

942 *Supplement to* “ **Towards Identifiability of Hierarchical Temporal**
943 **Causal Representation Learning** ”

944

945	A Useful Definitions and Lemmas	23
946	B Proof of Theoretical Results	24
947	B.1 Notations	24
948	B.2 Proof of Block-wise Identifiability of Latent Variables \mathbf{z}_t in Hierarchical Latent Process	24
949	B.3 Examples of injective linear operators	27
950	B.4 Monotonicity and Normalization Assumption	27
951	B.5 Block-wise Identifiability of Latent Variables \mathbf{z}_t^l in any l -th Layer	27
952	B.6 Extension to Multiple Lags	31
953	B.7 Component-wise Identifiability of Latent Variables $\mathbf{z}_{t,i}^l$ in any l -th Layer	31
954	C Real-world Implications of Modeling Time Series Data in a Hierarchical Manner	32
955	D Relationship between Time Series Generation and Hierarchical Dynamics	33
956	E Related Works	33
957	F Implementation Details	34
958	F.1 Prior Likelihood Derivation	34
959	F.2 Evident Lower Bound	35
960	F.3 Model Details	35
961	F.3.1 Reproducibility of Simulation Experiment	35
962	F.3.2 Reproducibility of Real-world Experiment	35
963	G Experiment Details	35
964	G.1 Simulation Experiment	35
965	G.1.1 Data Generation Process	35
966	G.1.2 Evaluation Metrics	36
967	G.1.3 More Simulation Experiments of Different Length of Observations	37
968	G.2 Real-world Experiment	37
969	G.2.1 Dataset Description	37
970	G.2.2 Evaluation Metric	38
971	G.3 More Experiment Results	38
972	G.3.1 Experiment results on other datasets	38
973	G.3.2 Ablation Study	38
974	H Broader Impacts	39

975

A Useful Definitions and Lemmas

Definition 4. (Diagonal Operator) Consider two random variable a and b , density functions p_a and p_b are defined on some support \mathcal{A} and \mathcal{B} , respectively. The diagonal operator $D_{b|a}$ maps the density function p_a to another density function $D_{b|a} \circ p_a$ defined by the pointwise multiplication of the function $p_{b|a}$ at a fixed point b :

$$p_{b|a}(b | \cdot) p_a = D_{b|a} \circ p_a, \text{ where } D_{b|a} = p_{b|a}(b | \cdot). \quad (\text{A1})$$

Definition 5. (Completeness) A family of distribution $p(a|b)$ is complete if the only solution $p(a)$ to

$$\int_{\mathcal{A}} p(a) p_{a|b}(a|b) da = 0 \quad \text{for all } b \in \mathcal{B}, \quad (\text{A2})$$

is $p(a) = 0$. In other words, no matter the range of an operator is on finite or infinite, it is complete if its null space³ or kernel is a zero set. Completeness is always used to phrase the sufficient and necessary condition for injective linear operator [57, 5].

Theorem A1. (Theorem XV 4.5 in [10] Part III) A bounded operator T is a spectral operator if and only if it is the sum $T = S + N$ of a bounded scalar type operator S and a quasi-nilpotent operator N commuting with S . Furthermore, this decomposition is unique and T and S have the same spectrum and the same resolution of the identity.

Lemma A1. (Lemma 1 in [24]) Under Assumption 1, if $L_{z|x}$ is injective, then $L_{x|z}^{-1}$ exists and is densely defined over $\mathcal{G}(\mathcal{X})$ (for $\mathcal{G} = \mathcal{L}^1, \mathcal{L}_{\text{bnd}}^1$).

Properties of linear operator. We outline useful properties of the linear operator to facilitate understanding of our proof:

- i. (Inverse) If linear operator: $L_{b|a}$ exists a left-inverse $L_{b|a}^{-1}$, such $L_{b|a}^{-1} \circ L_{b|a} \circ p_a = p_a$ for all $a \in \mathcal{A}$. Analogously, if $L_{b|a}$ exists a right-inverse $L_{b|a}^{-1}$, such $L_{b|a} \circ L_{b|a}^{-1} \circ p_a = p_a$ for all $a \in \mathcal{A}$. If $L_{b|a}$ is bijective, there exists left-inverse and right-inverse which are the same.
- ii. (Injective) $L_{b|a}$ is said to be an injective linear operator if its $L_{b|a}^{-1}$ is defined over the range of the operator $L_{b|a}$ [43]. If so, under assumption 1 (ii), $L_{a|b}^{-1}$ exists and is densely defined over $\mathcal{F}(\mathcal{A})$. [24].
- iii. (Composition) Given two linear operators $L_{c|b} : \mathcal{F}(\mathcal{B}) \rightarrow \mathcal{F}(\mathcal{C})$ and $L_{c|a} : \mathcal{F}(\mathcal{A}) \rightarrow \mathcal{F}(\mathcal{C})$, with the function space supports defined uniformly on the range of supports for the domain spaces as characterized by $L_{b|a}$, it follows that $L_{c|a} = L_{c|b} \circ L_{b|a}$. Furthermore, the properties of linearity and associativity are preserved in the operation of linear operators. However, it is crucial to note the non-commutativity of these operators, i.e., $L_{c|b} L_{b|a} \neq L_{b|a} L_{c|b}$, indicating the significance of the order of application.

Definition 6 (Markov Network). Markov network is an undirected graph $G = (V, E)$ with a set of random variables $\mathbf{x}_{v \in V}$, where any two non-adjacent variables like x_a and x_b are conditionally independent given all other variables. That is,

$$x_a \perp x_b | \mathbf{x}_{V \setminus \{a, b\}}, \quad \forall (a, b) \notin E. \quad (\text{A3})$$

Markov Networks and Directed Acyclic Graphs (DAGs) are both graphical models employed to represent joint distributions and to illustrate conditional independence properties. Based on the definition of Markov Network, we further define Vector-Node Markov Network as follows:

Definition 7 (Vector-Node Markov Network). A Markov network where each node represents a vector, rather than a scalar variable. In this structure, the joint distribution is factorized over a graph $G = (V, E)$, where V is the set of nodes, and each $v \in V$ corresponds to a multidimensional vector \mathbf{x}_v . Given any two non-adjacent vectors \mathbf{x}_a and \mathbf{x}_b , any $\mathbf{x}_{a,i} \in \mathbf{x}_a$ and $\mathbf{x}_{b,j} \in \mathbf{x}_b$ are conditionally independent given all other variables. That is,

$$\mathbf{x}_{a,i} \perp \mathbf{x}_{b,j} | \mathbf{x}_{V \setminus \{a, b\}}, \quad \forall (a, b) \notin E. \quad (\text{A4})$$

³The null space or kernel of an operator L to be the set of all vectors which L maps to the zero vector: $\text{null } L = \{v \in V : Lv = 0\}$.

1016 **Definition 8** (Isomorphism of Markov networks). *We let the $V(\cdot)$ be the vertical set of any graphs,*
1017 *an isomorphism of Markov networks M and \hat{M} is a bijection between the vertex sets of M and \hat{M}*

$$f : V(M) \rightarrow V(\hat{M})$$

1018 *such that any two vertices u and v of M are adjacent in G if and only if $f(u)$ and $f(v)$ are adjacent*
1019 *in \hat{M} .*

1020 **Definition 9** (Definition of left and right inverse.). *Let $f : \mathbf{x} \rightarrow \mathbf{y}$ and $g : \mathbf{y} \rightarrow \mathbf{x}$ be functions. We*
1021 *say that g is a left inverse to f , and that f is a right inverse of g , if $g \circ f = \text{id}_X$, where id_X is the*
1022 *identity function.*

1023 B Proof of Theoretical Results

1024 B.1 Notations

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, s, o	Index of latent variables	$i, j, k, s, o \in \{1, \dots, n\}$
t	Time index	$t \in \mathbb{N}^+$
L	Number of latent layers	$L \in \mathbb{N}^+$ and $L \geq 2$
l	Index of latent layers	$l = \{1, 2, \dots, L\}$
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	$\mathbf{z}_t \in \mathbb{R}^n$
$z_{t,i}^l$	The i -th, l -th-layer latent variable at t -th step	$z_{t,i}^l \in \mathbb{R}$
\mathbf{c}_t	$\{\mathbf{z}_{t-1}, \mathbf{z}_t\}$	$\mathbf{c}_t \in \mathbb{R}^{2 \times n \times L}$
ϵ_t^0	Independent noise of mixing procedure	$\epsilon_t^0 \sim p_{\epsilon_t^0}$
$\epsilon_{t,i}^l$	Independent noise of the latent transition of $\mathbf{z}_{t,i}^l$	$\epsilon_{t,i}^l \sim p_{\epsilon_{t,i}^l}$
Function		
$p_{a b}(\cdot b)$	Density function of a given b	/
$p_{a,b c}(a, \cdot c)$	Joint density function of (a, b) given a and c	/
$\mathbf{pa}_d(\cdot)$	Time-delayed parents	/
$\mathbf{pa}_h(\cdot)$	Hierarchical parents	/
$\mathbf{g}(\cdot)$	Nonlinear mixing function	$\mathbb{R}^n \rightarrow \mathbb{R}^n$
$f_i^l(\cdot)$	Transition function of $\mathbf{z}_{t,i}^l$	$\mathbb{R}^{ \mathbf{Pa}_d(\mathbf{z}_{t,i}^l) + \mathbf{Pa}_h(\mathbf{z}_{t,i}^l) + 1} \rightarrow \mathbb{R}$
$h(\cdot)$	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	/
$\mathcal{M}_{\mathbf{c}_t}$	Markov network over \mathbf{c}_t	/
ϕ_l	Contextual hierarchical encoder for $\hat{\mathbf{z}}_t^l$	/
ψ_l	Step-wise Decoder	/
$r_{t,i}^l$	Noise estimator of $\epsilon_{t,i}^l$.	/
Symbol		
\mathbf{J}_κ	Jacobian matrix of r_t^l	/

1025

1026 B.2 Proof of Block-wise Identifiability of Latent Variables \mathbf{z}_t in Hierarchical Latent Process

1027 **Theorem A2.** (Block-wise Identifiability of Latent Variables \mathbf{z}_t in Hierarchical Latent Process.)
1028 *Suppose the observed and L -layer latent variables follow the data generation process in Figure*

1029 1. By matching the true joint distribution of $2L + 1$ number of adjacent observed variables, i.e.,
 1030 $\{\mathbf{x}_{t-L}, \dots, \mathbf{x}_t, \dots, \mathbf{x}_{t+L}\}$, we further make the following assumptions:

- 1031 • (i) The joint distribution of X, Z , and their marginal and conditional densities are bounded
 1032 and continuous.
- 1033 • (ii) The linear operators $L_{\mathbf{x}_{t+L}, \dots, \mathbf{x}_{t+1} | \mathbf{z}_t}$ and $L_{\mathbf{x}_{t-1}, \dots, \mathbf{x}_{t-L} | \mathbf{x}_{t+1}, \dots, \mathbf{x}_{t+L}}$ are injective for
 1034 bounded function space.
- 1035 • (iii) For all $\mathbf{z}_t, \mathbf{z}'_t \in \mathcal{Z}_t (\mathbf{z}_t \neq \mathbf{z}'_t)$, the set $\{\mathbf{x}_t : P(\mathbf{x}_t | \mathbf{z}_t) \neq P(\mathbf{x}_t | \mathbf{z}'_t)\}$ has positive
 1036 probability.

1037 Suppose that we have learned $(\hat{g}, \hat{f}, P(\hat{\mathbf{z}}_t))$ to achieve Equation (1) and (2), then the latent variables
 1038 $\mathbf{z}_t = \{\mathbf{z}_t^1, \dots, \mathbf{z}_t^L\}$ are block-wise identifiable.

1039 *Proof.* We first follow Hu et al. [24] framework to prove that \mathbf{z}_t is block-wise identifiable given
 1040 sufficient observation. Sequentially, we prove that we require at least $2L + 1$ adjacent observed
 1041 variables to achieve block-wise identifiability.

1042 Given time series data with T timesteps $X = \{\mathbf{x}_1, \dots, \mathbf{x}_t, \dots, \mathbf{x}_T\}$ and L -layer latent variables,
 1043 we let $\mathbf{x}_{< t}$ and $\mathbf{x}_{> t}$ be $\{\mathbf{x}_1, \dots, \mathbf{x}_{t-1}\}$ and $\{\mathbf{x}_{t+1}, \dots, \mathbf{x}_T\}$, respectively. Note that the length of
 1044 $\mathbf{x}_{< t}$ and $\mathbf{x}_{> t}$ are larger than L , i.e., $|\mathbf{x}_{< t}| > L$ and $|\mathbf{x}_{> t}| > L$. Sequentially, according to the data
 1045 generation process in Figure 1, we have:

$$P(\mathbf{x}_{< t} | \mathbf{x}_t, \mathbf{z}_t) = P(\mathbf{x}_{< t} | \mathbf{z}_t), \quad P(\mathbf{x}_{> t} | \mathbf{x}_t, \mathbf{x}_{< t}, \mathbf{z}_t) = P(\mathbf{x}_{> t} | \mathbf{z}_t). \quad (\text{A5})$$

1046 Sequentially, the observed $P(\mathbf{x}_{t-1})$ and joint distribution $P(\mathbf{x}_{> t}, \mathbf{x}_t, \mathbf{x}_{< t})$ directly indicates
 1047 $P(\mathbf{x}_{> t}, \mathbf{x}_t | \mathbf{x}_{< t})$, and we have:

$$\begin{aligned} P(\mathbf{x}_{> t}, \mathbf{x}_t | \mathbf{x}_{< t}) &= \underbrace{\int_{\mathcal{Z}_t} P(\mathbf{x}_{> t}, \mathbf{x}_t, \mathbf{z}_t | \mathbf{x}_{< t}) d\mathbf{z}_t}_{\text{Integration over } \mathcal{Z}_t} = \underbrace{\int_{\mathcal{Z}_t} P(\mathbf{x}_{> t} | \mathbf{x}_t, \mathbf{z}_t, \mathbf{x}_{< t}) P(\mathbf{x}_t, \mathbf{z}_t | \mathbf{x}_{< t}) d\mathbf{z}_t}_{\text{Factorization of joint conditional probability}} \\ &= \underbrace{\int_{\mathcal{Z}_t} P(\mathbf{x}_{> t} | \mathbf{z}_t) P(\mathbf{x}_t, \mathbf{z}_t | \mathbf{x}_{< t}) d\mathbf{z}_t}_{\text{Conditional Independence}} = \underbrace{\int_{\mathcal{Z}_t} P(\mathbf{x}_{> t} | \mathbf{z}_t) P(\mathbf{x}_t | \mathbf{z}_t, \mathbf{x}_{< t}) P(\mathbf{z}_t | \mathbf{x}_{< t}) d\mathbf{z}_t}_{\text{Bayes Law}} \\ &= \int_{\mathcal{Z}_t} P(\mathbf{x}_{> t} | \mathbf{z}_t) P(\mathbf{x}_t | \mathbf{z}_t) P(\mathbf{z}_t | \mathbf{x}_{< t}) d\mathbf{z}_t. \end{aligned} \quad (\text{A6})$$

1048 We further incorporate the integration over $\mathcal{X}_{< t}$ as follows:

$$\int_{\mathcal{X}_{< t}} P(\mathbf{x}_{> t}, \mathbf{x}_t | \mathbf{x}_{< t}) P(\mathbf{x}_{< t}) d\mathbf{x}_{< t} = \int_{\mathcal{X}_{< t}} \int_{\mathcal{Z}_t} P(\mathbf{x}_{> t} | \mathbf{z}_t) P(\mathbf{x}_t | \mathbf{z}_t) P(\mathbf{z}_t | \mathbf{x}_{< t}) P(\mathbf{x}_{< t}) d\mathbf{z}_t d\mathbf{x}_{< t}. \quad (\text{A7})$$

1049 According to the definition of linear operator, we have:

$$\begin{aligned} \int_{\mathcal{X}_{< t}} P(\mathbf{x}_{> t}, \mathbf{x}_t | \mathbf{x}_{< t}) P(\mathbf{x}_{< t}) d\mathbf{x}_{< t} &= [L_{\mathbf{x}_{> t}, \mathbf{x}_t | \mathbf{x}_{< t}} \circ P](\mathbf{x}_{< t}), \\ \int_{\mathcal{X}_{< t}} P(\mathbf{z}_t | \mathbf{x}_{< t}) P(\mathbf{x}_{< t}) d\mathbf{x}_{< t} &= [L_{\mathbf{z}_t | \mathbf{x}_{< t}} \circ P](\mathbf{x}_{< t}) \\ \int_{\mathcal{Z}_t} P(\mathbf{x}_{> t} | \mathbf{z}_t) d\mathbf{z}_t &= L_{\mathbf{x}_{> t} | \mathbf{z}_t}. \end{aligned} \quad (\text{A8})$$

1050 By combining Equation (A7) and (A8), we have:

$$[L_{\mathbf{x}_{> t}, \mathbf{x}_t | \mathbf{x}_{< t}} \circ P](\mathbf{x}_{< t}) = [L_{\mathbf{x}_{> t} | \mathbf{z}_t} D_{\mathbf{x}_t | \mathbf{z}_t} L_{\mathbf{z}_t | \mathbf{x}_{< t}} \circ P](\mathbf{x}_{< t}), \quad (\text{A9})$$

1051 which implies the operator equivalence:

$$L_{\mathbf{x}_{> t}, \mathbf{x}_t | \mathbf{x}_{< t}} = L_{\mathbf{x}_{> t} | \mathbf{z}_t} D_{\mathbf{x}_t | \mathbf{z}_t} L_{\mathbf{z}_t | \mathbf{x}_{< t}}. \quad (\text{A10})$$

1052 Sequentially, we further integrate out \mathbf{x}_t and have:

$$\int_{\mathcal{X}_t} L_{\mathbf{x}_{>t}, \mathbf{x}_t | \mathbf{x}_{<t}} d\mathbf{x}_t = \int_{\mathcal{X}_t} L_{\mathbf{x}_{>t} | \mathbf{z}_t} D_{\mathbf{x}_t | \mathbf{z}_t} L_{\mathbf{z}_t | \mathbf{x}_{<t}} d\mathbf{x}_t, \quad (\text{A11})$$

1053 and it results in:

$$L_{\mathbf{x}_{>t} | \mathbf{x}_{<t}} = L_{\mathbf{x}_{>t} | \mathbf{z}_t} L_{\mathbf{z}_t | \mathbf{x}_{<t}}. \quad (\text{A12})$$

1054 According to assumption (ii), the linear operator $L_{\mathbf{x}_{>t} | \mathbf{z}_t}$ is injective, Equation (A12) can be rewritten
1055 as:

$$L_{\mathbf{x}_{>t} | \mathbf{z}_t}^{-1} L_{\mathbf{x}_{>t} | \mathbf{x}_{<t}} = L_{\mathbf{z}_t | \mathbf{x}_{<t}}. \quad (\text{A13})$$

1056 By combining Equation (A10) and (A13), we have

$$L_{\mathbf{x}_{>t}, \mathbf{x}_t | \mathbf{x}_{<t}} = L_{\mathbf{x}_{>t} | \mathbf{z}_t} D_{\mathbf{x}_t | \mathbf{z}_t} L_{\mathbf{x}_{>t} | \mathbf{z}_t}^{-1} L_{\mathbf{x}_{>t} | \mathbf{x}_{<t}}. \quad (\text{A14})$$

1057 By leveraging Lemma A1, if $L_{\mathbf{x}_{<t} | \mathbf{x}_{>t}}$ is injective, then $L_{\mathbf{x}_{>t} | \mathbf{x}_{<t}}^{-1}$ exists. Therefore, we have:

$$L_{\mathbf{x}_{>t}, \mathbf{x}_t | \mathbf{x}_{<t}} L_{\mathbf{x}_{>t} | \mathbf{x}_{<t}}^{-1} = L_{\mathbf{x}_{>t} | \mathbf{z}_t} D_{\mathbf{x}_t | \mathbf{z}_t} L_{\mathbf{x}_{>t} | \mathbf{z}_t}^{-1}. \quad (\text{A15})$$

1058 Then we can leverage assumption (iii) and the linear operator is bounded. Consequently,
1059 $L_{\mathbf{x}_{>t}, \mathbf{x}_t | \mathbf{x}_{<t}} L_{\mathbf{x}_{>t} | \mathbf{x}_{<t}}^{-1}$ is also bounded, which satisfies the condition of Theorem A1, and hence
1060 the the operator $L_{\mathbf{x}_{>t} | \mathbf{z}_t} D_{\mathbf{x}_t | \mathbf{z}_t} L_{\mathbf{x}_{>t} | \mathbf{z}_t}^{-1}$ have a unique spectral decomposition, where $L_{\mathbf{x}_{>t} | \mathbf{z}_t}$ and
1061 $D_{\mathbf{x}_t | \mathbf{z}_t}$ correspond to eigenfunctions and eigenvalues, respectively.

1062 Since both the marginal and conditional distributions of the observed variables are matched, the true
1063 model and the estimated model yield the same distribution over the observed variables. Therefore,
1064 we also have:

$$L_{\mathbf{x}_{>t}, \mathbf{x}_t | \mathbf{x}_{<t}} L_{\mathbf{x}_{>t} | \mathbf{x}_{<t}}^{-1} = L_{\hat{\mathbf{x}}_{>t}, \hat{\mathbf{x}}_t | \hat{\mathbf{x}}_{<t}} L_{\hat{\mathbf{x}}_{>t} | \hat{\mathbf{x}}_{<t}}^{-1}, \quad (\text{A16})$$

1065 where the L.H.S corresponds to the true model and the R.H.S corresponds to the estimated model.
1066 Moreover, $L_{\hat{\mathbf{x}}_{>t}, \hat{\mathbf{x}}_t | \hat{\mathbf{x}}_{<t}} L_{\hat{\mathbf{x}}_{>t} | \hat{\mathbf{x}}_{<t}}^{-1}$ also have the unique decomposition, so the L.H.S of the Equation
1067 (A16) can be written as:

$$L_{\mathbf{x}_{>t}, \mathbf{x}_t | \mathbf{x}_{<t}} L_{\mathbf{x}_{>t} | \mathbf{x}_{<t}}^{-1} = L_{\hat{\mathbf{x}}_{>t} | \hat{\mathbf{z}}_t} D_{\hat{\mathbf{x}}_t | \hat{\mathbf{z}}_t} L_{\hat{\mathbf{x}}_{>t} | \hat{\mathbf{z}}_t}^{-1}, \quad (\text{A17})$$

1068 Integrating Equation A15 and Equation A17, and noting that their L.H.S. are identical, it follows that
1069 they share the same spectral decomposition. This yields

$$L_{\mathbf{x}_{>t} | \mathbf{z}_t} = C L_{\hat{\mathbf{x}}_{>t} | \hat{\mathbf{z}}_t} P, \quad D_{\mathbf{x}_t | \mathbf{z}_t} = P^{-1} D_{\hat{\mathbf{x}}_t | \hat{\mathbf{z}}_t} P, \quad (\text{A18})$$

1070 where C is a scalar accounting for scaling indeterminacy and P is a permutation on the order of
1071 elements in $D_{\hat{\mathbf{x}}_t | \hat{\mathbf{z}}_t}$, as discussed in [10]. These forms of indeterminacy are analogous to those in
1072 eigendecomposition, which can be viewed as a finite-dimensional special case. P is a mapping from
1073 distribution to distribution

1074 Since the normalizing condition

$$\int_{\hat{\mathcal{X}}_{t+1}} p_{\hat{\mathbf{x}}_t | \hat{\mathbf{z}}_t} d\hat{\mathbf{x}}_t = 1 \quad (\text{A19})$$

1075 must hold for every $\hat{\mathbf{z}}_t$, one only solution is to set $C = 1$.

1076 Hence, $D_{\hat{\mathbf{x}}_t | \hat{\mathbf{z}}_t}$ and $D_{\mathbf{x}_t | \mathbf{z}_t}$ are identical up to a permutation on their repsective elements. We use
1077 unordered sets to express this equivalence:

$$\{p(\mathbf{x}_t | \mathbf{z}_t)\} = \{p(\mathbf{x}_t | \hat{\mathbf{z}}_t)\}, \quad \text{for all } \mathbf{z}_t, \hat{\mathbf{z}}_t. \quad (\text{A20})$$

1078 Due to the set being unordered, the only way to match the R.H.S. with the L.H.S. in a consistent
1079 order is to exchange the conditioning variables, that is,

$$\{p(\mathbf{x}_t | \mathbf{z}_t^{(1)}), p(\mathbf{x}_t | \mathbf{z}_t^{(2)}), \dots\} = \{p(\mathbf{x}_t | \hat{\mathbf{z}}_t^{(\pi(1))}), p(\mathbf{x}_t | \hat{\mathbf{z}}_t^{(\pi(2))}), \dots\}, \quad (\text{A21})$$

1080 where superscript (\cdot) denotes the index of a conditioning variable, and π is reindexing the conditioning
1081 variables. We use a relabeling map h to represent its corresponding value mapping:

$$p(\mathbf{x}_t | \mathbf{z}_t) = p(\mathbf{x}_t | h(\hat{\mathbf{z}}_t)), \quad \text{for all } \mathbf{z}_t, \hat{\mathbf{z}}_t \quad (\text{A22})$$

1082 Since $K_{\hat{\mathbf{z}}_t, \mathbf{z}_t}$, $L_{\mathbf{x}_{>t}|\hat{\mathbf{z}}_t}^{-1}$, and $L_{\mathbf{z}_t|\mathbf{x}_{>t}}$ are continuous, h is continuous and differentiable. Moreover, by
 1083 leveraging Assumption (iii), different values of \mathbf{z} , i.e., $\mathbf{z}_t^{(1)}, \mathbf{z}_t^{(2)}$ imply $p(\mathbf{x}_t|\mathbf{z}_t^{(1)}) \neq p(\mathbf{x}_t|\mathbf{z}_t^{(2)})$. So
 1084 we can construct a function $F: \mathcal{Z} \rightarrow p(\mathbf{x}_t|\mathbf{z}_t)$, and we have:

$$\mathbf{z}_t^{(1)} \neq \mathbf{z}_t^{(2)} \longrightarrow F(\mathbf{z}_t^{(1)}) \neq F(\mathbf{z}_t^{(2)}), \quad (\text{A23})$$

1085 implying that F is injective. Moreover, by using Equation (A22), we have $F(\mathbf{z}_t) = F(h(\hat{\mathbf{z}}_t))$, which
 1086 implies $\mathbf{z}_t = h(\hat{\mathbf{z}}_t)$.

1087 The aforementioned result leverage $\mathbf{x}_{<t}$, \mathbf{x}_t , and $\mathbf{x}_{>t}$ as three different measurement of \mathbf{z}_t , where
 1088 $|\mathbf{x}_{<t}| \gg |\mathbf{z}_t|$, $|\mathbf{x}_{>t}| \gg |\mathbf{z}_t|$ and $|\mathbf{x}_t| < |\mathbf{z}_t|$. It may imply that when the \mathbf{x}_t cannot provide enough
 1089 information to recover \mathbf{z}_t , we can seek more information from $\mathbf{x}_{<t}$ and $\mathbf{x}_{>t}$.

1090 Sequentially, we further prove that when the observed and L -layer latent variables follow the data
 1091 generation process in Equation (1) and (2), we require at least $2L + 1$ adjacent observed variables,
 1092 i.e., $\{\mathbf{x}_{t-L}, \dots, \mathbf{x}_t, \dots, \mathbf{x}_{t+L}\}$ to make \mathbf{z}_t block-wise identifiable. We prove it by contradiction as
 1093 follows.

1094 Suppose we have $2L$ adjacent observations, which can be divided into two cases: 1)
 1095 $\{\mathbf{x}_{t-L+1}, \dots, \mathbf{x}_t, \dots, \mathbf{x}_{t+L}\}$ and $\{\mathbf{x}_{t-L}, \dots, \mathbf{x}_t, \dots, \mathbf{x}_{t+L-1}\}$. In the first case, suppose the di-
 1096 mension of \mathbf{x}_t and that of any layer of latent variables \mathbf{z}_t^l are n , the dimension of $\mathbf{x}_{t-L+1}, \dots, \mathbf{x}_{t-1}$
 1097 is $(L-1) \times n$ and the dimension of \mathbf{z}_t is $L \times n$, conflicting with the assumption that $L_{\mathbf{x}_{t+1}, \dots, \mathbf{x}_{t+L}|\mathbf{z}_t}$
 1098 is injective. In the second case, the dimensions of $\mathbf{x}_{t-L}, \dots, \mathbf{x}_{t-1}$ and $\mathbf{x}_{t+1}, \dots, \mathbf{x}_{t+L-1}$ are $L \times n$
 1099 and $(L-1) \times n$, respectively, conflicting with the assumption that $L_{\mathbf{x}_{t-L}, \dots, \mathbf{x}_{t-1}|\mathbf{x}_{t+1}, \dots, \mathbf{x}_{t+L}}$ is in-
 1100 jective. As a result, we require at least $2L + 1$ adjacent observations, i.e., $\{\mathbf{x}_{t-L}, \dots, \mathbf{x}_t, \dots, \mathbf{x}_{t+L}\}$
 1101 to make \mathbf{z}_t block-wise identifiable. \square

1102 B.3 Examples of injective linear operators

1103 The assumption of the injectivity of a linear operator is commonly employed in the nonparametric
 1104 identification [24, 3, 26]. Intuitively, it means that different input distributions of a linear operator
 1105 correspond to different output distributions of that operator. For a better understanding of this
 1106 assumption, we provide several examples that describe the mapping from $p_{\mathbf{a}} \rightarrow p_{\mathbf{b}}$, where \mathbf{a} and \mathbf{b}
 1107 are random variables.

1108 **Example 1 (Inverse Transformation).** $\mathbf{b} = g(\mathbf{a})$, where g is an invertible function.

1109 **Example 2 (Additive Transformation).** $\mathbf{b} = \mathbf{a} + \epsilon$, where $p(\epsilon)$ must not vanish everywhere after the
 1110 Fourier transform (Theorem 2.1 in [53]).

1111 **Example 3.** $\mathbf{b} = g(\mathbf{a}) + \epsilon$, where the same conditions from Examples 1 and 2 are required.

1112 **Example 4 (Post-linear Transformation).** $\mathbf{b} = g_1(g_2(\mathbf{a}) + \epsilon)$, a post-nonlinear model with invertible
 1113 nonlinear functions g_1, g_2 , combining the assumptions in **Examples 1-3**.

1114 **Example 5 (Nonlinear Transformation with Exponential Family).** $\mathbf{b} = g(\mathbf{a}, \epsilon)$, where the joint
 1115 distribution $p(\mathbf{a}, \mathbf{b})$ follows an exponential family.

1116 **Example 6 (General Nonlinear Transformation).** $\mathbf{b} = g(\mathbf{a}, \epsilon)$, a general nonlinear formulation.
 1117 Certain deviations from the nonlinear additive model (**Example 3**), e.g., polynomial perturbations,
 1118 can still be tractable.

1119 B.4 Monotonicity and Normalization Assumption

1120 **Assumption 1** (Monotonicity and Normalization Assumption [26]). For any $\mathbf{x}_t \in \mathcal{X}_t$, there exists
 1121 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 normalize $\mathbf{z}_t =$
 1122 $G[p_{\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{z}_t}(\cdot|\mathbf{x}_t, \mathbf{z}_t)]$.

1123 B.5 Block-wise Identifiability of Latent Variables \mathbf{z}_t^l in any l -th Layer

1124 **Theorem A3. (Block-wise Identifiability of Latent Variables \mathbf{z}_t^l in any l -th Layer.)** For a series of
 1125 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
 1126 $\hat{f}_i, \hat{P}(\hat{\epsilon}), \hat{P}(\hat{\eta}), \hat{\mathbf{g}}$, suppose that the process subject to observational equivalence $\mathbf{x}_t = \hat{\mathbf{g}}(\hat{\mathbf{z}}_t, \hat{\eta}_t)$. We

1127 let $\mathbf{c}_t \triangleq \{\mathbf{z}_{t-1}, \mathbf{z}_t\} \in \mathbb{R}^{2 \times L \times n}$ and that $\mathcal{M}_{\mathbf{c}_t}$ be the variable set of two consecutive timestamps
 1128 and the corresponding node-vector Markov network respectively, and further employ following
 1129 assumptions:

- 1130 • (i) (Smooth and Positive Density): The conditional probability function of the latent variables
 1131 \mathbf{c}_t is smooth and positive, i.e., $p(\mathbf{c}_t|\mathbf{z}_{t-2})$ is third-order differentiable and $p(\mathbf{c}_t|\mathbf{z}_{t-2}) > 0$
 1132 over $\mathbb{R}^{2 \times L \times n}$.
- 1133 • (ii) (Sufficient Variability): Denote $|\mathcal{M}_{\mathbf{c}_t}|$ as the number of edges in Markov network $\mathcal{M}_{\mathbf{c}_t}$.
 1134 Let

$$w(m) = \left(\frac{\partial^3 \log p(\mathbf{c}_t|\mathbf{z}_{t-2})}{\partial c_{t,1}^2 \partial z_{t-2,m}}, \dots, \frac{\partial^3 \log p(\mathbf{c}_t|\mathbf{z}_{t-2})}{\partial c_{t,2n}^2 \partial z_{t-2,m}} \right) \oplus \left(\frac{\partial^2 \log p(\mathbf{c}_t|\mathbf{z}_{t-2})}{\partial c_{t,1} \partial z_{t-2,m}}, \dots, \frac{\partial^2 \log p(\mathbf{c}_t|\mathbf{z}_{t-2})}{\partial c_{t,2n} \partial z_{t-2,m}} \right) \oplus \left(\frac{\partial^3 \log p(\mathbf{c}_t|\mathbf{z}_{t-2})}{\partial c_{t,i} \partial c_{t,j} \partial z_{t-2,m}} \right)_{(i,j) \in \mathcal{E}(\mathcal{M}_{\mathbf{c}_t})}, \quad (\text{A24})$$

1135 where \oplus denotes concatenation operation and $(i, j) \in \mathcal{E}(\mathcal{M}_{\mathbf{c}_t})$ denotes all pairwise indice
 1136 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
 1137 values of $\mathbf{z}_{t-2,m}$, such that the $4n + |\mathcal{M}_{\mathbf{c}_t}|$ values of vector functions $w(m)$ are linearly
 1138 independent.

1139 Then for latent variables \mathbf{z}_t^l at the l -th layer, \mathbf{z}_t^l is block-wise identifiable, i.e., there exists $\hat{\mathbf{z}}_t^l$ and an
 1140 invertible function $h_t^l: \mathbb{R}^n \rightarrow \mathbb{R}^n$, such that $\mathbf{z}_t^l = h_t^l(\hat{\mathbf{z}}_t^l)$.

1141 *Proof.* By reusing Theorem A2 with more observed variables, $\{\mathbf{z}_{t-2}, \mathbf{z}_{t-1}, \mathbf{z}_t\}$ and $\{\mathbf{z}_{t-1}, \mathbf{z}_t\}$ are
 1142 also block-wise identifiable, implying that there exists invertible functions h_3 and h_2 , such that
 1143 $\hat{\mathbf{z}}_{t-2}, \hat{\mathbf{z}}_{t-1}, \hat{\mathbf{z}}_t = h_3(\mathbf{z}_{t-2}, \mathbf{z}_{t-1}, \mathbf{z}_t)$ and $\hat{\mathbf{z}}_{t-1}, \hat{\mathbf{z}}_t = h_2(\mathbf{z}_{t-1}, \mathbf{z}_t)$. So we have:

$$\begin{aligned} P(\hat{\mathbf{z}}_{t-2}, \hat{\mathbf{z}}_{t-1}, \hat{\mathbf{z}}_t) &= P(\hat{\mathbf{z}}_{t-2}, \mathbf{z}_{t-1}, \mathbf{z}_t) |\mathbf{J}_{h_2}| \iff P(\hat{\mathbf{z}}_{t-1}, \hat{\mathbf{z}}_t | \hat{\mathbf{z}}_{t-2}) = P(\mathbf{z}_{t-1}, \mathbf{z}_t | \hat{\mathbf{z}}_{t-2}) |\mathbf{J}_{h_2}| \\ &\iff P(\hat{\mathbf{c}}_t | \hat{\mathbf{z}}_{t-2}) = P(\mathbf{c}_t | \hat{\mathbf{z}}_{t-2}) |\mathbf{J}_{h_2}| \iff \ln P(\hat{\mathbf{c}}_t | \hat{\mathbf{z}}_{t-2}) = \ln P(\mathbf{c}_t | \hat{\mathbf{z}}_{t-2}) + \ln |\mathbf{J}_{h_2}|, \end{aligned} \quad (\text{A25})$$

1144 And then we partition the latent variables \mathbf{z}_t into two parts \mathbf{z}_t

1145 Let $\hat{\mathbf{z}}_{t-1,k}^{l_1}$ and $\hat{\mathbf{z}}_{t,o}^{l_2}$ be two variables that denote the k -th variable of $\mathbf{z}_{t-1}^{l_1}$ and o -th variable of $\mathbf{z}_t^{l_2}$,
 1146 respectively, where $l_1 > l_2$. According to the data generation process, it is not hard to find that $\hat{\mathbf{z}}_{t-1,k}^{l_1}$
 1147 and $\hat{\mathbf{z}}_{t,o}^{l_2}$ are not adjacent in the estimated Markov network $\mathcal{M}_{\hat{\mathbf{c}}_t}$ over $\hat{\mathbf{c}}_t = \{\hat{\mathbf{z}}_{t-1}, \hat{\mathbf{z}}_t\}$. We conduct
 1148 the first-order derivative w.r.t. $\hat{\mathbf{z}}_{t-1,k}^{l_1}$ and have

$$\frac{\partial \log p(\hat{\mathbf{c}}_t | \hat{\mathbf{z}}_{t-2})}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1}} = \sum_{l=1}^{2L} \sum_{i=1}^n \frac{\partial \log p(\mathbf{c}_t | \hat{\mathbf{z}}_{t-2})}{\partial c_{t,i}^l} \cdot \frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1}} + \frac{\partial \log |\mathbf{J}_{h_2}|}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1}}. \quad (\text{A26})$$

1149 We further conduct the second-order derivative w.r.t. $\hat{\mathbf{z}}_{t-1,k}^{l_1}$ and $\hat{\mathbf{z}}_{t,o}^{l_2}$, then we have:

$$\begin{aligned} \frac{\partial^2 \log p(\hat{\mathbf{c}}_t | \hat{\mathbf{z}}_{t-2})}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1} \partial \hat{\mathbf{z}}_{t,o}^{l_2}} &= \sum_{l=1}^{2L} \sum_{i=1}^n \sum_{s=1}^{2L} \sum_{j=1}^n \frac{\partial^2 \log p(\mathbf{c}_t | \hat{\mathbf{z}}_{t-2})}{\partial c_{t,i}^l \partial c_{t,j}^s} \cdot \frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1}} \cdot \frac{\partial c_{t,j}^s}{\partial \hat{\mathbf{z}}_{t,o}^{l_2}} \\ &\quad + \sum_{l=1}^{2L} \sum_{i=1}^n \frac{\partial \log p(\mathbf{c}_t | \hat{\mathbf{z}}_{t-2})}{\partial c_{t,i}^l} \cdot \frac{\partial^2 c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1} \partial \hat{\mathbf{z}}_{t,o}^{l_2}} + \frac{\partial^2 \log |\mathbf{J}_{h_2}|}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1} \partial \hat{\mathbf{z}}_{t,o}^{l_2}}. \end{aligned} \quad (\text{A27})$$

1150 Since $\hat{\mathbf{z}}_{t-1,k}^{l_1}$ and $\hat{\mathbf{z}}_{t,o}^{l_2}$ are not adjacent in $\mathcal{M}_{\hat{\mathbf{c}}_t}$, $\hat{\mathbf{z}}_{t-1,k}^{l_1}$ and $\hat{\mathbf{z}}_{t,o}^{l_2}$ are conditionally independent given
 1151 $\hat{\mathbf{c}}_t \setminus \{\hat{c}_{t,k}, \hat{c}_{t,l}\}$. Utilizing the fact that conditional independence can lead to zero cross derivative [51],
 1152 for each value of $\hat{\mathbf{z}}_{t-2}$, we have:

$$\begin{aligned} \frac{\partial^2 \log p(\hat{\mathbf{c}}_t | \hat{\mathbf{z}}_{t-2})}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1} \partial \hat{\mathbf{z}}_{t,o}^{l_2}} &= \frac{\partial^2 \log p(\hat{c}_{t,k} | \hat{\mathbf{c}}_t \setminus \{\hat{\mathbf{z}}_{t-1,k}^{l_1}, \hat{\mathbf{z}}_{t,o}^{l_2}\}, \hat{\mathbf{z}}_{t-2})}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1} \partial \hat{\mathbf{z}}_{t,o}^{l_2}} + \frac{\partial^2 \log p(\hat{c}_{t,l} | \mathbf{c}_t \setminus \{\hat{\mathbf{z}}_{t-1,k}^{l_1}, \hat{\mathbf{z}}_{t,o}^{l_2}\}, \hat{\mathbf{z}}_{t-2})}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1} \partial \hat{\mathbf{z}}_{t,o}^{l_2}} \\ &\quad + \frac{\partial^2 \log p(\hat{\mathbf{c}}_t \setminus \{\hat{\mathbf{z}}_{t-1,k}^{l_1}, \hat{\mathbf{z}}_{t,o}^{l_2}\} | \hat{\mathbf{z}}_{t-2})}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1} \partial \hat{\mathbf{z}}_{t,o}^{l_2}} = 0. \end{aligned} \quad (\text{A28})$$

1153 Bring in Equation (A28), Equation (A27) can be further derived as:

$$\begin{aligned}
0 = & \underbrace{\sum_{l=1}^{2L} \sum_{i=1}^n \frac{\partial^2 \log p(\mathbf{c}_t | \hat{\mathbf{z}}_{t-2})}{\partial (c_{t,i}^l)^2} \cdot \frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1}} \cdot \frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,o}^{l_2}}}_{\text{(i) } i=j} \\
& + \underbrace{\sum_{l=1}^{2L} \sum_{i=1}^n \sum_{s:(s,l) \in \mathcal{E}(\mathcal{M}_{\mathbf{c}_t})} \sum_{j=1}^n \frac{\partial^2 \log p(\mathbf{c}_t | \hat{\mathbf{z}}_{t-2})}{\partial c_{t,i}^l \partial c_{t,j}^s} \cdot \frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1}} \cdot \frac{\partial c_{t,j}^s}{\partial \hat{\mathbf{z}}_{t,o}^{l_2}}}_{\text{(ii) } c_{t,i} \text{ and } c_{t,j} \text{ are adjacent in } \mathcal{M}_{\mathbf{c}_t}} \\
& + \underbrace{\sum_{l=1}^{2L} \sum_{i=1}^n \sum_{s:(s,l) \notin \mathcal{E}(\mathcal{M}_{\mathbf{c}_t})} \sum_{j=1}^n \frac{\partial^2 \log p(\mathbf{c}_t | \hat{\mathbf{z}}_{t-2})}{\partial c_{t,i}^l \partial c_{t,j}^s} \cdot \frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1}} \cdot \frac{\partial c_{t,j}^s}{\partial \hat{\mathbf{z}}_{t,o}^{l_2}}}_{\text{(iii) } c_{t,i} \text{ and } c_{t,j} \text{ are not adjacent in } \mathcal{M}_{\mathbf{c}_t}} \\
& + \sum_{l=1}^{2L} \sum_{i=1}^n \frac{\partial \log p(\mathbf{c}_t | \hat{\mathbf{z}}_{t-2})}{\partial c_{t,i}^l} \cdot \frac{\partial^2 c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1} \partial \hat{\mathbf{z}}_{t,o}^{l_2}} + \frac{\partial \log |\mathbf{J}_{h_{\mathbf{c},t}}|}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1} \partial \hat{\mathbf{z}}_{t,o}^{l_2}},
\end{aligned} \tag{A29}$$

1154 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 (A28), we
1155 have $\frac{\partial^2 p(\mathbf{c}_t | \hat{\mathbf{z}}_{t-2})}{\partial c_{t,i} \partial c_{t,j}} = 0$ when $c_{t,i}$, $c_{t,j}$ are not adjacent in $\mathcal{M}_{\mathbf{c}_t}$. Thus, Equation (A29) can be rewritten
1156 as:

$$\begin{aligned}
0 = & \sum_{l=1}^{2L} \sum_{i=1}^n \frac{\partial^2 \log p(\mathbf{c}_t | \hat{\mathbf{z}}_{t-2})}{\partial (c_{t,i}^l)^2} \cdot \frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1}} \cdot \frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,o}^{l_2}} + \sum_{l=1}^{2L} \sum_{i=1}^n \sum_{s:(s,l) \in \mathcal{E}(\mathcal{M}_{\mathbf{c}_t})} \sum_{j=1}^n \frac{\partial^2 \log p(\mathbf{c}_t | \hat{\mathbf{z}}_{t-2})}{\partial c_{t,i}^l \partial c_{t,j}^s} \cdot \frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1}} \cdot \frac{\partial c_{t,j}^s}{\partial \hat{\mathbf{z}}_{t,o}^{l_2}} \\
& + \sum_{l=1}^{2L} \sum_{i=1}^n \frac{\partial \log p(\mathbf{c}_t | \hat{\mathbf{z}}_{t-2})}{\partial c_{t,i}^l} \cdot \frac{\partial^2 c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1} \partial \hat{\mathbf{z}}_{t,o}^{l_2}} + \frac{\partial \log |\mathbf{J}_{h_{\mathbf{c},t}}|}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1} \partial \hat{\mathbf{z}}_{t,o}^{l_2}},
\end{aligned} \tag{A30}$$

1157 Then for each $m = 1, 2, \dots, nL$ and each value of $\hat{\mathbf{z}}_{t-2,m}$, we conduct partial derivative on both
1158 sides of Equation (A30) and have:

$$\begin{aligned}
0 = & \sum_{l=1}^{2L} \sum_{i=1}^n \frac{\partial^3 \log p(\mathbf{c}_t | \hat{\mathbf{z}}_{t-2})}{\partial (c_{t,i}^l)^2 \partial \hat{\mathbf{z}}_{t-2,m}} \cdot \frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1}} \cdot \frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,o}^{l_2}} + \\
& \sum_{l=1}^{2L} \sum_{i=1}^n \sum_{s:(s,l) \in \mathcal{E}(\mathcal{M}_{\mathbf{c}_t})} \sum_{j=1}^n \frac{\partial^3 \log p(\mathbf{c}_t | \hat{\mathbf{z}}_{t-2})}{\partial c_{t,i}^l \partial c_{t,j}^s \partial \hat{\mathbf{z}}_{t-2,m}} \cdot \frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1}} \cdot \frac{\partial c_{t,j}^s}{\partial \hat{\mathbf{z}}_{t,o}^{l_2}} \\
& + \sum_{l=1}^{2L} \sum_{i=1}^n \frac{\partial^2 \log p(\mathbf{c}_t | \hat{\mathbf{z}}_{t-2})}{\partial c_{t,i}^l \partial \hat{\mathbf{z}}_{t-2,m}} \cdot \frac{\partial^2 c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1} \partial \hat{\mathbf{z}}_{t,o}^{l_2}},
\end{aligned} \tag{A31}$$

1159 Finally, we have

$$\begin{aligned}
0 = & \sum_{l=1}^{2L} \sum_{i=1}^n \frac{\partial^3 \log p(\mathbf{c}_t | \hat{\mathbf{z}}_{t-2})}{\partial (c_{t,i}^l)^2 \partial \hat{\mathbf{z}}_{t-2,m}} \cdot \frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1}} \cdot \frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,o}^{l_2}} + \sum_{l=1}^{2L} \sum_{i=1}^n \frac{\partial^2 \log p(\mathbf{c}_t | \hat{\mathbf{z}}_{t-2})}{\partial c_{t,i}^l \partial \hat{\mathbf{z}}_{t-2,m}} \cdot \frac{\partial^2 c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1} \partial \hat{\mathbf{z}}_{t,o}^{l_2}} \\
& + \sum_{l,s:(l,s) \in \mathcal{E}(\mathcal{M}_{\mathbf{c}_t})} \sum_{i=1}^n \sum_{j=1}^n \frac{\partial^3 \log p(\mathbf{c}_t | \hat{\mathbf{z}}_{t-2})}{\partial c_{t,i}^l \partial c_{t,j}^s \partial \hat{\mathbf{z}}_{t-2,m}} \cdot \left(\frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1}} \cdot \frac{\partial c_{t,j}^s}{\partial \hat{\mathbf{z}}_{t,o}^{l_2}} + \frac{\partial c_{t,j}^s}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1}} \cdot \frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,o}^{l_2}} \right).
\end{aligned} \tag{A32}$$

1160 According to Assumption A2, we can construct $4nL + |\mathcal{M}_{\mathbf{c}}| \times n^2$ different equations with different
1161 values of $\hat{\mathbf{z}}_{t-2,m}$, and the coefficients of the equation system they form are linearly independent. To
1162 ensure that the right-hand side of the equations is always 0, the only solution is

$$\frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1}} \cdot \frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,o}^{l_2}} = 0, \tag{A33}$$

1163

$$\frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1}} \cdot \frac{\partial c_{t,j}^s}{\partial \hat{\mathbf{z}}_{t,o}^{l_2}} + \frac{\partial c_{t,j}^s}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1}} \cdot \frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,o}^{l_2}} = 0, \quad (\text{A34})$$

1164

$$\frac{\partial (c_{t,i}^l)^2}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1} \partial \hat{\mathbf{z}}_{t,o}^{l_2}} = 0. \quad (\text{A35})$$

1165 Bringing Eq A33 into Eq A34, at least one product must be zero, and the other must be zero as well.
1166 That is,

$$\frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1}} \cdot \frac{\partial c_{t,j}^s}{\partial \hat{\mathbf{z}}_{t,o}^{l_2}} = 0. \quad (\text{A36})$$

1167 According to the aforementioned results, $\hat{\mathbf{z}}_{t-1,k}^{l_1}$ and $\hat{\mathbf{z}}_{t,o}^{l_2}$ be two variables that denote the k -th variable
1168 of $\mathbf{z}_{t-1}^{l_1}$ and o -th variable of $\mathbf{z}_t^{l_2}$, respectively, where $l_1 > l_2$, we draw the following conclusions.

1169 (i) Equation (A33) implies that, each ground-truth latent variable $c_{t,i}^l$ in l -th block is a function of at
1170 most one of $\hat{c}_{t,k}^{l_1}$ and $\hat{c}_{t,l}^{l_2}$, which are in l_1 -th and l_2 -th layer, respectively

1171 (ii) Equation (A36) implies that, for each pair of ground-truth latent variables $c_{t,i}^l$ and $c_{t,j}^s$ in l -th and
1172 l -th blocks, respectively, that are **adjacent** in $\mathcal{M}_{\mathbf{c}_t}$ over \mathbf{c}_t , they can not be a function of $\hat{c}_{t,k}^{l_1}$ and $\hat{c}_{t,l}^{l_2}$
1173 respectively.

1174 According to the data generation process, we can restrict the independent noises among blocks, and
1175 hence the estimated node-vector Markov network is isomorphic to the ground-truth node-vector
1176 Markov network.

1177 Sequentially, we further give the proof that under the same permutation \mathbf{z}_t^π , block \mathbf{z}_t^i is only a
1178 function of $\mathbf{z}_t^{\pi(i)}$. Since the permutation happens on each timestamp respectively, the cross-timestamp
1179 block-wise identifiability is presented clearly.

1180 Suppose there exists a pair of indices $l, s \in \{1, \dots, L\}$. Since h_2 is invertible, there exists a permuted
1181 version of the estimated blocks, denoted as \mathbf{c}_t^π , such that:

$$\frac{\partial \mathbf{z}_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,j}^{\pi(l)}} \neq 0, \quad l = 1, \dots, L, \text{ and } i, j = 1, \dots, n, \quad (\text{A37})$$

1182 we have $\frac{\partial \mathbf{z}_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,j}^{\pi(l)}} \neq 0$ and $\frac{\partial \mathbf{z}_{t,i}^s}{\partial \hat{\mathbf{z}}_{t,j}^{\pi(s)}} \neq 0$. Let us discuss it case by case. We can discuss it in the following
1183 case.

1184 • If \mathbf{z}_t^l is not adjacent to \mathbf{z}_t^s , we have $\hat{\mathbf{z}}_t^{\pi(l)}$ is not adjacent to $\hat{\mathbf{z}}_t^{\pi(s)}$. Using Equation (A33), we
1185 have $\frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1}} \cdot \frac{\partial c_{t,i}^s}{\partial \hat{\mathbf{z}}_{t,o}^{l_2}} = 0$, which leads to $\frac{\partial \mathbf{z}_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,j}^{\pi(j)}} = 0$.

1186 • If block \mathbf{z}_t^l is adjacent to block \mathbf{z}_t^s , we have $\hat{\mathbf{z}}_t^{\pi(i)}$ is adjacent to $\hat{\mathbf{z}}_t^{\pi(j)}$. The intimate neighbor
1187 set of \mathbf{z}_t^i is empty, there exists at least one index k such that \mathbf{z}_t^k is adjacent to \mathbf{z}_t^i but not
1188 adjacent to \mathbf{z}_t^j . Similarly, we have the same structure on the estimated Markov network,
1189 which means that $\hat{\mathbf{z}}_t^{\pi(k)}$ is adjacent to $\hat{\mathbf{z}}_t^{\pi(i)}$ but not adjacent to $\hat{\mathbf{z}}_t^{\pi(j)}$. Using Equation (A36)
1190 we have $\frac{\partial c_{t,i}^l}{\partial \hat{\mathbf{z}}_{t-1,k}^{l_1}} \cdot \frac{\partial c_{t,j}^s}{\partial \hat{\mathbf{z}}_{t,o}^{l_2}} = 0$, which leads to $\frac{\partial \mathbf{z}_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,j}^{\pi(s)}} = 0$.

1191 In conclusion, we always have $\frac{\partial \mathbf{z}_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,j}^{\pi(s)}} = 0$, meaning that there exists a permutation π of the estimated
1192 blocks, such that \mathbf{z}_t^l and $\mathbf{z}_t^{\pi(l)}$ is one-to-one corresponding.

1193 Sequentially, we further leverage Lemma A2 to show that there is exist an invertible function h^l , such
1194 that $\hat{\mathbf{z}}_t^l = h^l(\mathbf{z}_t^l)$. \square

1195 **Lemma A2.** (Hierarchical Structure Resolves Layer-Wise Permutation Indeterminacy) If the
1196 estimated latent causal graph represents a hierarchical structure, the layer-wise components
1197 $\mathbf{z}_t^1, \mathbf{z}_t^2, \dots, \mathbf{z}_t^L$ are block-wise identifiable without permutation.

1198 *Proof.* Let \mathbf{A}_t represent the true causal adjacency matrix of the latent causal graph at time t in
 1199 layer-wise level, and let $\hat{\mathbf{A}}_t$ represent its estimation. By definition, we have

$$\mathbf{A}_{t,k,l} = \begin{cases} \frac{\partial \mathbf{z}_t^l}{\partial \mathbf{z}_t^k}, & k = l + 1, \\ 0, & \text{otherwise,} \end{cases} \quad \hat{\mathbf{A}}_{t,k,l} = \begin{cases} \frac{\partial \hat{\mathbf{z}}_t^l}{\partial \hat{\mathbf{z}}_t^k}, & k = l + 1, \\ 0, & \text{otherwise,} \end{cases} \quad k, l = 1, \dots, L. \quad (\text{A38})$$

1200 Consider the nonzero elements of the adjacency matrices, which occur at positions where $k = l + 1$.
 1201 If a layer-wise permutation is applied, the rows and columns of the \mathbf{A}_t are both permuted according
 1202 to the same permutation. Specifically, we have

$$\hat{\mathbf{A}}_{l+1,l} = \mathbf{D}_{l+1,1} \mathbf{A}_{\pi(l+1),\pi(l)}, \quad (\text{A39})$$

1203 where $\mathbf{D}_{l+1,1}$ is a diagonal matrix representing the scaling indeterminacy, and $\pi : \{1, \dots, L\} \rightarrow$
 1204 $\{1, \dots, L\}$ is a permutation function. The subscripts of $\hat{\mathbf{A}}$ and \mathbf{A} above indicate the following
 1205 equation:

$$\pi(l+1) = \pi(l) + 1. \quad (\text{A40})$$

1206 The recursive formula in Equation (A40) implies:

$$\pi(l) = \pi(1) + (l - 1). \quad (\text{A41})$$

1207 For $\pi(l)$ to be a valid permutation covering all values in $\{1, \dots, L\}$, it is necessary that $\pi(1) = 1$. If
 1208 $\pi(1) \neq 1$, the sequence $\pi(l) = \pi(1) + (l - 1)$ would either exceed L or fail to include 1, violating
 1209 the bijective property of π . Hence, we conclude that

$$\pi(l) = l, \quad (\text{A42})$$

1210 indicating that the layer-wise components remain unpermuted. \square

1211 B.6 Extension to Multiple Lags

1212 For the sake of simplicity, we consider only one special case with $\tau = 1$ in Theorem 2. Our
 1213 theoretical results can actually be extended to arbitrary lags and subsequences easily. For any given
 1214 time lag τ , and future horizons which is centered at \mathbf{z}_t with historical τ_h and future τ_f steps, i.e.,
 1215 $\mathbf{c}_t = \{\mathbf{z}_{t-\tau_h}, \dots, \mathbf{z}_t, \dots, \mathbf{z}_{t+\tau_f}\} \in \mathbb{R}^{(\tau_h+\tau_f+n) \times n}$. In this case, the vector function $w(m)$ in the
 1216 Sufficient Variability assumption should be modified as

$$\begin{aligned} w(m) = & \left(\frac{\partial^3 \log p(\mathbf{c}_t | \mathbf{z}_{t-\tau_h-1}, \dots, \mathbf{z}_{t-\tau_h-\tau})}{\partial c_{t,1}^2 \partial z_{t-\tau_h-1,m}}, \dots, \frac{\partial^3 \log p(\mathbf{c}_t | \mathbf{z}_{t-\tau_h-1}, \dots, \mathbf{z}_{t-\tau_h-\tau})}{\partial c_{t,2n}^2 \partial z_{t-\tau_h-1,m}} \right) \oplus \\ & \left(\frac{\partial^2 \log p(\mathbf{c}_t | \mathbf{z}_{t-\tau_h-1}, \dots, \mathbf{z}_{t-\tau_h-\tau})}{\partial c_{t,1} \partial z_{t-\tau_h-1,m}}, \dots, \frac{\partial^2 \log p(\mathbf{c}_t | \mathbf{z}_{t-\tau_h-1}, \dots, \mathbf{z}_{t-\tau_h-\tau})}{\partial c_{t,2n} \partial z_{t-\tau_h-1,m}} \right) \oplus \\ & \left(\frac{\partial^3 \log p(\mathbf{c}_t | \mathbf{z}_{t-\tau_h-1}, \dots, \mathbf{z}_{t-\tau_h-\tau})}{\partial c_{t,i} \partial c_{t,j} \partial z_{t-\tau_h-1,m}} \right)_{(i,j) \in \mathcal{E}(\mathcal{M}_{\mathbf{c}_t})}. \end{aligned} \quad (\text{A43})$$

1217 Besides, $2 \times n \times (\tau_h + \tau_f + 1) + |\mathcal{M}_{\mathbf{c}_t}|$ values of linearly independent vector functions in $z_{t',m}^l$ for
 1218 $t' \in \{t - \tau_h - 1, \dots, t - \tau_h - \tau\}$ and $m \in \{1, \dots, n\}$ are required as well. The rest of the theorem
 1219 remains the same, and the proof can be easily extended in such a setting.

1220 B.7 Component-wise Identifiability of Latent Variables $\mathbf{z}_{t,i}^l$ in any l -th Layer

1221 **Lemma A3. (Component-wise Identifiability of Latent Variables $\mathbf{z}_{t,i}^l$ in any l -th Layer)** For a series
 1222 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
 1223 $\hat{f}_i, \hat{P}(\hat{\epsilon}), \hat{P}(\hat{\eta}), \hat{\mathbf{g}}$, suppose that the process subject to observational equivalence $\mathbf{x}_t = \hat{\mathbf{g}}(\hat{\mathbf{z}}_t, \hat{\eta}_t)$. We
 1224 employ sufficient variability assumptions as follows:

- 1225 • (i) (sufficient variability) For any $\mathbf{z}_t^l \in \mathcal{Z}_t^l \subseteq \mathbb{R}^n$ and $\hat{\mathbf{u}} = \{\hat{\mathbf{z}}_{t-1}^l, \hat{\mathbf{z}}_t^{l+1}\}$ there exist $2n + 1$
 1226 different values of $\hat{\mathbf{u}}$, $m = 0, \dots, 2n$, such that these $2n$ vectors $\mathbf{v}_{l,m} - \mathbf{v}_{l,0}$ are linearly
 1227 independent, where $\mathbf{v}_{l,m}$ is defined as:

$$\mathbf{v}_{l,m} = \left(\frac{\partial^2 \ln P(\mathbf{z}_{t,1}^l | \hat{\mathbf{u}})}{(\partial \mathbf{z}_{t,1}^l)^2}, \dots, \frac{\partial^2 \ln P(\mathbf{z}_{t,n}^l | \hat{\mathbf{u}})}{(\partial \mathbf{z}_{t,n}^l)^2}, \frac{\partial \ln P(\mathbf{z}_{t,1}^l | \hat{\mathbf{u}})}{\partial \mathbf{z}_{t,1}^l}, \dots, \frac{\partial \ln P(\mathbf{z}_{t,n}^l | \hat{\mathbf{u}})}{\partial \mathbf{z}_{t,n}^l} \right). \quad (\text{A44})$$

1228 Then for i -th latent variable $\mathbf{z}_{t,i}^l$ at the l -th layer, $\mathbf{z}_{t,i}^l$ is component-wise identifiability, i.e., there
 1229 exists $\hat{\mathbf{z}}_{t,j}^l$ and an invertible function $h_t^l : \mathbb{R} \rightarrow \mathbb{R}$, such that $\hat{\mathbf{z}}_{t,j}^l = h_t^l(\mathbf{z}_{t,i}^l)$.

1230 *Proof.* According to Theorem A3, we have achieved the block-identifiability of \mathbf{z}_t^l , by letting \mathbf{z}_{t-1}^l
 1231 and \mathbf{z}_t^{l-1} be the temporal and hierarchical parents, we further have :

$$\begin{aligned} P(\hat{\mathbf{z}}_t^l, \hat{\mathbf{z}}_{t-1}^l, \hat{\mathbf{z}}_t^{l-1}) &= P(h_t^l(\mathbf{z}_t^l), \hat{\mathbf{z}}_{t-1}^l, \hat{\mathbf{z}}_t^{l-1}) \iff P(\hat{\mathbf{z}}_t^l | \hat{\mathbf{z}}_{t-1}^l, \hat{\mathbf{z}}_t^{l-1}) = P(h_t^l(\mathbf{z}_t^l) | \hat{\mathbf{z}}_{t-1}^l, \hat{\mathbf{z}}_t^{l-1}) \\ &\iff \ln P(\hat{\mathbf{z}}_t^l | \hat{\mathbf{u}}) = \ln P(\mathbf{z}_t^l | \hat{\mathbf{u}}) + \ln |\mathbf{J}_{h_t^l}| \iff \sum_{i=1}^n \ln P(\hat{\mathbf{z}}_{t,i}^l | \hat{\mathbf{u}}) = \sum_{i=1}^n \ln P(\mathbf{z}_{t,i}^l | \hat{\mathbf{u}}) + \ln |\mathbf{J}_{h_t^l}|, \end{aligned} \quad (\text{A45})$$

1232 where $\mathbf{J}_{h_t^l}$ is the Jacobian matrix of the transformation associated with h_t^l . Sequentially, we dif-
 1233 ferentiate both sides of the Equation (A45) w.r.t $\hat{\mathbf{z}}_{t,j}^l$ and $\hat{\mathbf{z}}_{t,k}^l$, where $j, k \in \{1, \dots, n\}$ and $j \neq k$
 1234 yields

$$0 = \sum_{i=1}^n \left(\frac{\partial^2 \ln P(\mathbf{z}_{t,i}^l | \hat{\mathbf{u}})}{(\partial \mathbf{z}_{t,i}^l)^2} \cdot \frac{\partial \mathbf{z}_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,j}^l} \cdot \frac{\partial \mathbf{z}_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,k}^l} + \frac{\partial \ln P(\mathbf{z}_{t,i}^l | \hat{\mathbf{u}})}{\partial \mathbf{z}_{t,i}^l} \cdot \frac{\partial^2 \mathbf{z}_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,j}^l \partial \hat{\mathbf{z}}_{t,k}^l} \right) + \frac{\partial |\mathbf{J}_{h_t^l}|}{\partial \hat{\mathbf{z}}_{t,j}^l \partial \hat{\mathbf{z}}_{t,k}^l}. \quad (\text{A46})$$

1235 Therefore, for $\mathbf{u} = \mathbf{u}_0, \dots, \mathbf{u}_{2n}$, we have $2n + 1$ such equations. Subtracting each equation
 1236 corresponding to $\mathbf{u}_1, \dots, \mathbf{u}_{2n}$ with the equation corresponding to \mathbf{u}_0 results in $2n$ equations:

$$\begin{aligned} 0 &= \sum_{i=1}^n \left(\left(\frac{\partial^2 \ln P(\mathbf{z}_{t,i}^l | \hat{\mathbf{u}}_m)}{(\partial \mathbf{z}_{t,i}^l)^2} - \frac{\partial^2 \ln P(\mathbf{z}_{t,i}^l | \hat{\mathbf{u}}_0)}{(\partial \mathbf{z}_{t,i}^l)^2} \right) \cdot \frac{\partial \mathbf{z}_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,j}^l} \cdot \frac{\partial \mathbf{z}_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,k}^l} \right. \\ &\quad \left. + \left(\frac{\partial \ln P(\mathbf{z}_{t,i}^l | \hat{\mathbf{u}}_m)}{\partial \mathbf{z}_{t,i}^l} - \frac{\partial \ln P(\mathbf{z}_{t,i}^l | \hat{\mathbf{u}}_0)}{\partial \mathbf{z}_{t,i}^l} \right) \cdot \frac{\partial^2 \mathbf{z}_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,j}^l \partial \hat{\mathbf{z}}_{t,k}^l} \right), \end{aligned} \quad (\text{A47})$$

1237 where $m = 1, \dots, 2n$. As a result, under the sufficient variability assumption, the linear system is a
 1238 $2n \times 2n$ full-rank system, and the only solution is $\frac{\partial \mathbf{z}_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,j}^l} \cdot \frac{\partial \mathbf{z}_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,k}^l} = 0$ and $\frac{\partial^2 \mathbf{z}_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,j}^l \partial \hat{\mathbf{z}}_{t,k}^l} = 0$. And $\frac{\partial \mathbf{z}_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,j}^l} \cdot \frac{\partial \mathbf{z}_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,k}^l} = 0$
 1239 implies that for each $i = 1, \dots, n$, $\frac{\partial \mathbf{z}_{t,i}^l}{\partial \hat{\mathbf{z}}_{t,j}^l} \neq 0$ for at most one element $j \in \{1, \dots, n\}$. Therefore,
 1240 there is only at most one non-zero entry in each row indexed by i in the Jacobian $\mathbf{J}_{h_t^l}$. Moreover, the
 1241 invertibility of h_t^l necessitates $\mathbf{J}_{h_t^l}$ to be full-rank which implies that there is exactly one non-zero
 1242 component in each row of $\mathbf{J}_{h_t^l}$, implying that the $\mathbf{z}_{t,i}^l$ is component-wise identifiable. \square

1243 C Real-world Implications of Modeling Time Series Data in a Hierarchical 1244 Manner

1245 Real-world scenarios are usually governed by hierarchical layers of temporal latent dynamics instead
 1246 of a single one. Capturing the hierarchical temporal latent dynamics plays a critical role in modeling
 1247 time series data. Here is two real-world scenarios.

1248 **Human Motion Modeling.** Human locomotion unfolds over multiple temporal layers [16, 58, 2]:

- 1249 • **Higher-level latent variables:** Capture coarse movement categories—e.g., walking versus
 1250 running or resting—capturing the switch-like changes in overall gait dynamics.
- 1251 • **Lower-level latent variables:** Resolve fine-grained kinematics such as joint angles, stride
 1252 frequency, and instantaneous speed.

1253 This decomposition isolates semantically meaningful factors: the high-level code tells what activity
 1254 is occurring, while the low-level code details how it is executed. Such a structure supports down-
 1255 stream tasks, including activity recognition, early anomaly detection (e.g., pathological gaits), and
 1256 personalized coaching systems that adapt exercises to an individual’s movement profile.

1257 **Climate Modeling.** Atmospheric processes also exhibit a natural hierarchy [34, 67], which is
1258 shown as follows:

- 1259 • **Top-level latent variables** capture large-scale regimes like monsoon cycles or transitions
1260 between seasons, governing the dominant energy balance of a region.
- 1261 • **Middle-level latent variables** reflect regional or intra-seasonal patterns—e.g., the twelve
1262 traditional solar terms in East Asia or monthly precipitation phases that modulate local
1263 ecosystems.
- 1264 • **Bottom-level latent variables** track short-term fluctuations in temperature, humidity, and
1265 wind that drive daily weather variability.

1266 Modeling these tiers jointly enables scientists to disentangle long-term trends from transient dis-
1267 turbances, improving extended-range forecasts, pinpointing abnormal events (heat-wave onsets,
1268 unseasonal cold snaps), and facilitating attribution studies of climate change.

1269 D Relationship between Time Series Generation and Hierarchical Dynamics

1270 The goal of time series generation is to identify the joint distribution of observed variables of
1271 different time steps, i.e., $p(\mathbf{x}_1, \dots, \mathbf{x}_T)$. To achieve this goal, we should simultaneously capture
1272 long-range structure spanning hundreds of steps and fine-grained, moment-to-moment variability.
1273 Collapsing every timescale into a single latent layer forces one hidden state to shoulder both roles.
1274 This entangles long-term context with short-term detail, squanders model capacity on resolving
1275 incompatible dependencies, and obscures the semantics of the latent dimensions—making directed
1276 manipulation or counterfactual sampling nearly impossible.

1277 By contrast, an explicit hierarchical latent structure allocates a dedicated timescale to each layer: a
1278 slow-clock top tier encodes global intent or seasonality, intermediate tiers refine meso-level patterns,
1279 and a fast-clock bottom tier captures instantaneous noise. This separation prevents semantic mixing,
1280 lets the model reuse parameters efficiently, and provides clearly interpretable handles for controllable
1281 generation (e.g., editing only high-level intent while preserving fine detail).

1282 E Related Works

1283 **Time Series Generation** Generating realistic time series data [64, 33, 1, 12, 76] is important in
1284 numerous fields, including finance, healthcare, and engineering, where access to sufficient data
1285 can be challenging. Traditional generative methods, such as Generative Adversarial Networks
1286 (GANs) [17, 54, 11, 60, 35] and Variational Autoencoders (VAEs) [39, 46, 6], have been widely
1287 explored for time series generation. TimeGAN [71], a GAN-based framework, integrates adversarial
1288 and supervised losses to effectively capture temporal dynamics, demonstrating superior performance
1289 over previous methods in terms of similarity and predictive performance. However, such GAN-based
1290 models often face challenges such as training instability and mode collapse. To overcome these
1291 limitations, VAEs have been explored as a more robust alternative. TimeVAE [9] incorporates domain
1292 knowledge to model temporal patterns like trends and seasonalities, improving both interpretability
1293 and training efficiency. Recent advances have also introduced diffusion-based models [41, 15, 21],
1294 such as Diffusion-TS [72], which leverage disentangled temporal representations and Fourier-based
1295 training objectives to generate high-quality and interpretable time series while mitigating common
1296 limitations of autoregressive approaches. Furthermore, Koopman VAEs (KoVAE) [56] leverage
1297 linear latent dynamics inspired by Koopman theory, enabling the integration of domain knowledge
1298 and stability analysis. These methods collectively illustrate the evolving landscape of time series
1299 generation, with a focus on balancing realism, interpretability, and computational efficiency. Other
1300 relevant works include [36], which synthesize time series by manipulating existing time series.

1301 **Identifiability of Causal Representation Learning** Independent Component Analysis (ICA) has
1302 been widely used to identify latent variables with identifiability guarantees, traditionally assuming
1303 a linear mixing process [7, 30, 27]. However, this linearity constraint restricts its applicability
1304 in more complex scenarios. To handle nonlinear scenarios, researchers have proposed nonlinear
1305 ICA by introducing additional assumptions, such as auxiliary variables [37, 68, 47] or structural
1306 sparsity [75, 74]. Specifically, methods based on auxiliary variables [28, 29] typically assume that

latent sources are conditionally independent given auxiliary information, such as domain indices or temporal segments, thereby enabling identifiability. For unsupervised approaches, structural sparsity in the generative process has been exploited to recover causal latent factors without relying on auxiliary variables [75, 74, 65]. This approach ensures identifiability by enforcing sparsity constraints on the Jacobian of the mixing function, allowing for the identification up to component-wise transformations and permutations.

Temporal observations introduce unique challenges to identifiability, especially in the presence of instantaneous dependencies and dynamic latent structures. [28] achieves identifiability for stationary data using permutation-based contrastive learning, while [29] extends this approach to nonstationary time series by leveraging variability in data segments. Recent methods [69, 70] assume conditional independence of latent variables given their time-delayed parent to identify latent causal relations. However, these assumptions often fail in real-world scenarios where instantaneous dependencies are prevalent, such as in human motion data, making them of limited applicability in such contexts. Recently, [48] proposed an identification framework for instantaneous latent dynamics, introducing a sparse influence constraint to enforce sparsity in both time-delayed and instantaneous causal relationships among latent processes. While these advances enhance identifiability, addressing hierarchical latent dependencies remains a key challenge for further improving the identifiability of causal structures in temporal data.

F Implementation Details

F.1 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

$$p(z_{t-1,1}, z_{t-1,2}, z_{t,1}, z_{t,2}) = \frac{p(z_{t-1,1}, z_{t-1,2}, \epsilon_{t,1}, \epsilon_{t,2})}{|\det \mathbf{J}_{\mathbf{f}}|}, \quad (\text{A48})$$

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 $\prod_i \frac{\partial z_{t,i}}{\epsilon_{t,i}}$. 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 (A48) as follows

$$p(z_{t-1,1}, z_{t-1,2}, z_{t,1}, z_{t,2}) = p(z_{t-1,1}, z_{t-1,2}) \times \frac{p(\epsilon_{t,1}, \epsilon_{t,2})}{|\mathbf{J}_{\mathbf{f}}|} = p(z_{t-1,1}, z_{t-1,2}) \times \frac{\prod_i p(\epsilon_{t,i})}{|\mathbf{J}_{\mathbf{f}}|}. \quad (\text{A49})$$

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., $\epsilon_{t,i} = r_i(\hat{\mathbf{z}}_{t,i}, \{\hat{\mathbf{z}}_{t-\tau}\})$. Then we design a transformation $\mathbf{A} \rightarrow \mathbf{B}$ with low-triangular Jacobian as follows:

$$\underbrace{[\hat{\mathbf{z}}_{t-L}, \dots, \hat{\mathbf{z}}_{t-1}, \hat{\mathbf{z}}_t]^\top}_{\mathbf{A}} \text{ mapped to } \underbrace{[\hat{\mathbf{z}}_{t-L}, \dots, \hat{\mathbf{z}}_{t-1}, \hat{\epsilon}_{t,i}]^\top}_{\mathbf{B}}, \text{ with } \mathbf{J}_{\mathbf{A} \rightarrow \mathbf{B}} = \begin{bmatrix} \mathbb{I}_{n_s \times L} & 0 \\ * & \text{diag} \left(\frac{\partial r_{i,j}}{\partial \hat{\mathbf{z}}_{t,j}} \right) \end{bmatrix}. \quad (\text{A50})$$

Similar to Equation (A49), we can obtain the joint distribution of the estimated dynamics subspace as:

$$\log p(\mathbf{A}) = \underbrace{\log p(\hat{\mathbf{z}}_{t-L}, \dots, \hat{\mathbf{z}}_{t-1}) + \sum_{i=1}^{n_s} \log p(\hat{\epsilon}_{t,i})}_{\text{Because of mutually independent noise assumption}} + \log(|\det(\mathbf{J}_{\mathbf{A} \rightarrow \mathbf{B}})|) \quad (\text{A51})$$

Finally, we have:

$$\log p(\hat{\mathbf{z}}_t | \{\hat{\mathbf{z}}_{t-\tau}\}_{\tau=1}^L) = \sum_{i=1}^{n_s} p(\hat{\epsilon}_{t,i}) + \sum_{i=1}^{n_s} \log \left| \frac{\partial r_i}{\partial \hat{z}_{t,i}} \right| \quad (\text{A52})$$

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 assumption, we can estimate $p(\hat{\mathbf{z}}_{t+1:T} | \hat{\mathbf{z}}_{1:t})$ in a similar way.

F.2 Evident Lower Bound

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

$$\begin{aligned} \ln p(\mathbf{x}_{1:T}) &= \ln \frac{p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T})}{p(\mathbf{z}_{1:T} | \mathbf{x}_{1:T})} = \mathbb{E}_{q(\mathbf{z}_{1:T} | \mathbf{x}_{1:T})} \ln \frac{p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) q(\mathbf{z}_{1:T} | \mathbf{x}_{1:T})}{p(\mathbf{z}_{1:T} | \mathbf{x}_{1:T}) q(\mathbf{z}_{1:T} | \mathbf{x}_{1:T})} \\ &\geq \underbrace{\mathbb{E}_{q(\mathbf{z}_{1:T} | \mathbf{x}_{1:T})} \ln p(\mathbf{x}_{1:T} | \mathbf{z}_{1:T})}_{L_r} - \underbrace{D_{KL}(q(\mathbf{z}_{1:T} | \mathbf{x}_{1:T}) || p(\mathbf{z}_{1:T}))}_{L_{KLD}} = ELBO. \end{aligned} \quad (\text{A53})$$

F.3 Model Details

F.3.1 Reproducibility of Simulation Experiment

For the implementation of baseline models, we utilized publicly released code for TDRL, CaRiNG. Since the source code of IDOL is not available, we implemented it based on TDRL with the sparsity constraint. And the proposed CHiLD is also modified based on the code of TDRL, and shared hyperparameters remain the same.

F.3.2 Reproducibility of Real-world Experiment

We follow the setting of [72]. The model details of the proposed method are shown in Table A5. As for other baselines like KoVAE, Diffusion-TS, TimeGAN, and TimeVAE, we employ the official implementations and use the default hyperparameters for a fair comparison. To achieve the results from the well-fit baselines, we employed the default hyperparameters and tried different values of learning rate for the best models. And the number of latent variables on each real-world dataset is shown in Table A6.

G Experiment Details

G.1 Simulation Experiment

G.1.1 Data Generation Process

As for the temporally latent processes, we use MLPs with the activation function of LeakyReLU to model the sparse time-delayed. That is:

$$\begin{aligned} z_{t,i}^1 &= (\text{LeakyReLU}(W_{i,:}^1 \cdot \mathbf{z}_{t-1}^1, 0.2) + V_{<i,i} \cdot \mathbf{z}_{t,<i}^2) + \epsilon_{t,i}^1 \\ z_{t,i}^2 &= \text{LeakyReLU}(W_{i,:}^2 \cdot \mathbf{z}_{t-1}^2, 0.2) + \epsilon_{t,i}^2, \end{aligned} \quad (\text{A54})$$

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, each independent noise $\epsilon_{t,i}$ is sampled from the distribution of normal distribution. We further let the data generation process from latent variables to observed variables be MLPs with the LeakyReLU units. And the generation procedure can be formulated as follows:

$$\mathbf{x}_t = \text{LeakyReLU}((\text{LeakyReLU}(\mathbf{z}_t^1, 0.2)) + \epsilon_t^0, 0.2), \quad (\text{A55})$$

Table A5: Architecture details. T , length of time series. $|\mathbf{x}_t|$: input dimension. n : latent dimension. LeakyReLU: Leaky 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
Convolution neural networks	$ \mathbf{x}_t $ neurons	Batch Size $\times t \times \mathbf{x}_t $
Concat zero	concatenation	Batch Size $\times T \times \mathbf{x}_t $
Dense	n neurons	Batch Size $\times T \times n$
ψ	Decoder	
Input: $\mathbf{z}_{1:T}$	Latent Variable	Batch Size $\times T \times n$
Dense	$ \mathbf{x}_t $ neurons, Tanh	Batch Size $\times T \times \mathbf{x}_t $
\mathbf{r}	Modular Prior Networks	
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 A6: Number of latent variables on each real-world dataset

	ETTh	Stocks	MuJoco	fMRI	Box	Gestures	Throwcatch	Discussion	Purchases	WalkDog
Sample Size	7524	3686	10000	10000	1889	1687	1022	13760	7481	10087
Dimension of High-level Layer	3	2	4	20	15	15	15	25	25	25
Dimension of Low-level Layer	4	4	10	30	30	30	30	26	26	26
Number of Layer	2	2	2	2	2	2	2	2	2	2

where the dimension of ϵ_t^0 is 1.

We generate six different synthetic datasets. For dataset A, the first layer contains 4 latent variables and the second layer contains 1 latent variable. For dataset B, all the latent variables are in the first layer, since there is only one latent layer. For dataset C, the first layer contains 8 latent variables and the second layer contains 2 latent variables. For dataset D, we consider the second-order latent markov process, and the first and second layers contain 4 and 1 latent variables, respectively. For dataset E, we consider a more complex generation procedure. Specifically, each causal relation and generation using a 2-layer MLP with LeakyReLU activation, which are shown as follows:

$$\begin{aligned}
z_{t,i}^2 &= M^1(M^2(\mathbf{z}_{t-1}^2)) + \epsilon_{t,i}^2 \\
z_{t,i}^1 &= 0.5 * M^3(M^4(\mathbf{z}_{t-1}^1)) + 0.5 * M^4(M^5(\mathbf{z}_{t-1}^2)) + \epsilon_{t,i}^1 \\
\mathbf{x}_{t,i} &= M^6(M^7(\mathbf{z}_t^1)) + \epsilon_t^0,
\end{aligned} \tag{A56}$$

where M^* is a linear layer followed by a LeakyReLU activation. For dataset F, we consider three latent layers with 1, 2, and 4 latent variables. For dataset G, we consider the case with large number of latent layers and dimensions, which contains three latent layers with 8, 8, and 8 latent variables, respectively. The total size of the dataset is 100,000, with 1,024 samples designated as the validation set. The remaining samples are the training set.

G.1.2 Evaluation Metrics

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

G.1.3 More Simulation Experiments of Different Length of Observations

To evaluate our theoretical results that L -layer latent variables require $2L + 1$ consecutive observations, we designed an experiment on a synthetic two-layer model with 8 latent variables in each layer. In our experimental setting, we vary the number of consecutive observations, i.e., 1, 3, 5, denoted by settings J1, J3, and J5, respectively. The results are shown in Table A7:

Table A7: MCC performance under different numbers of consecutive observations.

	Setting	MCC (mean \pm std)
1	J1	0.6859 ± 0.024
3	J3	0.7001 ± 0.008
5	J5	0.7622 ± 0.004

According to the experiment results, we can find that we observe that as the length of the consecutive observation window increases, the recovery performance of the latent variables improves. In particular, when the length is $2L + 1$, i.e., 5, the identifiability result is optimal.

G.2 Real-world Experiment

G.2.1 Dataset Description

The detailed descriptions of the datasets are shown as follows:

- Stock is the Google stock price data from 2004 to 2019. Each observation represents one day and has 6 features.
- ETTh1 dataset contains the data collected from electricity transformers, including load and oil temperature that is recorded every 15 minutes between July 2016 and July 2018.
- fMRI is a benchmark for causal discovery, which consists of realistic simulations of blood-oxygen-level-dependent (BOLD) time series.
- MuJoCo (multi-joint dynamics with contact) is a time series dataset generated by the physics engine.
- Huaman 3.6 is collected over 3.6 million different human poses, viewed from 4 different angles, using an accurate human motion capture system. The motions were executed by 11 professional actors, and covered a diverse set of everyday scenarios including conversations, eating, greeting, talking on the phone, posing, sitting, smoking, taking photos, waiting, and walking in various non-typical scenarios. We randomly use four motions, i.e., Gestures (Ge), Jog (J), CatchThrow (CT), and Walking (W).
- HuamnEVA-I comprises 3 subjects, each performing several action categories. Each pose has 15 joints with three axis. We choose 6 motions, i.e., Discussion (D), Greeting (Gr), Purchases (P), SittingDown (SD), Walking (W), and WalkTogether (WT) for the task of human motion forecasting. Specifically, the ground truth motion of the body was captured using a commercial motion capture (MoCap) system from ViconPeak 5. The system uses reflective markers and six 1M-pixel cameras to recover the 3D position of the markers and thereby estimate the 3D articulated pose of the body. We consider the joints as latent variables and the signals recorded from the system as observations.
- Weather⁴ dataset offers 10-minute summaries from an automated rooftop station at the Max Planck Institute for Biogeochemistry in Jena, Germany.
- WeatherBench⁵ is a benchmark dataset for data-driven medium-range weather forecasting. It repackages forty years (1979-2018) of ERA5 global reanalysis into machine-learning-ready NetCDF tensors sampled every six hours. Fourteen core surface and pressure-level variables—geopotential height, temperature, humidity, wind components, vorticity, potential vorticity, cloud cover, precipitation, solar radiation, and more—are provided on latitude-longitude grids of 0.25° , 1.406° , 2.812° , and 5.625° . Static fields such as land-sea mask, soil type, orography, and grid coordinates are included.

⁴<https://www.bgc-jena.mpg.de/wetter/>

⁵<https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2020MS002203>

- CESM2 ⁶ delivers 100 fully coupled Earth system simulations at 1° resolution spanning 1850-2100 under CMIP6 historical and SSP3-7.0 forcing. Ensemble spread arises from distinct ocean-atmosphere initial states: ten members seeded from evenly spaced low-drift periods; eighty members created by tiny 10-14 K atmospheric perturbations applied to four AMOC-phased control years; and ten MOAR members useful for regional downscaling.

G.2.2 Evaluation Metric

- Context-Frechet Inception Distance (Context-FID) score [35] quantifies the quality of the synthetic time series samples by computing the difference between representations of time series that fit into the local context.
- Correlational score [49] uses the absolute error between cross-correlation matrices by real data and synthetic data to assess the temporal dependency.

G.3 More Experiment Results

G.3.1 Experiment results on other datasets

We would like to clarify that while some absolute improvements may seem small, e.g., 0.079 (ours) vs. 0.098 (IDOL) in the fmri dataset, it represents a 19% relative improvement. Focusing on percentage improvements, we observe significant gains of 34.07% and 23.91% of two metrics, showing the effectiveness of CHiLD.

Table A8: Experiment results on other datasets.

		ETTh	Stocks	Mujoco	fmri
Context-FID Score	CHiLD	0.111(0.029)	0.001(0.000)	0.238(0.055)	0.025(0.001)
	KoVAE	0.120(0.009)	0.095(0.013)	0.024(0.009)	1.086(0.142)
	Diffusion-TS	0.116(0.010)	0.147(0.025)	0.013(0.001)	0.105(0.006)
	TimeGAN	0.300(0.013)	0.103(0.013)	0.563(0.052)	1.292(0.218)
	IDOL	0.077(0.013)	0.022(0.002)	0.062(0.012)	0.065(0.002)
	cwVAE	0.892(0.059)	0.807(0.252)	1.010(0.195)	0.896(0.103)
	TimeVAE	0.805(0.186)	0.215(0.035)	0.251(0.015)	14.449(0.969)
Correlational Score	CHiLD	0.008(0.000)	0.001(0.000)	0.001(0.000)	0.079(0.003)
	KoVAE	0.045(0.006)	0.007(0.002)	0.203(0.031)	11.832(0.007)
	Diffusion-TS	0.049(0.008)	0.004(0.001)	0.193(0.027)	1.411(0.042)
	TimeGAN	0.210(0.006)	0.063(0.005)	0.886(0.039)	23.502(0.039)
	IDOL	0.002(0.000)	0.006(0.000)	0.002(0.000)	0.098(0.003)
	cwVAE	0.070(0.001)	0.053(0.001)	0.050(0.001)	18.434(0.030)
	TimeVAE	0.111(0.020)	0.095(0.008)	0.388(0.041)	17.296(0.526)

G.3.2 Ablation Study

To evaluate the effectiveness of each module, we further devise two variants of the proposed CHiLD as follows. 1) **CHiLD-KL**: we remove the *KL* divergence restriction of prior estimation. 2) **CHiLD-C**: we do not use the context information for the proposed CHiLD. Experiment results are shown in Table A9, we can find that incorporating the contextual observation and KL divergence has a positive impact on the overall performance of the model.

Table A9: Experiment results of different model variants on the Humaneva-Box and Humaneva-Throwcatch datasets.

	Box		Throwcatch	
Model	Context-FID Score	Correlational Score	Context-FID Score	Correlational Score
CHiLD	0.007(0.001)	0.021(0.002)	0.063(0.056)	0.662(0.007)
CHiLD-KL	0.038(0.003)	0.637(0.006)	0.077(0.0044)	0.974(0.0039)
CHiLD-C	0.017(0.0034)	0.036(0.0015)	0.093(0.011)	1.077(0.0098)

⁶<https://climatedata.ibs.re.kr/data/cesm2-lens>

1458 H Broader Impacts

1459 The proposed **CHiLD** method offers a significant advancement in identifying hierarchical latent
1460 dynamics, which is crucial for generating realistic and interpretable time series data across a wide
1461 range of domains such as healthcare, finance, climate science, and autonomous systems. By enabling
1462 the generation of time series that are both data-driven and causally grounded, our method not only
1463 enhances the credibility of the generated sequences but also improves the performance of downstream
1464 tasks like forecasting, anomaly detection, and decision-making. These improvements foster increased
1465 trust and adoption of machine learning models in critical applications, particularly where reliable
1466 data-driven insights are of paramount importance.

1467 Furthermore, our work contributes to the broader field of generative modeling by providing a
1468 principled framework for hierarchical latent dynamics. This framework mitigates issues like mode
1469 collapse, a common problem in generative models, and enhances the model’s generalizability across
1470 various datasets. As a result, **CHiLD** stands as a powerful tool for both researchers and practitioners
1471 looking to explore complex time series data and apply generative modeling techniques in real-
1472 world scenarios. This method paves the way for future research in more complex domains, such
1473 as controllable video generation, neuroscience, and predictive analytics in environmental systems,
1474 thereby further extending its applicability.