_{t,l}$ respectively.

1297 **Step2: Isomorphism of Markov Networks** First, we demonstrate that there always exists a row
 1298 permutation for each invertible matrix such that the permuted diagonal entries are non-zero [Zhang
 1299 et al., 2024a]. By contradiction, if the product of the diagonal entry of an invertible matrix A is zero
 1300 for every row permutation, then we have Equation

$$\det(A) = \sum_{\sigma \in \mathcal{S}_n} \left(\text{sgn}(\sigma) \prod_{i=1}^n a_{\sigma(i),i} \right), \quad (\text{A53})$$

1301 by the Leibniz formula, where \mathcal{S}_n is the set of n -permutations. Thus, we have

$$\prod_{i=1}^n a_{\sigma(i),i} = 0, \quad \forall \sigma \in \mathcal{S}_n, \quad (\text{A54})$$

1302 which indicates that $\det(A) = 0$ and A is non-invertible. It contradicts the assumption that A is
 1303 invertible, and a row permutation where the permuted diagonal entries are non-zero must exist. Since

1304 h_z is invertible, for \mathbf{z}_t at time step t , there exists a permuted version of the estimated latent variables,
 1305 such that

$$\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_t(i)}} \neq 0, \quad i = 1, \dots, n, \quad (\text{A55})$$

1306 where π_t is the corresponding permutation at time step t . Since $\mathbf{c}_t = \{\mathbf{z}_{t-1}, \mathbf{z}_t, \mathbf{z}_{t+1}\}$, by applying
 1307 $\pi_{t-1}, \pi_t, \pi_{t+1}$, we have π' such that

$$\frac{\partial c_{t,i}}{\partial \hat{c}_{t,\pi'(i)}} \neq 0, \quad i = 1, \dots, 3n. \quad (\text{A56})$$

1308 Second, we demonstrate that $\mathcal{M}_{\mathbf{c}_t}$ is identical to $\mathcal{M}_{\hat{\mathbf{c}}_t^{\pi'}}$, where $\mathcal{M}_{\hat{\mathbf{c}}_t^{\pi'}}$ denotes the Markov network of
 1309 the permuted version of $\pi'(\hat{\mathbf{c}}_t)$.

1310 **Step3: Component-wise Identification of Latent Variables** Finally, we prove that the latent
 1311 variables are component-wise identifiable. On the one hand, for any pair of (i, j) such that $c_{t,i}, c_{t,j}$
 1312 are **adjacent** in $\mathcal{M}_{\mathbf{c}_t}$ while $\hat{c}_{t,\pi'(i)}, \hat{c}_{t,\pi'(j)}$ are **not adjacent** in $\mathcal{M}_{\hat{\mathbf{c}}_t^{\pi'}}$, according to Equation (A52),
 1313 we have $\frac{\partial c_{t,i}}{\partial \hat{c}_{t,\pi'(i)}} \cdot \frac{\partial c_{t,j}}{\partial \hat{c}_{t,\pi'(j)}} = 0$, which is a contradiction with how π' is constructed. Thus, any
 1314 edge presents in $\mathcal{M}_{\mathbf{c}_t}$ must exist in $\mathcal{M}_{\hat{\mathbf{c}}_t^{\pi'}}$. On the other hand, since observational equivalence can
 1315 be achieved by the true latent process $(g, f, p_{\mathbf{c}_t})$, the true latent process is clearly the solution with
 1316 minimal edges.

1317 Under the sparsity constraint on the edges of $\mathcal{M}_{\hat{\mathbf{c}}_t^{\pi'}}$, the permuted estimated Markov network $\mathcal{M}_{\hat{\mathbf{c}}_t^{\pi'}}$
 1318 must be identical to the true Markov network $\mathcal{M}_{\mathbf{c}_t}$. Thus, we claim that

1319 (i) the estimated Markov network $\mathcal{M}_{\hat{\mathbf{c}}_t}$ is isomorphic to the ground-truth Markov network $\mathcal{M}_{\mathbf{c}_t}$.

1320 Sequentially, under the same permutation π_t , we further give the proof that $z_{t,i}$ is only the function of
 1321 $\hat{z}_{t,\pi_t(i)}$. Since the permutation happens on each time step respectively, the cross-time disentanglement
 1322 is prevented clearly.

1323 Now let us focus on instantaneous disentanglement. Suppose there exists a pair of indices $i, j \in$
 1324 $\{1, \dots, n\}$. According to Equation (A55), we have $\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_t(i)}} = 0$ and $\frac{\partial z_{t,j}}{\partial \hat{z}_{t,\pi_t(j)}} = 0$. Let us discuss it
 1325 case by case.

- 1326 • If $z_{t,i}$ is not adjacent to $z_{t,j}$, we have $\hat{z}_{t,\pi_t(i)}$ is not adjacent to $\hat{z}_{t,\pi_t(j)}$ as well according
 1327 to the conclusion of identical Markov network. Using Equation (A49), we have $\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_t(i)}} \cdot$
 1328 $\frac{\partial z_{t,j}}{\partial \hat{z}_{t,\pi_t(j)}} = 0$, which leads to $\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_t(j)}} = 0$.
- 1329 • If $z_{t,i}$ is adjacent to $z_{t,j}$, we have $\hat{z}_{t,\pi_t(i)}$ is adjacent to $\hat{z}_{t,\pi_t(j)}$. When the Assumption A3
 1330 (Sparse Latent Process) is assured, i.e., the intimate neighbor set of $z_{t,i}$ is empty, there
 1331 exists at least one pair of (t', k) such that $z_{t',k}$ is adjacent to $z_{t,i}$ but not adjacent to $z_{t,j}$.
 1332 Similarly, we have the same structure on the estimated Markov network, which means that
 1333 $\hat{z}_{t',\pi_{t'}(k)}$ is adjacent to $\hat{z}_{t,\pi_t(i)}$ but not adjacent to $\hat{z}_{t,\pi_t(j)}$. Using Equation (A52) we have
 1334 $\frac{\partial z_{t,k}}{\partial \hat{z}_{t',\pi_{t'}(k)}} \cdot \frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_t(j)}} = 0$, which leads to $\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_t(j)}} = 0$.

1335 In conclusion, we always have $\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_t(j)}} = 0$. Thus, we have reached the conclusion that

1336 (ii) there exists a permutation π of the estimated latent variables, such that $z_{t,i}$ and $\hat{z}_{t,\pi(i)}$ is one-to-one
 1337 corresponding, i.e., $z_{t,i}$ is component-wise identifiable.

1338 □

1339 **Theorem A3. (Component-wise Identification of \mathbf{z}_t under sparse mixing procedure.)** For a series
 1340 of observations $\mathbf{x}_t \in \mathbb{R}^n$ and estimated latent variables $\hat{\mathbf{z}}_t \in \mathbb{R}^n$ with the corresponding process
 1341 $\hat{f}_i, \hat{p}(\epsilon), \hat{g}$, suppose the marginal distribution of observed variables is matched. Let $\mathcal{M}_{\mathbf{u}_t}$ be the
 1342 Markov network over $\mathbf{u}_t \triangleq \{\mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \mathbf{z}_t, \mathbf{x}_t\}$ and $\mathcal{M}_{\mathbf{u}_t}$. Besides the similar assumptions like
 1343 smooth, positive density, and sufficient variability assumptions, we further assume:

1344 • A7 (Sparse Mixing Procedure): For any $\mathbf{z}_{it} \in \mathbf{z}_t$, the intimate neighbor set of \mathbf{z}_{it} is an empty set.

1345 When the observational equivalence is achieved with the minimal number of edges of the estimated
 1346 mixing procedure, there exists a permutation π of the estimated latent variables, such that \mathbf{z}_{it} and
 1347 $\hat{\mathbf{z}}_{\pi(i)t}$ is one-to-one corresponding, i.e., \mathbf{z}_{it} is component-wise identifiable.

1348 *Proof.* By reusing Theorem 2 with more observations, $(\mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \mathbf{z}_t, \mathbf{x}_t)$ is also block-wise identi-
 1349 fiable. So we have:

$$\begin{aligned} p(\hat{\mathbf{z}}_t, \hat{\mathbf{x}}_t, \hat{\mathbf{z}}_{t-1}, \hat{\mathbf{x}}_{t-1}) &= p(\mathbf{z}_t, \mathbf{x}_t, \hat{\mathbf{z}}_{t-1}, \hat{\mathbf{x}}_{t-1}) |\mathbf{J}_h| \\ \iff p(\hat{\mathbf{z}}_t, \hat{\mathbf{x}}_t | \hat{\mathbf{z}}_{t-1}, \hat{\mathbf{x}}_{t-1}) &= p(\mathbf{z}_t, \mathbf{x}_t | \hat{\mathbf{z}}_{t-1}, \hat{\mathbf{x}}_{t-1}) |\mathbf{J}_h| \\ \iff \ln p(\hat{\mathbf{z}}_t, \hat{\mathbf{x}}_t | \hat{\mathbf{z}}_{t-1}, \hat{\mathbf{x}}_{t-1}) &= \ln p(\mathbf{z}_t, \mathbf{x}_t | \hat{\mathbf{z}}_{t-1}, \hat{\mathbf{x}}_{t-1}) + \ln |\mathbf{J}_h|, \end{aligned} \quad (\text{A57})$$

1350 where $h : \mathcal{X}, \mathcal{Z} \rightarrow \mathcal{X}, \mathcal{Z}$ denotes the invertible transformation. $|\mathbf{J}_h|$ stands for the absolute value of
 1351 the Jacobian matrix determinant of h . For any $\hat{\mathbf{z}}_{t,j}$, suppose that there exist $\hat{\mathbf{x}}_{t,k}$ that $\hat{\mathbf{z}}_{t,j}$ does not
 1352 contribute to the mixture of $\hat{\mathbf{x}}_{t,k}$.

1353 By using the sparse mixing procedure assumption (A7), we can constrain the sparsity of the estimated
 1354 mixing function, such that there exist two different estimated latent variables $\hat{\mathbf{u}}_{t,k}$ and $\hat{\mathbf{u}}_{t,l}$ that are
 1355 not adjacent in the estimated Markov networks $\mathcal{M}_{\mathbf{u}_t}$ and $\frac{\partial^2 \log p(\hat{\mathbf{u}}_t | \hat{\mathbf{z}}_{t-1}, \hat{\mathbf{x}}_{t-1})}{\partial \hat{\mathbf{u}}_{t,k} \partial \hat{\mathbf{u}}_{t,l}} = 0$. Sequentially, we
 1356 can replace $p(\mathbf{c} | \mathbf{z}_{t-1})$ in Lemma 1 with $p(\hat{\mathbf{u}}_t | \hat{\mathbf{z}}_{t-1}, \hat{\mathbf{x}}_{t-1})$, and then by reusing the proof process of
 1357 Lemma 1, we can prove that \mathbf{z} is component-wise identifiable.

1358 □

1359 C.8 More Discussion on the Sparse Mixing Procedure

1360 Although recent works like Zheng et al. [2022], Zheng and Zhang [2023] also utilize the sparse
 1361 mixing process from \mathbf{z}_t to \mathbf{x}_t to achieve identifiability, our assumption is easier to satisfy compared
 1362 to these methods. The primary reason for this is that our generative process allows for noise in the
 1363 mixing process from \mathbf{z}_t to \mathbf{x}_t , thereby accounting for measurement errors in the observed data. In
 1364 contrast, methods like Zheng et al. [2022], Zheng and Zhang [2023] require the additional assumption
 1365 that the mixing process is invertible and free from noise.

1366 D Experiment Details

1367 D.1 Synthetic Experiment

1368 D.1.1 Data Generation Process

1369 We follow Equation (1) to generate the synthetic data. As for the temporally latent processes, we use
 1370 MLPs with the activation function of LeakyReLU to model the sparse time-delayed. That is:

$$\mathbf{z}_{t,i} = (\text{LeakyReLU}(W_{i,:} \cdot \mathbf{z}_{t-1}, 0.2) + V_{<i,i} \cdot \mathbf{z}_{t,<i}) \cdot \epsilon_{t,i} + \epsilon_{t,i}^{\mathbf{z}} \quad (\text{A58})$$

1371 where $W_{i,:}$ is the i -th row of W^* and $V_{<i,i}$ is the first $i - 1$ columns in the i -th row of V . Moreover,
 1372 each independent noise $\epsilon_{t,i}$ is sampled from the distribution of normal distribution. We further let the
 1373 data generation process from latent variables to observed variables be MLPs with LeakyReLU units.
 1374 And the generation procedure can be formulated as follows:

$$\mathbf{x}_t = \text{LeakyReLU}(\text{LeakyReLU}(0.2 \times \text{LeakyReLU}(\mathbf{x}_{t-1} \cdot W_{\mathbf{x}}, 0.2) + \mathbf{z}_t + \epsilon_t^{\mathbf{o}}, 0.2) \cdot W_m), \quad (\text{A59})$$

1375 where $W_{\mathbf{x}}$ and W_m denote the weights of mixing function. We provide 4 datasets from A to D, whose
 1376 settings are shown in Table A5.

1377 The total size of the dataset is 100,000, with 1,024 samples designated as the validation set. The
 1378 remaining samples are the training set.

1379 D.1.2 Evaluation Metric

1380 To evaluate the identifiability performance of our method under instantaneous dependencies, we
 1381 employ the Mean Correlation Coefficient (MCC) between the ground-truth \mathbf{z}_t and the estimated $\hat{\mathbf{z}}_t$.

Table A5: Details of different synthetic datasets.

	Dimension of Latent Variables	Time Lag	Causal Edge among Observations
A	5	1	yes
B	5	1	no
C	5	2	yes
D	10	1	yes

A higher MCC denotes a better identification performance the model can achieve. In addition, we also draw the estimated latent causal process to validate our method. Since the estimated transition function will be a transformation of the ground truth, we do not compare their exact values, but only the activated entries.

D.1.3 Prior Likelihood Derivation

We first consider the prior of $\ln p(\mathbf{z}_{1:t})$. We start with an illustrative example of stationary latent causal processes with two time-delay latent variables, i.e. $\mathbf{z}_t = [z_{t,1}, z_{t,2}]$ with maximum time lag $L = 1$, i.e., $z_{t,i} = f_i(\mathbf{z}_{t-1}, \epsilon_{t,i})$ with mutually independent noises. Then we write this latent process as a transformation map \mathbf{f} (note that we overload the notation f for transition functions and for the transformation map):

$$\begin{bmatrix} z_{t-1,1} \\ z_{t-1,2} \\ z_{t,1} \\ z_{t,2} \end{bmatrix} = \mathbf{f} \left(\begin{bmatrix} z_{t-1,1} \\ z_{t-1,2} \\ \epsilon_{t,1} \\ \epsilon_{t,2} \end{bmatrix} \right).$$

By applying the change of variables formula to the map \mathbf{f} , we can evaluate the joint distribution of the latent variables $p(z_{t-1,1}, z_{t-1,2}, z_{t,1}, z_{t,2})$ as

$$p(z_{t-1,1}, z_{t-1,2}, z_{t,1}, z_{t,2}) = \frac{p(z_{t-1,1}, z_{t-1,2}, \epsilon_{t,1}, \epsilon_{t,2})}{|\det \mathbf{J}_{\mathbf{f}}|}, \quad (\text{A60})$$

where $\mathbf{J}_{\mathbf{f}}$ is the Jacobian matrix of the map \mathbf{f} , where the instantaneous dependencies are assumed to be a low-triangular matrix:

$$\mathbf{J}_{\mathbf{f}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{\partial z_{t,1}}{\partial z_{t-1,1}} & \frac{\partial z_{t,1}}{\partial z_{t-1,2}} & \frac{\partial z_{t,1}}{\partial \epsilon_{t,1}} & 0 \\ \frac{\partial z_{t,2}}{\partial z_{t-1,1}} & \frac{\partial z_{t,2}}{\partial z_{t-1,2}} & \frac{\partial z_{t,2}}{\partial \epsilon_{t,1}} & \frac{\partial z_{t,2}}{\partial \epsilon_{t,2}} \end{bmatrix}.$$

Given that this Jacobian is triangular, we can efficiently compute its determinant as $\prod_i \frac{\partial z_{t,i}}{\partial \epsilon_{t,i}}$. Furthermore, because the noise terms are mutually independent, and hence $\epsilon_{t,i} \perp \epsilon_{t,j}$ for $j \neq i$ and $\epsilon_t \perp \mathbf{z}_{t-1}$, so we can with the RHS of Equation (A60) as follows

$$p(z_{t-1,1}, z_{t-1,2}, z_{t,1}, z_{t,2}) = p(z_{t-1,1}, z_{t-1,2}) \times \frac{p(\epsilon_{t,1}, \epsilon_{t,2})}{|\mathbf{J}_{\mathbf{f}}|} = p(z_{t-1,1}, z_{t-1,2}) \times \frac{\prod_i p(\epsilon_{t,i})}{|\mathbf{J}_{\mathbf{f}}|}. \quad (\text{A61})$$

Finally, we generalize this example and derive the prior likelihood below. Let $\{r_i\}_{i=1,2,3,\dots}$ be a set of learned inverse transition functions that take the estimated latent causal variables, and output the noise terms, i.e., $\hat{\epsilon}_{t,i} = r_i(\hat{\mathbf{z}}_{t,i}, \{\hat{\mathbf{z}}_{t-\tau}\})$. Then we design a transformation $\mathbf{A} \rightarrow \mathbf{B}$ with low-triangular Jacobian as follows:

$$\underbrace{[\hat{\mathbf{z}}_{t-L}, \dots, \hat{\mathbf{z}}_{t-1}, \hat{\mathbf{z}}_t]^\top}_{\mathbf{A}} \text{ mapped to } \underbrace{[\hat{\mathbf{z}}_{t-L}, \dots, \hat{\mathbf{z}}_{t-1}, \hat{\epsilon}_{t,i}]^\top}_{\mathbf{B}}, \text{ with } \mathbf{J}_{\mathbf{A} \rightarrow \mathbf{B}} = \begin{bmatrix} \mathbb{I}_{n_s \times L} & 0 \\ * & \text{diag} \left(\frac{\partial r_{i,j}}{\partial \hat{\mathbf{z}}_{t,j}} \right) \end{bmatrix}. \quad (\text{A62})$$

Similar to Equation (A61), we can obtain the joint distribution of the estimated dynamics subspace as:

$$\log p(\mathbf{A}) = \underbrace{\log p(\hat{\mathbf{z}}_{t-L}, \dots, \hat{\mathbf{z}}_{t-1})}_{\text{Because of mutually independent noise assumption}} + \sum_{i=1}^{n_s} \log p(\hat{\epsilon}_{t,i}) + \log(|\det(\mathbf{J}_{\mathbf{A} \rightarrow \mathbf{B}})|) \quad (\text{A63})$$

1405 Finally, we have:

$$\log p(\hat{\mathbf{z}}_t | \{\hat{\mathbf{z}}_{t-\tau}\}_{\tau=1}^L) = \sum_{i=1}^{n_s} p(\hat{e}_{t,i}) + \sum_{i=1}^{n_s} \log \left| \frac{\partial r_i}{\partial \hat{z}_{t,i}} \right| \quad (\text{A64})$$

1406 Since the prior of $p(\hat{\mathbf{z}}_{t+1:T} | \hat{\mathbf{z}}_{1:t}) = \prod_{i=t+1}^T p(\hat{\mathbf{z}}_i | \hat{\mathbf{z}}_{i-1})$ with the assumption of first-order Markov
 1407 assumption, we can estimate $p(\hat{\mathbf{z}}_{t+1:T} | \hat{\mathbf{z}}_{1:t})$ in a similar way.

1408 D.1.4 Evident Lower Bound

1409 In this subsection, we show the evident lower bound. We first factorize the conditional distribution
 1410 according to the Bayes theorem.

$$\begin{aligned} \ln p(\mathbf{x}_{1:T}) &= \ln \frac{p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T})}{p(\mathbf{z}_{1:T} | \mathbf{x}_{1:T})} = \mathbb{E}_{q(\mathbf{z}_{1:T} | \mathbf{x}_{1:t})} \ln \frac{p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) q(\mathbf{z}_{1:T} | \mathbf{x}_{1:t})}{p(\mathbf{z}_{1:T} | \mathbf{x}_{1:T}) q(\mathbf{z}_{1:T} | \mathbf{x}_{1:t})} \\ &\geq \underbrace{\mathbb{E}_{q(\mathbf{z}_{1:T} | \mathbf{x}_{1:t})} \ln p(\mathbf{x}_{1:T} | \mathbf{z}_{1:T})}_{L_r \text{ and } L_y} - \underbrace{D_{KL}(q(\mathbf{z}_{1:T} | \mathbf{x}_{1:t}) || p(\mathbf{z}_{1:T}))}_{L_{KL}^z} = ELBO. \end{aligned} \quad (\text{A65})$$

1411 D.1.5 More Synthetic Experiment Results

1412 We repeat each experiment with different random seeds. We further consider CariNG as baselines,
 1413 experiment results are shown in Table A6.

Table A6: MCC results of synthetic datasets.

	TOT	IDOL	CariNG	TDRL
A	0.9258(0.0034)	0.3788(0.0245)	0.7354(0.0346)	0.3572(0.0523)
B	0.9324(0.0078)	0.8593(0.0092)	0.0823(0.0092)	0.8073(0.0786)
C	0.9322(0.0052)	0.6073(0.0952)	0.7084(0.0361)	0.7134(0.0346)
D	0.8433(0.0140)	0.7800(0.0387)	0.7371(0.0804)	0.7747(0.0690)

1414 D.2 Real-world Experiment

1415 D.2.1 Dataset Description

- 1416 • **ETT** [Zhou et al. \[2021\]](#) is an electricity transformer temperature dataset collected from two
 1417 separated counties in China, which contains two separate datasets {ETTh2, ETTm1} for
 1418 one hour level.
- 1419 • **Exchange** [Lai et al. \[2018\]](#) is the daily exchange rate dataset from of eight foreign coun-
 1420 tries including Australia, British, Canada, Switzerland, China, Japan, New Zealand, and
 1421 Singapore ranging from 1990 to.
- 1422 • **ECL** ³ is an electricity consuming load dataset with the electricity consumption (kWh)
 1423 collected from 321 clients.
- 1424 • **Traffic** ⁴ is a dataset of traffic speeds collected from the California Transportation Agencies
 1425 (CalTrans) Performance Measurement System (PeMS), which contains data collected from
 1426 325 sensors located throughout the Bay Area.
- 1427 • **Weather** ⁵ provides 10-minute summaries from an automated rooftop station at the Max
 1428 Planck Institute for Biogeochemistry in Jena, Germany.

1429 D.2.2 Implementation Details

1430 The implementations of our method based on different backbones are shown in Table A7 to A11.

³<https://archive.ics.uci.edu/dataset/321/electricityloadaddiagrams20112014>

⁴<https://pems.dot.ca.gov/>

⁵<https://www.bgc-jena.mpg.de/wetter/>

Table A7: LSTD+ToT Architecture details. T , length of time series. $|\mathbf{x}_t|$: input dimension. n : latent dimension. LeakyReLU: Leaky Rectified Linear Unit. ReLU: Rectified Linear Unit. Tanh: Hyperbolic tangent function.

Configuration	Description	Output
ϕ	Latent Variable Encoder	
Input: $\mathbf{x}_{1:t}$	Observed time series	Batch Size $\times t \times \mathbf{x}$ dimension
Dense	Conv1d	Batch Size $\times 640 \times \mathbf{x}_t $
Dense	t neurons, LeakyReLU	Batch Size $\times t \times \mathbf{x}_t $
Dense	T neurons, LeakyReLU	Batch Size $\times T \times \mathbf{x}_t $
Dense	512 neurons, LeakyReLU	Batch Size $\times 512 \times \mathbf{x}_t $
Dense	t neurons, LeakyReLU	Batch Size $\times t \times \mathbf{x}_t $
Dense	T neurons, LeakyReLU	Batch Size $\times T \times \mathbf{x}_t $
ψ	Latent Variable Decoder	
Input: $\mathbf{z}_{1:t}$	Latent Variable	Batch Size $\times t \times 2 \times \mathbf{x}_t $
Dense	$ \mathbf{x}_t $ neurons, LeakyReLU	Batch Size $\times t \times \mathbf{x}_t $
η	Dimensionality reduction	
Input: $\mathbf{x}_{1:t}$	Observed time series	Batch Size $\times t \times \mathbf{x}_t $
Linear	t neurons, ReLU	Batch Size $\times t \times \mathbf{x}_t $
φ	Regressor	
Input: $[\mathbf{z}_{1:T}; \eta(\mathbf{x}_{1:t})]$	Latent Variable	Batch Size $\times T \times 2 \times \mathbf{x}_t $
Dense	512 neurons, LeakyReLU	Batch Size $\times T \times 512$
Dense	$ \mathbf{x}_t $ neurons, LeakyReLU	Batch Size $\times T \times \mathbf{x}_t $
$r_i^{\mathbf{z}}$	Latent Transition Estimator	
Input: $\mathbf{z}_{1:T}$	Latent Variable	Batch Size $\times (n+1)$
Dense	128 neurons, LeakyReLU	$(n+1) \times 128$
Dense	128 neurons, LeakyReLU	128×128
Dense	128 neurons, LeakyReLU	128×128
Dense	1 neuron	Batch Size $\times 1$
Jacobian Compute	Compute $\log(\det(\mathbf{J}))$	Batch Size
$r_i^{\mathbf{o}}$	Observed Transition Estimator	
Input: $\mathbf{z}_{1:T}$	Latent Variable	Batch Size $\times (n+1)$
Dense	128 neurons, LeakyReLU	$(n+1) \times 128$
Dense	128 neurons, LeakyReLU	128×128
Dense	128 neurons, LeakyReLU	128×128
Dense	1 neuron	Batch Size $\times 1$
Jacobian Compute	Compute $\log(\det(\mathbf{J}))$	Batch Size

Table A8: OneNet+ToT Architecture details. T , length of time series. $|\mathbf{x}_t|$: input dimension. n : latent dimension. LeakyReLU: Leaky Rectified Linear Unit. ReLU: Rectified Linear Unit. Tanh: Hyperbolic tangent function.

Configuration	Description	Output
ϕ	Latent Variable Encoder	
Input: $\mathbf{x}_{1:t}$	Observed time series	Batch Size $\times t \times \mathbf{x}$ dimension
Linear	n neurons	Batch Size $\times n \times \mathbf{x}_t $
Convolution neural networks	320 neurons	Batch Size $\times 320 \times \mathbf{x}_t $
ψ	Latent Variable Decoder	
Input: $\mathbf{z}_{1:t}$	Latent Variable	Batch Size $\times 320 \times \mathbf{x}_t $
Linear	t neurons, ReLU	Batch Size $\times t \times \mathbf{x}_t $
η	Dimensionality reduction	
Input: $\mathbf{x}_{1:t}$	Observed time series	Batch Size $\times t \times \mathbf{x}_t $
Linear	320 neurons, ReLU	Batch Size $\times 320 \times \mathbf{x}_t $
φ	Regressor	
Input: $[\mathbf{z}_{1:T}; \eta(\mathbf{x}_{1:t})]$	Latent Variable	Batch Size $\times 640 \times \mathbf{x}_t $
Linear	$ \mathbf{x}_t $ neurons, ReLU	Batch Size $\times T \times \mathbf{x}_t $
Linear	n neurons, ReLU	Batch Size $\times n \times \mathbf{x}_t $
Convolution neural networks	320 neurons	Batch Size $\times 320 \times \mathbf{x}_t $
Linear	T neurons, ReLU	Batch Size $\times T \times \mathbf{x}_t $
$r_i^{\mathbf{z}}$	Latent Transition Estimator	
Input: $\mathbf{z}_{1:T}$	Latent Variable	Batch Size $\times (n+1)$
Dense	128 neurons, LeakyReLU	$(n+1) \times 128$
Dense	128 neurons, LeakyReLU	128×128
Dense	128 neurons, LeakyReLU	128×128
Dense	1 neuron	Batch Size $\times 1$
Jacobian Compute	Compute $\log(\det(\mathbf{J}))$	Batch Size
$r_i^{\mathbf{o}}$	Observed Transition Estimator	
Input: $\mathbf{z}_{1:T}$	Latent Variable	Batch Size $\times (n+1)$
Dense	128 neurons, LeakyReLU	$(n+1) \times 128$
Dense	128 neurons, LeakyReLU	128×128
Dense	128 neurons, LeakyReLU	128×128
Dense	1 neuron	Batch Size $\times 1$
Jacobian Compute	Compute $\log(\det(\mathbf{J}))$	Batch Size

Table A9: OneNet-T+TOT Architecture details. T , length of time series. $|\mathbf{x}_t|$: input dimension. n : latent dimension. LeakyReLU: Leaky Rectified Linear Unit. ReLU: Rectified Linear Unit. Tanh: Hyperbolic tangent function.

Configuration	Description	Output
ϕ	Latent Variable Encoder	
Input: $\mathbf{x}_{1:t}$	Observed time series	Batch Size $\times t \times \mathbf{x}$ dimension
Linear	n neurons, ReLU	Batch Size $\times n \times \mathbf{x}_t $
Dilation convolution	T neurons, 10 layers	Batch Size $\times T \times \mathbf{x}_t $
ψ	Latent Variable Decoder	
Input: $\mathbf{z}_{1:t}$	Latent Variable	Batch Size $\times 320 \times \mathbf{x}_t $
Linear	t neurons, ReLU	Batch Size $\times t \times \mathbf{x}_t $
η	Dimensionality reduction	
Input: $\mathbf{x}_{1:t}$	Observed time series	Batch Size $\times t \times \mathbf{x}_t $
Linear	n neurons, ReLU	Batch Size $\times n \times \mathbf{x}_t $
Dilation convolution	T neurons, 10 layers	Batch Size $\times T \times \mathbf{x}_t $
φ	Regressor	
Input: $[\mathbf{z}_{1:T}; \eta(\mathbf{x}_{1:t})]$	Latent Variable	Batch Size $\times T \times 2 \times \mathbf{x}_t $
Linear	$ \mathbf{x}_t $ neurons, ReLU	Batch Size $\times T \times \mathbf{x}_t $
Linear	n neurons, ReLU	Batch Size $\times n \times \mathbf{x}_t $
Convolution neural networks	320 neurons	Batch Size $\times 320 \times \mathbf{x}_t $
Linear	T neurons, ReLU	Batch Size $\times T \times \mathbf{x}_t $
$r_i^{\mathbf{z}}$	Latent Transition Estimator	
Input: $\mathbf{z}_{1:T}$	Latent Variable	Batch Size $\times (n+1)$
Dense	128 neurons, LeakyReLU	$(n+1) \times 128$
Dense	128 neurons, LeakyReLU	128×128
Dense	128 neurons, LeakyReLU	128×128
Dense	1 neuron	Batch Size $\times 1$
Jacobian Compute	Compute $\log(\det(J))$	Batch Size
$r_i^{\mathbf{o}}$	Observed Transition Estimator	
Input: $\mathbf{z}_{1:T}$	Latent Variable	Batch Size $\times (n+1)$
Dense	128 neurons, LeakyReLU	$(n+1) \times 128$
Dense	128 neurons, LeakyReLU	128×128
Dense	128 neurons, LeakyReLU	128×128
Dense	1 neuron	Batch Size $\times 1$
Jacobian Compute	Compute $\log(\det(J))$	Batch Size

Table A10: online-T+TOT Architecture details. T , length of time series. $|\mathbf{x}_t|$: input dimension. n : latent dimension. LeakyReLU: Leaky Rectified Linear Unit. ReLU: Rectified Linear Unit. Tanh: Hyperbolic tangent function.

Configuration	Description	Output
ϕ	Latent Variable Encoder	
Input: $\mathbf{x}_{1:t}$	Observed time series	Batch Size $\times t \times \mathbf{x}$ dimension
Linear	n neurons, ReLU	Batch Size $\times n \times \mathbf{x}_t $
Convolution neural networks	320 neurons	Batch Size $\times 320 \times \mathbf{x}_t $
Linear	t neurons, ReLU	Batch Size $\times t \times \mathbf{x}_t $
ψ	Latent Variable Decoder	
Input: $\mathbf{z}_{1:t}$	Latent Variable	Batch Size $\times 320 \times \mathbf{x}_t $
Linear	t neurons, ReLU	Batch Size $\times t \times \mathbf{x}_t $
η	Dimensionality reduction	
Input: $\mathbf{x}_{1:t}$	Observed time series	Batch Size $\times t \times \mathbf{x}_t $
Linear	t neurons, ReLU	Batch Size $\times t \times \mathbf{x}_t $
φ	Regressor	
Input: $[\mathbf{z}_{1:T}; \eta(\mathbf{x}_{1:t})]$	Latent Variable	Batch Size $\times 2t \times \mathbf{x}_t $
Moving average	kernel size, stride=1	Batch Size $\times 2t \times \mathbf{x}_t $
Dilation convolution	320 neurons, 5 layers, ReLU	Batch Size $\times 320 \times \mathbf{x}_t $
Padding	patch length=6, ReLU	Batch Size $\times \mathbf{x}_t \times \text{patch length} \times \text{patch num}$
Transformer	n neurons	Batch Size $\times \mathbf{x}_t \times n \times \text{patch num}$
Linear	T neurons, ReLU	Batch Size $\times T \times \mathbf{x}_t $
r_i^z	Latent Transition Estimator	
Input: $\mathbf{z}_{1:T}$	Latent Variable	Batch Size $\times (n+1)$
Dense	128 neurons, LeakyReLU	$(n+1) \times 128$
Dense	128 neurons, LeakyReLU	128×128
Dense	128 neurons, LeakyReLU	128×128
Dense	1 neuron	Batch Size $\times 1$
Jacobian Compute	Compute $\log(\det(J))$	Batch Size
r_i^o	Observed Transition Estimator	
Input: $\mathbf{z}_{1:T}$	Latent Variable	Batch Size $\times (n+1)$
Dense	128 neurons, LeakyReLU	$(n+1) \times 128$
Dense	128 neurons, LeakyReLU	128×128
Dense	128 neurons, LeakyReLU	128×128
Dense	1 neuron	Batch Size $\times 1$
Jacobian Compute	Compute $\log(\det(J))$	Batch Size

Table A11: Proceed-T+TOT Architecture details. T , length of time series. $|\mathbf{x}_t|$: input dimension. n : latent dimension. LeakyReLU: Leaky Rectified Linear Unit. ReLU: Rectified Linear Unit. Tanh: Hyperbolic tangent function.

Configuration	Description	Output
ϕ	Latent Variable Encoder	
Input: $\mathbf{x}_{1:t}$	Observed time series	Batch Size $\times t \times$ dimension
Linear	n neurons, ReLU	Batch Size $\times n \times \mathbf{x}_t $
Dilation convolution	T neurons, 10 layers	Batch Size $\times T \times \mathbf{x}_t $
ψ	Latent Variable Decoder	
Input: $\mathbf{z}_{1:t}$	Latent Variable	Batch Size $\times 320 \times \mathbf{x}_t $
Linear	t neurons, ReLU	Batch Size $\times t \times \mathbf{x}_t $
η	Dimensionality reduction	
Input: $\mathbf{x}_{1:t}$	Observed time series	Batch Size $\times t \times \mathbf{x}_t $
Linear	n neurons, ReLU	Batch Size $\times n \times \mathbf{x}_t $
Dilation convolution	T neurons, 10 layers	Batch Size $\times T \times \mathbf{x}_t $
φ	Regressor	
Input: $[\mathbf{z}_{1:T}; \eta(\mathbf{x}_{1:t})]$	Latent Variable	Batch Size $\times T \times 2 \times \mathbf{x}_t $
Linear	$ \mathbf{x}_t $ neurons, ReLU	Batch Size $\times T \times \mathbf{x}_t $
r_i^z	Latent Transition Estimator	
Input: $\mathbf{z}_{1:T}$	Latent Variable	Batch Size $\times (n+1)$
Dense	128 neurons, LeakyReLU	$(n+1) \times 128$
Dense	128 neurons, LeakyReLU	128×128
Dense	128 neurons, LeakyReLU	128×128
Dense	1 neuron	Batch Size $\times 1$
Jacobian Compute	Compute $\log(\det(J))$	Batch Size
r_i^o	Observed Transition Estimator	
Input: $\mathbf{z}_{1:T}$	Latent Variable	Batch Size $\times (n+1)$
Dense	128 neurons, LeakyReLU	$(n+1) \times 128$
Dense	128 neurons, LeakyReLU	128×128
Dense	128 neurons, LeakyReLU	128×128
Dense	1 neuron	Batch Size $\times 1$
Jacobian Compute	Compute $\log(\det(J))$	Batch Size

1431 D.2.3 More Experiment Results

1432 We further consider MIR Aljundi et al. [2019a] and TFCL Aljundi et al. [2019b] as the backbone
1433 networks, experimental results are shown in Table A12.

Table A12: MSE and MAE results of different datasets on TFCL and MIR backbone.

Models	Len	TFCL		TFCL+TOT		MIR		MIR+TOT		Models	Len	TFCL		TFCL+TOT		MIR		MIR+TOT	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE			MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh2	1	0.557	0.472	0.463	0.382	0.486	0.41	0.447	0.378	ECL	1	2.732	0.524	3.815	0.44	2.575	0.504	3.396	0.589
	24	0.846	0.548	0.825	0.554	0.812	0.541	0.652	0.465		24	12.094	1.256	10.083	1.105	9.265	1.066	6.142	1.041
	48	1.208	0.592	0.87	0.555	1.103	0.565	0.842	0.526		48	12.11	1.303	10.685	1.075	9.411	1.079	6.479	1.090
ETTm1	1	0.087	0.198	0.081	0.187	0.085	0.197	0.083	0.188	Traffic	1	0.306	0.297	0.304	0.263	0.298	0.284	0.294	0.267
	24	0.211	0.341	0.186	0.32	0.192	0.325	0.132	0.267		24	0.441	0.493	0.389	0.314	0.451	0.443	0.39	0.339
	48	0.236	0.363	0.196	0.331	0.210	0.342	0.129	0.265		48	0.438	0.531	0.393	0.316	0.502	0.397	0.419	0.345
WTH	1	0.177	0.24	0.154	0.197	0.179	0.244	0.154	0.199	Exchange	1	0.106	0.153	0.045	0.142	0.095	0.118	0.056	0.162
	24	0.301	0.363	0.295	0.359	0.291	0.355	0.184	0.265		24	0.098	0.227	0.062	0.166	0.104	0.204	0.067	0.178
	48	0.323	0.382	0.294	0.36	0.297	0.361	0.195	0.278		48	0.101	0.183	0.098	0.207	0.101	0.209	0.047	0.137

1434 To demonstrate that the improvements of our approach are not due to an increase in parameters, we
1435 increase the number of parameters of the baseline methods by adding additional layers to the neural

Table A16: Standard deviation of MSE on different datasets.

Models	Len	LSTD	LSTD+ToT	Proceed-T	Proceed-T+ToT	OneNet	OneNet+TOT	OneNet-T	OneNet-T+TOT	MIR	MIR+TOT	Online-T	Online-T+TOT	TFCL	TFCL+TOT
ETTh2	1	0.0246	0.0031	0.1038	0.1351	0.0085	0.0041	0.0076	0.0233	0.1229	0.0168	0.0149	0.0535	0.1827	0.0049
	24	0.0260	0.0078	0.0413	0.0620	0.0137	0.0128	0.1563	0.0087	0.1618	0.0309	0.0098	0.0341	0.0699	0.0070
	48	0.0295	0.0494	0.1030	0.0414	0.0263	0.0145	0.0829	0.0137	0.0827	0.0561	0.0172	0.1233	0.1320	0.0206
ETTm1	1	0.00097	0.0004	0.0008	0.0006	0.0025	0.0007	0.0085	0.0023	0.1382	0.0019	0.0093	0.0007	0.1259	0.0018
	24	0.0003	0.0014	0.0197	0.0129	0.0018	0.0014	0.0025	0.0341	0.1107	0.0142	0.0080	0.0051	0.1643	0.0079
	48	0.0157	0.0012	0.0108	0.0049	0.0012	0.0044	0.0068	0.0093	0.0370	0.0173	0.0142	0.0151	0.0764	0.0063
WTH	1	0.00084	0.0003	0.0043	0.0040	0.0006	0.0004	0.0004	0.0157	0.1747	0.0006	0.0162	0.0001	0.1241	0.0057
	24	0.0023	0.0024	0.0025	0.0043	0.0026	0.0020	0.0023	0.0007	0.1212	0.0031	0.0167	0.0036	0.1782	0.0028
	48	0.0055	0.0044	0.0081	0.0050	0.0041	0.0191	0.0100	0.0112	0.1338	0.0035	0.0108	0.0035	0.1790	0.0035
ECL	1	0.0197	0.0189	0.0360	0.0673	0.0449	0.0415	0.0344	0.0731	0.0119	0.1042	0.0152	0.0582	0.0816	0.0484
	24	0.0256	0.0662	0.0232	0.0612	0.0205	0.0120	0.0326	0.1260	0.0051	0.0907	0.0178	0.0669	0.0104	0.7811
	48	0.2065	0.0278	0.2163	0.0867	0.1002	0.0908	0.0152	0.0246	0.0126	0.2878	0.0092	0.0799	0.0081	0.5162
Traffic	1	0.0027	0.0008	0.0186	0.0079	0.0010	0.0017	0.0015	0.0097	0.0149	0.0014	0.0462	0.0125	0.0052	0.0002
	24	0.0070	0.0065	0.0135	0.0034	0.0068	0.0654	0.0046	0.0186	0.0148	0.0064	0.0134	0.0271	0.0103	0.0045
	48	0.0024	0.0007	0.0063	0.0020	0.0289	0.0415	0.0015	0.0274	0.0159	0.0345	0.0180	0.0021	0.0175	0.0011
Exchange	1	0.0005	0.0002	0.0005	0.00003	0.0011	0.0008	0.0021	0.0017	0.0130	0.0101	0.0086	0.0008	0.0158	0.0084
	24	0.0004	0.0030	0.0027	0.0033	0.0064	0.0097	0.0010	0.0027	0.0070	0.0077	0.0137	0.0079	0.0139	0.0076
	48	0.0054	0.0023	0.0037	0.0035	0.0170	0.0137	0.0022	0.0062	0.0038	0.0130	0.0067	0.0031	0.0126	0.0017

Table A17: Standard deviation of MAE on different datasets.

Models	Len	LSTD	LSTD+ToT	Proceed-T	Proceed-T+ToT	OneNet	OneNet+TOT	OneNet-T	OneNet-T+TOT	MIR	MIR+TOT	Online-T	Online-T+TOT	TFCL	TFCL+TOT
ETTh2	1	0.0073	0.0004	0.0105	0.0031	0.0053	0.0008	0.0066	0.0120	0.0130	0.0040	0.0175	0.0049	0.0183	0.0013
	24	0.0032	0.0035	0.0094	0.0023	0.0061	0.0029	0.0174	0.0024	0.0408	0.0109	0.0066	0.0068	0.0470	0.0005
	48	0.0060	0.0226	0.0104	0.0038	0.0099	0.0004	0.0125	0.0019	0.0098	0.0092	0.0115	0.0222	0.0037	0.0065
ETTm1	1	0.0021	0.0008	0.0013	0.0006	0.0045	0.0008	0.0123	0.0044	0.0143	0.0033	0.0175	0.0017	0.0517	0.0046
	24	0.0008	0.0026	0.0096	0.0048	0.0022	0.0016	0.0020	0.0317	0.0526	0.0148	0.0048	0.0043	0.0571	0.0074
	48	0.0118	0.0021	0.0012	0.0020	0.0012	0.0049	0.0062	0.0087	0.0086	0.0173	0.0126	0.0115	0.0565	0.0067
WTH	1	0.0011	0.0006	0.0010	0.0020	0.0005	0.0008	0.0004	0.0222	0.0091	0.0013	0.0078	0.0003	0.0163	0.0045
	24	0.0010	0.0017	0.0004	0.0011	0.0027	0.0002	0.0023	0.0009	0.0078	0.0031	0.0157	0.0028	0.0089	0.0048
	48	0.0044	0.0039	0.0014	0.0019	0.0039	0.0183	0.0050	0.0060	0.0132	0.0031	0.0063	0.0034	0.0184	0.0075
ECL	1	0.0055	0.0010	0.0006	0.0012	0.0013	0.0016	0.0045	0.0391	0.0085	0.0084	0.0013	0.0014	0.0575	0.0044
	24	0.0044	0.0065	0.0004	0.0064	0.0025	0.0021	0.0001	0.0130	0.0581	0.0105	0.0152	0.0085	0.0176	0.0768
	48	0.0686	0.0079	0.0018	0.0032	0.0047	0.0038	0.0013	0.0027	0.0125	0.0132	0.0179	0.0022	0.0056	0.0226
Traffic	1	0.0028	0.0013	0.0044	0.0053	0.0006	0.0014	0.0014	0.0082	0.0072	0.0020	0.0105	0.0041	0.0125	0.0010
	24	0.0042	0.0046	0.0142	0.0105	0.0027	0.0411	0.0027	0.0088	0.0106	0.0077	0.0182	0.0137	0.0150	0.0066
	48	0.0021	0.0012	0.0044	0.0019	0.0144	0.0241	0.0013	0.0162	0.0028	0.0188	0.0017	0.0016	0.0165	0.0012
Exchange	1	0.0018	0.0006	0.0022	0.0006	0.0031	0.0024	0.0034	0.0048	0.0090	0.0072	0.0172	0.0034	0.0010	0.0061
	24	0.0004	0.0047	0.0018	0.0024	0.0095	0.0154	0.0014	0.0051	0.0139	0.0119	0.0012	0.0152	0.0051	0.0098
	48	0.0033	0.0036	0.0034	0.0018	0.0241	0.0210	0.0031	0.0089	0.0014	0.0153	0.0002	0.0054	0.0174	0.0025

E Broader Impacts

The proposed method for online time series forecasting presents a novel approach to address the challenges posed by distribution shifts in temporal data. By leveraging the identification of latent variables and their causal transitions, our framework demonstrates a provable reduction in Bayes risk, with significant improvements in forecasting accuracy, making the method highly applicable to real-time forecasting tasks in fields such as finance, healthcare, and energy management.

Our method not only outperforms existing models like IDOL and TDRL in both synthetic and real-world experiments, but it also enhances the scalability and adaptability of forecasting systems in dynamic environments. The theoretical advancements, coupled with the plug-and-play architecture, facilitate seamless integration into existing forecasting pipelines, further promoting the use of causal modeling in practical scenarios.

Furthermore, the broader implications of this work extend beyond time series forecasting. The ability to identify and utilize latent variables in real-time systems opens the door to new applications in domains such as anomaly detection, predictive maintenance, and environmental monitoring. With its potential for improving decision-making processes in critical industries, our method sets a strong foundation for future advancements in online forecasting and causal representation learning, thus contributing to the evolution of machine learning models that can effectively handle complex, dynamic data in real-world environments.