CaRiNG: Learning Temporal Causal Representation under Non-Invertible Generation Process

Guangyi Chen [*1 2] **Yifan Shen** [*1] **Zhenhao Chen** [*1] **Xiangchen Song** [2]
Yuewen Sun [1] **Weiran Yao** [3] **Xiao Liu** [1] **Kun Zhang** [1 2]

Abstract

Identifying the underlying time-delayed latent causal processes in sequential data is vital for grasping temporal dynamics and making downstream reasoning. While some recent methods can robustly identify these latent causal variables, they rely on strict assumptions about the invertible generation process from latent variables to observed data. However, these assumptions are often hard to satisfy in real-world applications containing information loss. For instance, the visual perception process translates a 3D space into 2D images, or the phenomenon of persistence of vision incorporates historical data into current perceptions. To address this challenge, we establish an identifiability theory that allows for the recovery of independent latent components even when they come from a nonlinear and non-invertible mix. Using this theory as a foundation, we propose a principled approach, **CaRiNG**, to learn the **Ca**usal **R**epresentat**i**on of **N**on-invertible **G**enerative temporal data with identifiability guarantees. Specifically, we utilize temporal context to recover lost latent information and apply the conditions in our theory to guide the training process. Through experiments conducted on synthetic datasets, we validate that our **CaRiNG** method reliably identifies the causal process, even when the generation process is non-invertible. Moreover, we demonstrate that our approach considerably improves temporal understanding and reasoning in practical applications. Code can be accessed through https://github.com/sanshuiii/CaRiNG.

[*]Equal contribution [1]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE [2]Carnegie Mellon University, Pittsburg, US [3]Salesforce, San Francisco, US. Correspondence to: Kun Zhang <kunz1@cmu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

Sequential data, including video, stock, and climate observations, are integral to our daily lives. Gaining an understanding of the causal dynamics in such time series data has always been a crucial challenge (Berzuini et al., 2012; Ghysels et al., 2016; Friston, 2009) and has attracted considerable attention. The core of this task is to identify the underlying causal dynamics in the data we observe.

Towards this goal, we focus on Independent Component Analysis (ICA) (Hyvärinen & Oja, 2000), which is a classical method for decomposing the latent signals from mixed observation. Recent advancements in nonlinear ICA (Hyvarinen & Morioka, 2016; 2017; Hyvarinen et al., 2019; Khemakhem et al., 2020; Sorrenson et al., 2020; Hälvä & Hyvarinen, 2020) have yielded robust theoretical evidence for the identifiability of latent variables, and enabled the use of deep neural networks to address complex scenarios. For example, by assuming the latent variables in the data generation process are mutually independent, and leveraging the auxiliary side information such as time index, domain index, or class label, (Hyvarinen & Morioka, 2017; Hyvarinen et al., 2019; Hälvä & Hyvarinen, 2020) have demonstrated the strong identifiability results. (Hälvä et al., 2021; Klindt et al., 2020; Yao et al., 2022b;a; Lachapelle et al., 2022) further extend this nonlinear ICA framework into scenarios of the time-delayed dynamical systems, which allows the temporal transitions among the latent variables.

However, these nonlinear ICA-based methods usually assume that the mixing function (the generation process from sources to observations) is invertible, which may be difficult to satisfy in real-world scenarios, such as the 3D to 2D projection in the visual process. As shown in Figure 1 (a) and (b), we provide two intuitive instances of the real videos to illustrate how the non-invertibility happens. In (a), when object occlusions occur, information from the obstructed object is lost in the generation process of the current time step, which causes non-invertibility. In (b), the persistence of vision introduces non-invertibility, since the mixing process of the current time step utilizes the history information. We further found that the violation of this invertibility assumption may cause the nonlinear ICA method to yield

poor identification performance. In part (c) of Figure 1, we demonstrate that TDRL, one of the typical nonlinear ICA-based methods making the invertibility assumption, markedly degrades its performance in identifying the latent variables with increasing non-invertibility. It motivates us to extend the current nonlinear ICA methods to consider non-invertible mixing function.

In this paper, to tackle the challenges above, we propose to leverage the temporal context for retrieving missing information caused by the non-invertible mixing function, mirroring the intuitive mechanisms of human perception. For instance, when we encounter an object with occlusion, our natural inclination is to draw from historical data to reconstruct the obscured portion. We demonstrate that, even when the generation process is non-invertible, the derived latent causal representation remains identifiable if the latent variables can be expressed as an arbitrary function combining the current observation with its history. Built upon this identification theorem, we introduce a principled approach, named **CaRiNG**, that learns the function to integrate historical data to compensate for the latent information lost due to non-invertibility. This approach extends the Sequential Variational Autoencoder (Sequential VAE (Chung et al., 2015; Li & Mandt, 2018)) with two distinct modifications. Firstly, it incorporates history (or context) information directly into the encoder. Specifically, we transform step-to-step mapping (from current observation to the current latent variable) into sequence-to-step mapping (from current observation and temporal context to the current latent variable). Secondly, a specialized prior module is introduced to determine the prior distribution of latent variables using the normalizing flow (Dinh et al., 2016), ensuring the imposition of an independent noise condition. We evaluate our method using both synthetic and real-world data. Using synthetic data, we design datasets with a non-invertible mixing function to measure identifiability. For real-world applications, **CaRiNG** is deployed in a traffic accident reasoning task, a scenario in which the intricate traffic dynamics introduce considerable non-invertibility. Experimental outcomes reveal that our method significantly outperforms other temporal representation learning methods for identifying causal representations amid non-invertible generation processes. Furthermore, this causal representation has proven instrumental in enhancing video reasoning tasks.

Key Insights and Contributions of our research include:

- To the best of our understanding, this paper presents the first identifiability theorem that accommodates a non-invertible generation process, which complements the existing body of the nonlinear ICA theory.
- We present a principled approach, **CaRiNG**, to learn the latent causal representation from temporal data under non-invertible generation processes with identifiability guarantees, by integrating temporal context information to recover the lost information.
- Our evaluations across synthetic and real-world datasets demonstrate the **CaRiNG**'s effectiveness for learning the identifiable latent causal representation, leading to enhancements in video reasoning tasks.

2. Problem Setup

2.1. Non-invertible Temporal Generative Process

Denote $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ as the observed d-dimensional time series data at T discrete time steps. Each observation $\mathbf{x}_t \in \mathbb{R}^d$ is generated from a nonlinear mixing function $\mathbf{g}$ that maps $r+1$ adjacent latent variables $\mathbf{z}_{t:t-r}$ to $\mathbf{x}_t$, where $\mathbf{z}_{t:t-r}$ refers to $\{\mathbf{z}_t, \mathbf{z}_{t-1}, \cdots, \mathbf{z}_{t-r}\}$. We have $\mathbf{z}_t \in \mathbb{R}^n$. For every $i \in 1, \ldots, n$, the variable z_{it} of $\mathbf{z}_t$ is derived from a stationary, non-parametric time-delayed causal relation:

$$\begin{aligned} &\underbrace{\mathbf{x}_t = \mathbf{g}(\mathbf{z}_{t:t-r})}_{\text{Nonlinear mixing}}, \\ &\underbrace{z_{it} = f_i\left(\{z_{j,t'} | z_{j,t'} \in \mathbf{Pa}(z_{it})\}, \epsilon_{it}\right)}_{\text{Stationary non-parametric transition}}. \end{aligned} \tag{1}$$

Note that with non-parametric causal transitions, the noise term $\epsilon_{it} \sim p_{\epsilon_i}$ (where p_{ϵ_i} denotes the distribution of ϵ_{it}) and the time-delayed parents $\mathbf{Pa}(z_{it})$ of z_{it} (i.e., the set of latent factors that directly cause z_{it}) are interacted and transformed in an arbitrarily nonlinear way to generate z_{it}. τ denotes the transition time lag. The components of $\mathbf{z}_t$ are mutually independent conditional on history variables $\mathbf{Pa}(\mathbf{z}_t)$.

In this case, one cannot recover $\mathbf{z}_t$ from $\mathbf{x}_t$ alone due to the non-invertibility of $\mathbf{g}$. Without extra assumptions, it is definitely non-identifiable. As a result, we assume that there exists a time lag μ and a nonlinear function $\mathbf{m}$ which can map a series of observations to latent variable $\mathbf{z}_t$, i.e.,

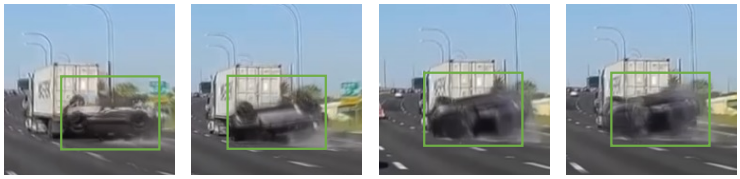
$$\mathbf{z}_t = \mathbf{m}(\mathbf{x}_{t:t-\mu}). \tag{2}$$

Once we successfully recover the information lost due to non-invertibility from the context, the classical nonlinear ICA algorithm can be used to solve this problem.

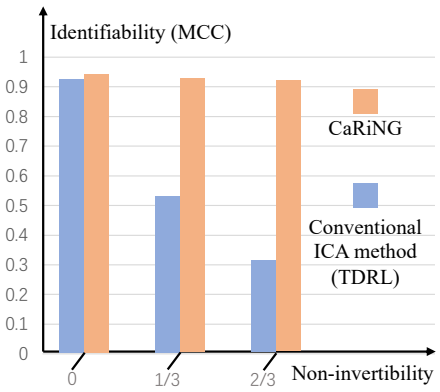
2.2. Identification of the Latent Causal Processes

Definition 1 (Identifiable Latent Causal Process)**.** *Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ be a sequence of observed variables generated by the true temporally causal latent processes specified by $(f_i, p(\epsilon_i), \mathbf{g})$ given in Eq 1. A learned generative model $(\hat{f}_i, \hat{p}(\epsilon_i), \hat{\mathbf{g}})$ is observational equivalent to $(f_i, p(\epsilon_i), \mathbf{g})$ if the model distribution $p_{\hat{f}_i, \hat{p}_\epsilon, \hat{\mathbf{g}}}(\mathbf{x}_{1:T})$ matches the data distribution $p_{f_i, p_\epsilon, \mathbf{g}}(\mathbf{x}_{1:T})$ for any value of*

(a) Non-invertibility by occlusion



(b) Non-invertibility by vision persistence



(c) Identifiablity V.S. non-invertibility

Figure 1: **Motivations of the non-invertible generation process.** (a) The occlusions raise the non-invertibility since the measured observation cannot cover the obstructed objects. (b) The vision persistence, shown with the high-speed movement of a crashing car, describes the generation process that jointly involves the current state and previous, and causes the non-invertibility. (c) The identifiability of conventional methods, such as TDRL (Yao et al., 2022a) (blue), drops drastically with the increase of non-invertibility, while the identifiability of our method (marked in orange) still holds. The levels of non-invertibility are defined by removing $0, 1/3$, and $2/3$ dimensions of $\mathbf{z_t}$ when generating $\mathbf{x_t}$. For example, when the dimension of $\mathbf{z_t}$ is 6, $2/3$ non-invertibility means that we remove 4 variables of $\mathbf{z_t}$ and use only 2 variables to generate $\mathbf{x_t}$.

$\mathbf{x}_t$. *We say latent causal processes are identifiable if observational equivalence can lead to a version of latent variable* $\mathbf{z}_t = \mathbf{m}(\mathbf{x}_{t:t-\mu})$ *up to permutation* π *and component-wise invertible transformation* $\mathcal{T}$*:*

$$\begin{aligned} & p_{\hat{f}_i, \hat{p}_{\epsilon_i}, \hat{\mathbf{g}}}(\mathbf{x}_{1:T}) = p_{f_i, p_{\epsilon_i}, \mathbf{g}}(\mathbf{x}_{1:T}) \\ \Rightarrow & \hat{\mathbf{m}}(\mathbf{x}_{t:t-\mu}) = (\mathcal{T} \circ \pi \circ \mathbf{m})(\mathbf{x}_{t:t-\mu}), \forall \mathbf{x}_{t:t-\mu} \in \mathbb{R}^{\mu+1}. \end{aligned} \tag{3}$$

Different from the existing literature, we involve $\mathbf{m}$ in the above definition since it serves implicitly as a property of the mixing function $\mathbf{g}$, although it does not explicitly participate in the generation process. Furthermore, the identifiability of $\mathbf{g}$ is different. In previous nonlinear ICA methods (Yao et al., 2022a; Hyvarinen & Morioka, 2017), the mixing function $\mathbf{g}$ is identifiable. However, in our case, we cannot find the identifiable mixing function since the information loss is caused by non-invertibility. Instead, we can obtain a component-wise transformation of a permuted version of latent variables $\hat{\mathbf{z}}_t = \mathbf{m}(\mathbf{x}_{t:t-\mu})$. The latent causal relations are also identifiable, up to a permutation π and component-wise invertible transformation $\mathcal{T}$, i.e., $\hat{\mathbf{f}} = \mathcal{T} \circ \pi \circ \mathbf{f}$, once $\mathbf{z}_t$ is identifiable. Because in the time-delayed causally sufficient system, the conditional independence relations fully characterize time-delayed causal relations when we assume no latent causal confounders in the (latent) causal processes.

2.3. Illustrations of the Problem Setup

Intuitive Illustration with Visual Persistence. Consider a rapidly moving ball on a two-dimensional plane as described in figure 2. The horizontal and vertical coordinates of the ball's position at any given moment can be represented by the latent variable $\mathbf{z}_t \in \mathbb{R}^2$. We assume that the ball follows a curved trajectory constrained by the nonlinear function $\mathbf{f}$ as it moves.

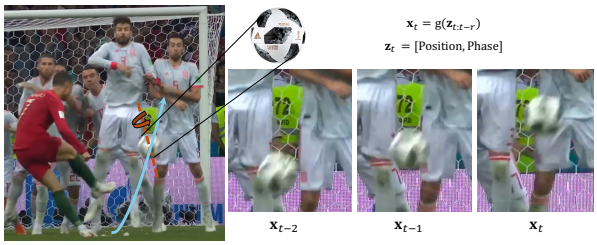


Figure 2: **An intuitive illustration of a moving football with a visual persistence effect.** Considering the generating process $\mathbf{x}_t = \mathbf{g}(\mathbf{z}_{t:t-r})$, $\mathbf{x}_t$ denotes the observed football with motion blur, and $\mathbf{z}_t$ denotes the position and phase of the ball. Recovering the latent variables from a single observation will be difficult, which introduces non-invertibility.

Suppose that we observe the ball with a visual persistence effect, where each observation $\mathbf{x}_t$ captures several consecutive latent variables as $\mathbf{x}_t = \mathbf{g}(\mathbf{z}_{<t})$. The mixing function $\mathbf{g}$ refers to the weighted sum of the images obtained through multiple exposures, which is what a person ultimately observes as $\mathbf{x}_t$. In this case, the invertibility of the mapping from $\mathbf{z}_t$ to $\mathbf{x}_t$ is compromised since the current frame also contains the latent information from previous frames.

Mathematical Illustration. Besides, we provide a mathematical example to demonstrate the existence of function $\mathbf{m}$ in Eq 2. Following the concept of visual persistence, let the current observation be a weakened previous observation overlaid with the current image of the object, i.e., $\mathbf{x}_t = \mathbf{z}_t + \frac{1}{2}\mathbf{x}_{t-1} = \sum_{i=1}^{\infty} \left(\frac{1}{2}\right)^i \mathbf{z}_{t-i}$ (Wolford, 1993).

Given an extra observation, the current latent variable can be rewritten as $\mathbf{z}_t = \mathbf{x}_t - \frac{1}{2}\mathbf{x}_{t-1}$. Thereby we can easily recover latent variables that cannot be obtained from a single observation, i.e., $\mathbf{z}_t = \mathbf{m}(\mathbf{x}_{t:t-1}) = \mathbf{x}_t - \frac{1}{2}\mathbf{x}_{t-1}$.

Illustration of time-delayed temporal relations. Here, we assume that there are only time-delayed temporal relations in the time series system. In other words, any instantaneous relations will not fall into the discussion. Generally speaking, a group of objects that have instantaneous relations with each other would be treated as one single variable. For example, within a video sequence, a ball in motion may be conceptualized as a cluster of pixels that move consistently and simultaneously (instantaneous relations). This pattern can help distinguish the ball from the others, which potentially provides a principle to extract concepts from time series data like video, motion sequence, etc.

3. Identifiability Theory

In this section, we demonstrate that, given certain mild conditions, the learned causal representation $\mathbf{z}_t$ is identifiable up to permutation and a component-wise transformation. This holds even if the mixing function $\mathbf{g}$ is non-invertible. Firstly, we present the identifiability results when faced with a non-invertible mixing function and stationary transitions. Subsequently, we address the gap between permutation-scaling Jacobian and identifiability. Lastly, by leveraging side information such as the domain index and label, we illustrate how identifiability can be achieved even in a non-stationary context. The proofs are available in Appendix A1.

3.1. Identifiability under Non-Invertible Generative Process

W.L.O.G., we first consider a simplified case with $\tau = r + 1$ and context length μ, which infers such process:

$$\mathbf{x}_t = \mathbf{g}(\mathbf{z}_{t:t-r}), \quad z_{it} = f_i\left(\mathbf{z}_{t-1:t-r-1}, \epsilon_{it}\right), \tag{4}$$

where a function $\mathbf{m}$ satisfying $\mathbf{z}_t = \mathbf{m}(\mathbf{x}_{t:t-\mu})$ exists. When taking $r = 0$, the time delay is present only in transitions and is absent in the generation process. Taking $r > 0$ leads us to a more intricate scenario, where the mixing function encompasses not just the latent causal variables of the current time step, but also the information of previous steps, termed the Time-delayed Mixing Process. Such a scenario is compelling, acknowledging that the mixing process can be influenced by time-delayed effects. To illustrate, human visual perception provides a fitting example: the phenomenon known as the persistence of vision reveals that humans retain impressions of a visual stimulus even after its cessation (Coltheart, 1980). The extensions for any time lag τ will be discussed in Appendix A1.4.

Theorem 1 (Identifiability under Non-invertible Generative Process)**.** *For a series of observations $\mathbf{x}_t \in \mathbb{R}^d$ and estimated latent variables $\hat{\mathbf{z}}_t \in \mathbb{R}^n$, suppose there exists function $\hat{\mathbf{g}}, \hat{\mathbf{m}}$ which is subject to observational equivalence,*

$$\mathbf{x}_t = \hat{\mathbf{g}}(\hat{\mathbf{z}}_{t:t-r}), \quad \hat{\mathbf{z}}_t = \hat{\mathbf{m}}(\mathbf{x}_{t:t-\mu}). \tag{5}$$

If assumptions

- *(Smooth and Positive Density) the probability density function of latent variables is third-order differentiable and positive in R^n,*
- *(conditional independence) the components of $\hat{\mathbf{z}}_t$ are mutually independent conditional on $\hat{\mathbf{z}}_{t-1:t-r-1}$,*
- *(sufficiency) let $\eta_{kt} \triangleq \log p(z_{kt}|\mathbf{z}_{t-1:t-r-1})$, and*

$$\mathbf{v}_{lt} \triangleq \Big(\frac{\partial^2 \eta_{1t}}{\partial z_{1t}\partial z_{l,t-r-1}}, ..., \frac{\partial^2 \eta_{nt}}{\partial z_{nt}\partial z_{l,t-r-1}}, \frac{\partial^3 \eta_{1t}}{\partial z_{1t}^2\partial z_{l,t-r-1}}, ..., \frac{\partial^3 \eta_{nt}}{\partial z_{nt}^2\partial z_{l,t-r-1}}\Big)^\intercal, \tag{6}$$

  for $l = 1, 2, \cdots, n$. For each value of $\mathbf{z}_t$, there exists $2n$ different values of $z_{l,t-r-1}$ such that the $2n$ vector functions $\mathbf{v}_{lt} \in \mathbf{R}^{2n}$ are linearly independent,

are satisfied, then $\mathbf{z}_t$ must be a component-wise transformation of a permuted version of $\hat{\mathbf{z}}_t$ with regard to context $\{\mathbf{x}_j \mid \forall j = t, t-1, \cdots, t-\mu-r\}$.

The proof of Theorem 1 can be found in Appendix A1.1. It is inspired by (Yao et al., 2022a), which follows the line of (Hyvarinen et al., 2019).

Besides, the nonstationary transition can also help to improve the identifiability of **CaRiNG**. As shown in the sufficiency assumption in Theorem 1, the identifiability relies on the sufficient changes of the conditional distribution $p(z_{kt}|\mathbf{z}_{t-1:t-r-1})$. When the distribution of the noise term varies between different domains, the domain index can serve as an auxiliary variable to improve this sufficiency since both domain dynamics and history variables can provide changes. More discussions are in Appendix A1.6.

3.2. Continuity for Permutation Invariance

In this subsection, we will introduce permutation invariance for further discussion.

Definition 2 (Permutation Invariance)**.** *Following Definition 1, if π is a fixed permutation and $\mathcal{T}$ is a component-wise invertible transformation which may vary across different time steps, we call this identifiability under Permutation Invariance.*

Let us further consider a more general scenario, with $\mathbf{x}_t \in \mathcal{X} \subseteq \mathbb{R}^d$ and $\mathbf{z}_t \in \mathcal{Z} \subseteq \mathbb{R}^n$, i.e., the probability density of $\mathbf{z}_t$ does not have to be non-zero everywhere in

$\mathbb{R}^n$. To establish identifiability, numerous existing nonlinear ICA-based methods (Yao et al., 2022b;a; Hyvarinen et al., 2019; Hälvä et al., 2021) utilize the Jacobian matrix, denoted by $\mathbf{H} = \frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}$, which captures the relationship between ground truth and estimated latent variables. These methods propose that the learned latent variables are identifiable if $\mathbf{H}_{ij} \cdot \mathbf{H}_{ik} = 0$ for $j \neq k$ (with only a single non-zero element in each row or column). $\mathbf{H}$ corresponds to the Jacobian matrix of the function $\mathbf{h} \triangleq \mathbf{m} \circ \hat{\mathbf{g}}$ in our scenario (or $\mathbf{g}^{-1} \circ \hat{\mathbf{g}}$ for the general scenario). However, it is crucial to highlight an often overlooked shortcoming: this condition alone is insufficient to establish identifiability when dealing with non-linear generation processes. Concurrently to our work, (Lachapelle et al., 2023) also arrived at the difference between local and global disentanglement, and achieved the global disentanglement under the additive decoding case. Alternatively, we demonstrate the identifiability under the permutation invariance and focus on a more general case without the block-specific decoder assumptions. While in linear ICA, given that the Jacobian remains constant, this condition indeed equates to identifiability. Yet, in nonlinear ICA, the Jacobian matrix, being a function of $\hat{\mathbf{z}}$, can vary with different $\hat{\mathbf{z}}$ values, potentially rendering the mapping unpredictable. A comprehensive discussion is available in Appendix A1.5.

To solve this issue in the nonlinear system, we provide two more assumptions. The domain $\hat{\mathcal{Z}}$ of $\hat{\mathbf{z}}$ should be path-connected, i.e., for any $\mathbf{a}, \mathbf{b} \in \hat{\mathcal{Z}}$, there exists a continuous path connecting $\mathbf{a}$ and $\mathbf{b}$ with all points of the path in $\hat{\mathcal{Z}}$. In addition, function $\mathbf{h}$ is second-order differentiable and holds the non-degeneracy condition.

For clarification, the condition that a function $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^n$ is invertible, or equivalently the non-vanishing of the determinant of the Jacobian matrix $\mathbf{H}_h$, is called the non-degeneracy condition. We first define the partially invertible function, and then give the non-degeneracy condition on it.

Definition 3 (Partially Invertiblility)**.** *A function $\mathbf{z} = \mathbf{h}(\hat{\mathbf{z}}, \mathbf{c})$, where $\mathbf{z}, \hat{\mathbf{z}} \in \mathbb{R}^n$ and $\mathbf{c} \in \mathbb{R}^m$, is partially invertible, if and only if for any given $\mathbf{c}$, the rest part $\mathbf{h}_{\mathbf{c}} : \mathbb{R}^n \to \mathbb{R}^n$ is always invertible.*

Definition 4 (Non-degeneracy Condition of Partially Invertible Functions)**.** *The non-degeneracy condition of a partially invertible function $\mathbf{z} = \mathbf{h}(\hat{\mathbf{z}}, \mathbf{c})$ is that for any given $\mathbf{c}$, the determinant of the Jacobian matrix $\mathbf{H}_{\mathbf{h}_{\mathbf{c}}}$ of $\mathbf{h}_{\mathbf{c}}$ is always non-zero.*

Lemma 1 (Disentanglement with Continuity)**.** *For second-order differentiable invertible function $\mathbf{h}$ defined on a path-connected domain $\hat{\mathcal{Z}} \subseteq \mathbb{R}^n$ which satisfies $\mathbf{z} = \mathbf{h}(\hat{\mathbf{z}})$, suppose the non-degeneracy condition holds. If there exists at most one non-zero entry in each row of the Jacobian matrix $\mathbf{H} = \frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}$, the identifiability under Permutation Invariance can be established.*

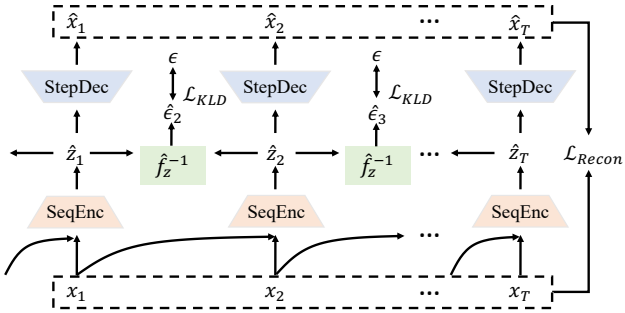


Figure 3: **The overall framework of CaRiNG.** It consists of three main modules, including the sequence-to-step encoder, step-to-step decoder, and the transition prior module, which is represented as SeqEnc, StepDec, and $\hat{\mathbf{f}}_{\mathbf{z}}^{-1}$ in a different color, respectively. The model is trained with both $\mathcal{L}_{Recon}$ and $\mathcal{L}_{KLD}$.

Furthermore, when the Jacobian matrix is more than a function of $\hat{\mathbf{z}}$, but also is influenced by a side information $\mathbf{c}$, the identifiability can be guaranteed under mild extra conditions.

Lemma 2 (Disentanglement with Continuity under Side Information)**.** *For second-order differentiable invertible function $\mathbf{h}$ defined on a path-connected domain $\hat{\mathcal{Z}} \times \mathcal{C} \subseteq \mathbb{R}^{n+m}$ which satisfies $\mathbf{z} = \mathbf{h}(\hat{\mathbf{z}}, \mathbf{c})$, suppose the non-degeneracy condition holds. If there exists at most one non-zero entry in each row of the Jacobian matrix $\mathbf{H}(\mathbf{c}) = \frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}$, the identifiability under Permutation Invariance can be established.*

With Lemma 2, we can further extend Theorem 1 to guarantee permutation invariance even when the probability density of $\mathbf{z}$ is not positive everywhere on $\mathbb{R}^n$, as long as appropriate continuity conditions are satisfied. This serves as a valuable complement to the existing theory of nonlinear ICA, which further relaxes the required assumptions. This relaxation enhances the robustness of **CaRiNG** and makes it more adaptable to diverse and complex data, thus improving its applicability in practical settings.

Proposition 1. *For a series of observations $\mathbf{x}_t \in \mathcal{X} \subseteq \mathbb{R}^d$ and estimated latent variables $\hat{\mathbf{z}}_t \in \mathcal{Z} \subseteq \mathbb{R}^n$, suppose there exists function $\hat{\mathbf{g}}, \hat{\mathbf{m}}$ which subject to observational equivalence, i.e.,*

$$\mathbf{x}_t = \hat{\mathbf{g}}(\hat{\mathbf{z}}_{t:t-r}), \quad \hat{\mathbf{z}}_t = \hat{\mathbf{m}}(\mathbf{x}_{t:t-\mu}). \tag{7}$$

where $\mathbf{g}, \hat{\mathbf{g}}, \mathbf{m}, \hat{\mathbf{m}}$ are second-order differentiable. In addition, if assumptions the same as Theorem 1 are satisfied, then the identifiability of $\mathbf{z}_t$ under Permutation Invariance can be established.

4. Approach

Given our results on identifiability, we introduce our **CaRiNG** approach. This aims to estimate the latent causal

dynamics presented in Eq 1, even when faced with a non-invertible mixing procedure. To achieve this, **CaRiNG** builds upon the Sequential Variational Auto-Encoders (Sequential VAE (Chung et al., 2015; Li & Mandt, 2018)) and incorporates three primary modules: the sequence-to-step encoder (SeqEnc), the step-to-step decoder (StepDec), and the transition prior module ($\hat{\mathbf{f}}_{\mathbf{z}}^{-1}$). Through Sequential VAE, we ensure the reconstruction capability from latent variables to observed variables. Meanwhile, in contrast to the Gaussian prior in VAEs, our method employs normalizing flow to control the prior distribution, ensuring that the latent variables satisfy the assumed conditional independence. During the training phase, we integrate the conditions from Sec. 3 as constraints and adopt two corresponding loss functions.

Overall Framework. As visualized in Figure 3, our framework starts by acquiring the latent causal representation via a sequence-to-step encoder, whose input and output are a sequence of observations $\mathbf{x}_{t:t-\mu}$ and the estimated latent variable $\hat{\mathbf{z}}_t$. Formally, it denotes the inference process of $q(\hat{\mathbf{z}}_t|\mathbf{x}_{t:t-\mu})$, which corresponds to the function $\mathbf{m}$ in Eq 2. Following this, observations are generated from the latent space through a step-to-step decoder $p(\hat{\mathbf{x}}_t|\hat{\mathbf{z}}_t)$, which implies the mixing function $\mathbf{g}$ as mentioned in Eq 1. To learn the independent latent variables, we apply a constraint using the KL divergence between the posterior distribution of learned latent variables and a prior distribution which is subject to our conditional independence assumption in Theorem 1. The estimation of the prior distribution motivates us to utilize a normalizing flow, converting the prior distribution into Gaussian noise, represented as $\hat{\epsilon}_{it} = f_i^{-1}(\hat{z}_{it}, \hat{\mathbf{z}}_{t-1:t-\tau})$. Moreover, a reconstruction loss between the ground truth and generated observations is integrated for model training. A detailed exploration of all modules and losses is forthcoming.

Sequence-to-Step Encoder and Step-to-Step Decoder. Drawing inspiration from the capability of the human visual system, we utilize temporal context to reclaim the information lost due to non-invertible generation. The human visual system adeptly fills in occluded segments by recognizing coherent motion cues (Palmer, 1999; Wertheimer, 1938; Spelke, 1990). Assuming there's a function that captures all latent information from the current observation and its temporal context, we can retrieve the latent causal process with identifiability, i.e. $\mathbf{m}$ exists. Various non-linear models are suitable for estimating this function, taking a sequence of observations, $\mathbf{x}_{t:t-\mu}$, with a lag of μ as inputs, and yielding the estimated latent representation of the current time step as output. In our experiments, we utilize both Multi-Layer Perceptron (MLP) (Werbos, 1974) and Transformer (Vaswani et al., 2017), catering to different complexities. Given the estimated latent variable $\hat{\mathbf{z}}_t$, a step-to-step decoder is employed to generate the current observation $\mathbf{x}_t$. For practical implementation, one MLP is sufficient.

Transition Prior Module. To uphold the conditional independence assumption, we propose to minimize the KL divergence between the posterior distribution and a hard-coding prior distribution with such property. The constraint indicates that current latent variables are mutually independent, conditioned on historical latent variables. Formally, by hard-coding the prior distribution we enforce $\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1:t-\tau}$ to be mutually independent. By minimizing the KL divergence, we expect the posterior to be subject to the assumption as well, such as $\hat{\mathbf{z}}_t|\hat{\mathbf{x}}_{t:t-\mu}, \hat{\mathbf{z}}_{t-1:t-\tau}$ are mutually independent. Direct estimation of the prior, which has an arbitrary density function, poses challenges. As a solution, we introduce a transition prior module that facilitates the estimation of the prior using normalizing flow. Specifically, the prior is represented through a Gaussian distribution combined with the Jacobian matrix of the transition module.

Formally presented, the transition prior module is represented as $\hat{\epsilon}_{it} = \hat{f}_i^{-1}(\hat{z}_{it}, \hat{\mathbf{z}}_{t-1:t-\tau})$. Subsequently, the joint distribution is decomposed as a product of the noise distribution and the determinant of the Jacobian matrix, formulated as $p([\hat{\mathbf{z}}_{t-1:t-\tau}, \hat{\mathbf{z}}_t]) = p([\hat{\mathbf{z}}_{t-1:t-\tau}, \hat{\epsilon}_t]) \times |\mathbf{J}|$, with $\mathbf{J} = \begin{bmatrix} \mathbb{I}_{n\tau} & \mathbf{0} \\ \mathbf{0} & diag(\frac{\partial \hat{\epsilon}_{it}}{\partial \hat{z}_{it}}) \end{bmatrix}$, where $[\cdot]$ denotes concatenation. Leveraging this joint distribution, we derive the prior as

$$\begin{aligned}
&\log p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1:t-\tau}) \\
&= \log p([\hat{\mathbf{z}}_t, \hat{\mathbf{z}}_{t-1:t-\tau}]) - \log p(\hat{\mathbf{z}}_{t-1:t-\tau}) \\
&= \log p([\hat{\epsilon}_t, \hat{\mathbf{z}}_{t-1:t-\tau}]) + \log |\mathbf{J}| - \log p(\hat{\mathbf{z}}_{t-1:t-\tau}) \\
&= \log p(\hat{\epsilon}_t|\hat{\mathbf{z}}_{t-1:t-\tau}) + \log |\mathbf{J}| \\
&= \log p(\hat{\epsilon}_t) + \log |\mathbf{J}| \\
&= \sum_i \log p(\hat{\epsilon}_{it}) + \log |\mathbf{J}| \quad : \text{Conditional independence} \\
&= \sum_i \left(\log p(\hat{\epsilon}_{it}) + \log \frac{\partial \hat{\epsilon}_{it}}{\partial \hat{z}_{t,i}} \right) : \text{Lower-triangular.}
\end{aligned} \tag{8}$$

The transition prior module can be efficiently executed using an MLP, transforming the latent variables $\hat{\mathbf{z}}_{t:t-\tau}$ into $\hat{\epsilon}_t$.

Optimization. We train **CaRiNG** using the Evidence Lower BOund (ELBO) objective, which is written as follows:

$$\begin{aligned}
&\text{ELBO} \triangleq \mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{X})}[\log p_\theta(\mathbf{X}|\mathbf{Z})] - D_{KL}(q_\phi(\mathbf{Z}|\mathbf{X})||p_\theta(\mathbf{Z})) \\
&= \underbrace{\mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{X})} \sum_{t=1}^{T} \log p_\theta(\mathbf{x}_t|\mathbf{z}_t)}_{-\mathcal{L}_{\text{Recon}}} \\
&+ \underbrace{\mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{X})} \left[\sum_{t=1}^{T} \log p_\theta(\mathbf{z}_t|\mathbf{z}_{t-1:t-\tau}) - \sum_{t=1}^{T} \log q_\phi(\mathbf{z}_t|\mathbf{x}_{t:t-\mu}) \right]}_{-\mathcal{L}_{\text{KLD}}}.
\end{aligned} \tag{9}$$

For the reconstruction likelihood $\mathcal{L}_{\text{Recon}}$, we utilize the mean-squared error (MSE) to measure the discrepancy be-

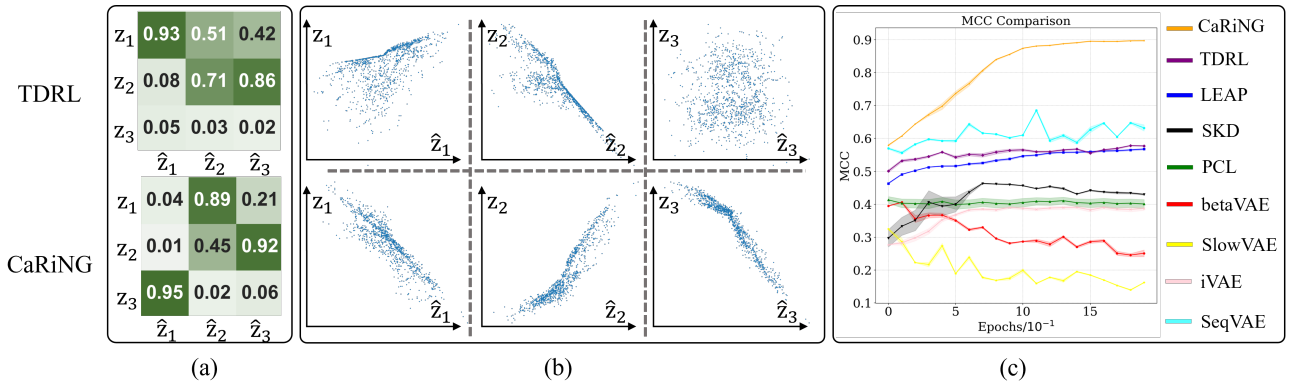


Figure 4: **Qualitative comparisons between baselines (especially TDRL) and CaRiNG in the setting of Non-invertible Generation.** (a) MCC matrix for all 3 latent variables; (b) The scatter plots between the estimated and ground-truth latent variables (only the aligned variables are plot); (c) The validation MCC curves of **CaRiNG** and other baselines.

tween the generated and original observations. When computing the KL divergence $\mathcal{L}_{\text{KLD}}$, we resort to a sampling method, given that the prior distribution lacks an explicit form. To elaborate, the posterior is produced by the encoder, while the prior is defined as in Eq 8.

5. Experiments

We conducted the experiments in two simulated environments, utilizing the available ground truth latent variables to evaluate identifiability. Subsequently, we assessed **CaRiNG** on a real-world VideoQA task, SUTD-TrafficQA (Xu et al., 2021), to verify its capability in representing complex and non-invertible traffic events.

5.1. Simulation Experiments

Dataset and experimental settings. To evaluate whether **CaRiNG** can learn the causal process and identify the latent variables under a non-invertible scenario, we design a series of simulation experiments based on a random causal structure with a given sample size and variable size. We provide two experimental settings, including NG and NG-TDMP, which simulate the scenarios in Theorem 1 with $r = 0$ (non-invertible generation) and $r > 0$ (time-delayed mixing process), respectively. In particular, for NG, we simulate the visual perception system that uses the ground-truth dimension as 3 to represent the 3D real world and apply 2 measured variables to represent the 2D observation, which indicates the generation is non-invertible. For NG-TDMP, we simulate the persistence of vision that involves the previous latent variables in the current mixing process. It denotes that even if the dimension of the observation is not reduced, the generation process is still non-invertible due to the time-delay mixing. More details of the data generation process can be found in Appendix A2.1.

Evaluation metrics. We apply the standard evaluation metric in the field of ICA, Mean Correlation Coefficient (MCC), to evaluate the identifiability of our **CaRiNG**. MCC measures the recovery of latent factors by calculating the absolute values of the correlation coefficient between every ground-truth factor against every estimated latent variable. It first calculates the Pearson correlation coefficients to measure the relationship and then adjusts the order with an assignment algorithm. The MCC score is a value from 0 to 1, where the higher score denotes better identifiability.

Baseline methods. We compare **CaRiNG** with a series of baseline methods. BetaVAE (Higgins et al., 2017) is the most basic baseline which ignores the temporal dependency and cannot utilize any auxiliary information. SlowVAE (Klindt et al., 2020), and PCL (Hyvarinen & Morioka, 2017) show the identifiability results but are limited by the assumption of independent sources. iVAE (Khemakhem et al., 2020) leverage nonstationarity (auxiliary information) to achieve identifiability. It is important to note that iVAE requires additional domain labels as input. In our experiments, we simply used time indices as the domain label. In addition, LEAP (Yao et al., 2022b) and TDRL (Yao et al., 2022a) allow for learning causal processes but assume an invertible generation process. Besides, we also compare **CaRiNG** with other temporal representation learning methods that are not based on ICA, such as Sequential VAE (Chung et al., 2015) and SKD (Berman et al., 2022), in which the disentangled representation has no identifiability guarantee.

Quantitative results. The performance of **CaRiNG** and other baseline methods in both the NG and NG-TDMP scenarios is presented in Table 1. Initially, it's evident that all baseline Nonlinear ICA methods yield unsatisfactory MCC scores in both scenarios, including the strong TDRL baseline, which previously obtained good results in invertible settings, as shown in Figure 4 (c). As shown in Figure 4 (a), TDRL cannot recover the lost latent variables caused by

Table 1: **MCC scores (with standard deviations over 4 seeds) of CaRiNG and baselines on NG and NG-TDMP settings.**

Setting	Method								
	CaRiNG	TDRL	LEAP	SlowVAE	PCL	betaVAE	SKD	iVAE	SequentialVAE
NG	**0.933 ±0.010**	0.627 ±0.009	0.651 ±0.019	0.362 ±0.041	0.507 ±0.091	0.551 ±0.007	0.489 ±0.077	0.391 ±0.686	0.750 ±0.035
NG-TDMP	**0.921 ±0.010**	0.837 ±0.068	0.704 ±0.005	0.398 ±0.037	0.489 ±0.095	0.437 ±0.021	0.381 ±0.084	0.553 ±0.097	0.847 ±0.019

Table 2: **Results on the SUTD-TrafficQA dataset.** The cross-modality matching parts of TDRL and **CaRiNG** are based on HCRN.

Method	Year	Accuracy(%)
I3D+LSTM	CVPR2017	33.21
HCRN	CVPR2020	36.26
VQAC	ICCV2021	36.00
MASN	ACL2021	36.03
DualVGR	TMM2021	36.07
Eclipse	CVPR2021	37.05
CMCIR	TPAMI2023	38.58
TDRL	NeurIPS2022	37.32
CaRiNG	-	**41.22**

non-invertible generation (MCC=0.03 for that variable). It is also illustrated by the scatter plots in Figure 4 (b), which show the independence between the estimated and ground truth variables on that dimension. Interestingly, we find that the Sequential VAE method works better than other methods that don't use the temporal context, which also demonstrates the necessity of temporal context to solve the invertibility issue. However, we still find that constraining the conditional independence benefits better performance, which shows the effect of the KL part. Furthermore, **CaRiNG** consistently delivers robust identifiability outcomes in both settings. This suggests that leveraging temporal context significantly enhances identifiability when faced with non-invertible generation processes. Lastly, performance in the NG scenario is better than that in the NG-TDMP scenario, showing the increased complexity introduced by the time-delayed mixing process.

5.2. Real-world Experiments

Dataset and experimental settings. The SUTD-TrafficQA dataset (Xu et al., 2021) is a comprehensive resource tailored for video event understanding in traffic scenarios, notably characterized by numerous occlusions among traffic agents. It consists of 10,090 videos and provides over 62,535 human-annotated QA pairs. Among them, 56,460 QA pairs are used for training and the rest 6,075 QA pairs are used for testing. The dataset challenges models with six reasoning tasks: "Basic Understanding" is designed for grasping traffic dynamics. "Event Forecasting" and "Reverse Reasoning" evaluate the temporal prediction ability. "Introspection", "Attribution", and "Counterfactual Inference" require the model to understand the causal dynamic and conduct reasoning. All tasks are formulated as multiple-choice forms (evaluation with accuracy) without limiting the number of candidate answers, and demand a deep comprehension of traffic events and their underlying causality.

Baseline methods. The primary method we benchmark against is TDRL (Yao et al., 2022a), to evaluate the representation ability of the complex and non-invertible traffic environment. Additionally, we evaluate **CaRiNG** in comparison with state-of-the-art VideoQA methods, including I3D+LSTM (Carreira & Zisserman, 2017), HCRN (Le et al., 2020), VQAC (Kim et al., 2021), MASN (Seo et al., 2021), DualVGR (Wang et al., 2021), Eclipse (Xu et al., 2021), and CMCIR (Liu et al., 2023). In our approach, **CaRiNG** is leveraged to identify latent causal dynamics, while HCRN serves as the basic model for question answering. Further implementation details are provided in the Appendix.

Quantitative results. Performance comparisons for the six question types on SUTD-TrafficQA are summarized in Table 2. **CaRiNG** achieves a score of 41.22, which demonstrates a significant improvement which is nearly 6.8% over the next best method. Notably, when compared to TDRL, which lacks temporal context, **CaRiNG** exhibits significant advancements in representing complex, non-invertible traffic events. When benchmarked against the HCRN baseline, which employs the same cross-modality matching module, our approach further escalates the score by 4.96 through causal representation learning. Though CMCIR (Liu et al., 2023) applies the Swin-Transformer-L (Liu et al., 2021) pretrained on ImageNet-22K dataset as the frame-level appearance extractor and employs the video Swin-B (Liu et al., 2022) pretrained on Kinetics-600 as the clip-level motion feature extractor (more powerful than ours), **CaRiNG** with sample ResNet101 (He et al., 2016) features still outperforms it with 2.64 in average. More analysis on TrafficQA and another evaluation on Volleyball (Ibrahim et al., 2016) can be found in Appendix A3 and A4, respectively.

6. Conclusion

In this paper, we have proposed to consider learning temporal causal representation under the non-invertible generation process, which is motivated by the common requirement of the temporal system, such as the visual perception process. We have established identifiability theories that allow for recovering the latent causal process with the nonlinear and

non-invertible mixing function. Furthermore, based on this theorem, we proposed our approach, **CaRiNG**, to leverage the temporal context to estimate the lost latent information. We have conducted a series of simulated experiments to verify the identifiability results of **CaRiNG** under the non-invertible generations and evaluated the learned representation in a complex and non-invertible traffic environment with real-world VideoQA tasks.

Impact Statement

This study introduces both a theoretical framework and a practical approach for extracting causal representations from time-series data. Such advancements enable the development of more transparent and interpretative models, enhancing our grasp of causal dynamics in real-world settings. This approach may benefit many real-world applications, including healthcare, auto-driving, and finance, but it could also be used illegally. For example, within the financial sphere, it can be harnessed to decipher ever-evolving market trends, optimizing predictions and thereby influencing investment and risk management decisions. However, it's imperative to note that any misjudgment of causal relationships could lead to detrimental consequences in these domains. Thus, establishing causal links must be executed with precision to prevent skewed or biased inferences.

Theoretically, though allowing for the non-invertible generation process, our theoretical assumptions still fall short of fully capturing the intricacies of real-world scenarios. For example, identifiability requires the absence of instantaneous causal relations, i.e., relying solely on time-delayed influences within the latent causal dynamics. Furthermore, we operate under the presumption that the number of variables remains consistent across different time steps, signifying that no agents enter or exit the environment. Moving forward, we aim to broaden our framework to ensure identifiability in more general settings, embracing instantaneous causal dynamics and the flexibility for variables to either enter or exit.

In our experiments, we evaluate our approach with both simulated and real-world datasets. However, our simulation relies predominantly on data points, creating a gap from real-world data. Concurrently, the real datasets lack the presence of ground truth latent variables. In the future, we plan to develop a benchmark specifically tailored for the causal representation learning task. This benchmark will harness the capabilities of game engines and renderers to produce videos embedded with ground-truth latent variables.

Acknowledgments

We would like to acknowledge the support from NSF Grant 2229881, the National Institutes of Health (NIH) under Contract R01HL159805, and grants from Apple Inc., KDDI Research Inc., Quris AI, and Florin Court Capital.

References

Berman, N., Naiman, I., and Azencot, O. Multifactor sequential disentanglement via structured koopman autoencoders. In *The Eleventh International Conference on Learning Representations*, 2022.

Berzuini, C., Dawid, P., and Bernardinell, L. *Causality: Statistical perspectives and applications*. John Wiley & Sons, 2012.

Cai, R. and Xie, F. Triad constraints for learning causal structure of latent variables. *Advances in neural information processing systems*, 2019.

Carreira, J. and Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Choi, M. J., Tan, V. Y., Anandkumar, A., and Willsky, A. S. Learning latent tree graphical models. *Journal of Machine Learning Research*, 12:1771–1812, 2011.

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C., and Bengio, Y. A recurrent latent variable model for sequential data. *Advances in neural information processing systems*, 28, 2015.

Coltheart, M. The persistences of vision. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 290(1038):57–69, 1980.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Drton, M., Lin, S., Weihs, L., and Zwiernik, P. Marginal likelihood and model selection for gaussian latent tree and forest models. 2017.

Fraccaro, M., Sønderby, S. K., Paquet, U., and Winther, O. Sequential neural models with stochastic layers, 2016.

Friston, K. Causal modelling and brain connectivity in functional magnetic resonance imaging. *PLoS biology*, 7 (2):e1000033, 2009.

Ghysels, E., Hill, J. B., and Motegi, K. Testing for granger causality with mixed frequency data. *Journal of Econometrics*, 192(1):207–230, 2016.

Hälvä, H. and Hyvarinen, A. Hidden markov nonlinear ica: Unsupervised learning from nonstationary time series. In *Conference on Uncertainty in Artificial Intelligence*, pp. 939–948. PMLR, 2020.

Hälvä, H., Corff, S. L., Lehéricy, L., So, J., Zhu, Y., Gassiat, E., and Hyvarinen, A. Disentangling identifiable features from noisy data with structured nonlinear ica. *arXiv preprint arXiv:2106.09620*, 2021.

Hartford, J., Ahuja, K., Bengio, Y., and Sridhar, D. Beyond the injective assumption in causal representation learning. 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Huang, B., Low, C. J. H., Xie, F., Glymour, C., and Zhang, K. Latent hierarchical causal structure discovery with rank constraints. *arXiv preprint arXiv:2210.01798*, 2022.

Hyvarinen, A. and Morioka, H. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in Neural Information Processing Systems*, 29: 3765–3773, 2016.

Hyvarinen, A. and Morioka, H. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pp. 460–469. PMLR, 2017.

Hyvärinen, A. and Oja, E. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5): 411–430, 2000.

Hyvarinen, A., Sasaki, H., and Turner, R. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 859–868. PMLR, 2019.

Ibrahim, M. S., Muralidharan, S., Deng, Z., Vahdat, A., and Mori, G. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1971–1980, 2016.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217. PMLR, 2020.

Kim, N., Ha, S. J., and Kang, J.-W. Video question answering using language-guided deep compressed-domain video feature. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1708–1717, 2021.

Klindt, D., Schott, L., Sharma, Y., Ustyuzhaninov, I., Brendel, W., Bethge, M., and Paiton, D. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*, 2020.

Kong, L., Xie, S., Yao, W., Zheng, Y., Chen, G., Stojanov, P., Akinwande, V., and Zhang, K. Partial disentanglement for domain adaptation. In *International conference on machine learning*, pp. 11455–11472. PMLR, 2022.

Kummerfeld, E. and Ramsey, J. Causal clustering for 1-factor measurement models. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1655–1664, 2016.

Lachapelle, S., Rodriguez, P., Sharma, Y., Everett, K. E., Le Priol, R., Lacoste, A., and Lacoste-Julien, S. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pp. 428–484. PMLR, 2022.

Lachapelle, S., Mahajan, D., Mitliagkas, I., and Lacoste-Julien, S. Additive decoders for latent variables identification and cartesian-product extrapolation, 2023.

Le, T. M., Le, V., Venkatesh, S., and Tran, T. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9972–9981, 2020.

Li, Y. and Mandt, S. Disentangled sequential autoencoder. *arXiv preprint arXiv:1803.02991*, 2018.

Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., and Gavves, E. icitris: Causal representation learning for instantaneous temporal effects. *arXiv preprint arXiv:2206.06169*, 2022a.

Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., and Gavves, S. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pp. 13557–13603. PMLR, 2022b.

Liu, Y., Li, G., and Lin, L. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the*

IEEE/CVF conference on computer vision and pattern recognition, pp. 12009–12019, 2022.

Palmer, S. E. *Vision science: Photons to phenomenology*. MIT press, 1999.

Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Seo, A., Kang, G.-C., Park, J., and Zhang, B.-T. Attend what you need: Motion-appearance synergistic networks for video question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 6167–6177, 2021.

Shimizu, S., Hoyer, P. O., and Hyvärinen, A. Estimation of linear non-gaussian acyclic models for latent factors. *Neurocomputing*, 72(7-9):2024–2027, 2009.

Silva, R., Scheines, R., Glymour, C., Spirtes, P., and Chickering, D. M. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7 (2), 2006.

Sorrenson, P., Rother, C., and Köthe, U. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). *arXiv preprint arXiv:2001.04872*, 2020.

Spearman, C. Pearson's contribution to the theory of two factors. *British Journal of Psychology*, 19(1):95, 1928.

Spelke, M. Principles of object perception, cognitive science 14. 1990.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, J., Bao, B., and Xu, C. Dualvgr: A dual-visual graph reasoning unit for video question answering. *IEEE Transactions on Multimedia*, 2021.

Werbos, P. Beyond regression: New tools for prediction and analysis in the behavioral sciences. *PhD thesis, Committee on Applied Mathematics, Harvard University, Cambridge, MA*, 1974.

Wertheimer, M. Laws of organization in perceptual forms. 1938.

Wolford, G. A model of visible persistence based on linear systems. *Canadian Psychology/Psychologie canadienne*, 34(2):162, 1993.

Xie, F., Cai, R., Huang, B., Glymour, C., Hao, Z., and Zhang, K. Generalized independent noise condition for estimating latent variable causal graphs. *arXiv preprint arXiv:2010.04917*, 2020.

Xie, F., Huang, B., Chen, Z., He, Y., Geng, Z., and Zhang, K. Identification of linear non-gaussian latent hierarchical structure. In *International Conference on Machine Learning*, pp. 24370–24387. PMLR, 2022.

Xu, L., Huang, H., and Liu, J. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *CVPR*, pp. 9878–9888, 2021.

Yao, W., Chen, G., and Zhang, K. Temporally disentangled representation learning. In *Advances in Neural Information Processing Systems*, 2022a. URL https://openreview.net/forum?id=Vi-sZWNA_Ue.

Yao, W., Sun, Y., Ho, A., Sun, C., and Zhang, K. Learning temporally causal latent processes from general temporal data. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=RDlLMjLJXdq.

Zhang, N. L. Hierarchical latent class models for cluster analysis. *The Journal of Machine Learning Research*, 5: 697–723, 2004.

Zhou, H., Kadav, A., Shamsian, A., Geng, S., Lai, F., Zhao, L., Liu, T., Kapadia, M., and Graf, H. P. Composer: compositional reasoning of group activity in videos with keypoint-only modality. In *European Conference on Computer Vision*, pp. 249–266. Springer, 2022.

Ziegler, Z. and Rush, A. Latent normalizing flows for discrete sequences. In *International Conference on Machine Learning*, pp. 7673–7682. PMLR, 2019.

Appendix for

"Learning Temporal Causal Representation under Non-Invertible Generation Process"

A1. Identifiability Theory

A1.1. Proof for Theorem 1

Let us first shed light on the identifiability theory on the special case with $\tau = r + 1$, i.e.,

$$\mathbf{x}_t = \mathbf{g}(\mathbf{z}_{t:t-r}), \quad z_{it} = f_i\left(\mathbf{z}_{t-1:t-r-1}, \epsilon_{it}\right), \quad \mathbf{z}_t = \mathbf{m}(\mathbf{x}_{t:t-\mu}). \tag{A1}$$

Theorem A1 (Identifiability under Non-invertible Generative Process)**.** *For a series of observations $\mathbf{x}_t \in \mathbb{R}^d$ and estimated latent variables $\hat{\mathbf{z}}_t \in \mathbb{R}^n$, suppose there exists function $\hat{\mathbf{g}}, \hat{\mathbf{m}}$ which is subject to observational equivalence,*

$$\mathbf{x}_t = \hat{\mathbf{g}}(\hat{\mathbf{z}}_{t:t-r}), \quad \hat{\mathbf{z}}_t = \hat{\mathbf{m}}(\mathbf{x}_{t:t-\mu}). \tag{A2}$$

If assumptions

- *{(Smooth and Positive Density) the probability density function of latent variables is third-order differentiable and positive in R^n,*
- *(conditional independence) the components of $\hat{\mathbf{z}}_t$ are mutually independent conditional on $\hat{\mathbf{z}}_{t-1:t-r-1}$,*
- *(sufficiency) let $\eta_{kt} \triangleq \log p(z_{kt}|\mathbf{z}_{t-1:t-r-1})$, and*

$$\mathbf{v}_{lt} \triangleq \Big(\frac{\partial^2 \eta_{1t}}{\partial z_{1t} \partial z_{l,t-r-1}}, ..., \frac{\partial^2 \eta_{nt}}{\partial z_{nt} \partial z_{l,t-r-1}}, \frac{\partial^3 \eta_{1t}}{\partial z_{1t}^2 \partial z_{l,t-r-1}}, ..., \frac{\partial^3 \eta_{nt}}{\partial z_{nt}^2 \partial z_{l,t-r-1}}\Big)^\intercal, \tag{A3}$$

  for $l = 1, 2, \cdots, n$. For each value of $\mathbf{z}_t$, there exists $2n$ different values of $z_{l,t-r-1}$ such that the $2n$ vector functions $\mathbf{v}_{lt} \in \mathbf{R}^{2n}$ are linearly independent,

are satisfied, then $\mathbf{z}_t$ must be a component-wise transformation of a permuted version of $\hat{\mathbf{z}}_t$ with regard to context $\{\mathbf{x}_j \mid \forall j = t, t-1, \cdots, t-\mu-r\}$.

Proof. For any t, combining Eq A1 and Eq A2 gives

$$\begin{aligned}\mathbf{z}_t &= \mathbf{m}(\mathbf{x}_{t:t-\mu}) \\ &= \mathbf{m}(\hat{\mathbf{g}}(\hat{\mathbf{z}}_t, \hat{\mathbf{z}}_{t-1:t-r}), \mathbf{x}_{t-1:t-\mu}) \\ &= \mathbf{m}(\hat{\mathbf{g}}(\hat{\mathbf{z}}_t, \hat{\mathbf{m}}(\mathbf{x}_{t-1:t-\mu-1}), \cdots, \hat{\mathbf{m}}(\mathbf{x}_{t-r:t-\mu-r})), \mathbf{x}_{t-1:t-\mu}),\end{aligned} \tag{A4}$$

as well as $\hat{\mathbf{z}}_t = \hat{\mathbf{m}}(\mathbf{g}(\mathbf{z}_t, \mathbf{m}(\mathbf{x}_{t-1:t-\mu-1}), \cdots, \mathbf{m}(\mathbf{x}_{t-r:t-\mu-r})), \mathbf{x}_{t-1:t-\mu})$ similarly. Upon Eq A4, we have an unified partially invertible function $\mathbf{z}_t = \mathbf{h}(\hat{\mathbf{z}}_t|\mathbf{x}_{t-1:t-\mu-r})$ where $\mathbf{h} = \mathbf{m} \circ \hat{\mathbf{g}}$ with Jacobian $\frac{\partial \mathbf{z}_t}{\partial \hat{\mathbf{z}}_t} = \mathbf{H}_t(\hat{\mathbf{z}}_t; \mathbf{x}_{t-1:t-\mu-r})$. By *partially invertible* it means that $\mathbf{z}$ and $\hat{\mathbf{z}}$ are in one-to-one correspondence for any context observations $\mathbf{x}_{t-1:t-\mu-r}$ that are fixed. One more thing to notify is that since $\mathbf{g}, \hat{\mathbf{g}}, \mathbf{m}, \hat{\mathbf{m}}$ are second-order differentiable, the nested $\mathbf{h}$ is also second-order differentiable. Let us consider the mapping from joint distribution $(\hat{\mathbf{z}}_t, \mathbf{x}_{t-1:t-\mu-r-1})$ to $(\mathbf{z}_t, \mathbf{x}_{t-1:t-\mu-r-1})$, i.e.,

$$P(\mathbf{z}_t, \mathbf{x}_{t-1:t-\mu-r-1}) = P(\hat{\mathbf{z}}_t, \mathbf{x}_{t-1:t-\mu-r-1}) \,/\, |\mathbf{J}_t|, \tag{A5}$$

where

$$\mathbf{J}_t = \begin{bmatrix} \frac{\partial \mathbf{z}_t}{\partial \hat{\mathbf{z}}_t} & \mathbf{0} \\ * & \mathbf{I} \end{bmatrix}, \tag{A6}$$

which is a lower triangle matrix, where $\mathbf{I}$ infers eye matrix and $*$ infers any possible matrix. Thus, we have determinant $|\mathbf{J}_t| = |\frac{\partial \mathbf{z}_t}{\partial \hat{\mathbf{z}}_t}| = |\mathbf{H}_t|$. Dividing both sides of Eq A5 by $P(\mathbf{x}_{t-1:t-\mu-r-1})$ gives

$$\mathbf{LHS} = P(\mathbf{z}_t|\mathbf{x}_{t-1:t-\mu-r-1}) = P(\mathbf{z}_t|\mathbf{z}_{t-1:t-r-1}), \tag{A7}$$

since $\mathbf{z}_t$ and $\mathbf{x}_{t-1:t-\mu-r-1}$ are independent conditioned on $\mathbf{z}_{t-1:t-r-1}$. Similarly, **RHS** $= P(\hat{\mathbf{z}}_t|\mathbf{x}_{t-1:t-\mu-r-1}) = P(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-r-1})$ holds true as well, which yields to

$$P(\mathbf{z}_t|\mathbf{z}_{t-1:t-r-1}) = P(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1:t-r-1}) \,/\, |\mathbf{H}_t|. \tag{A8}$$

From a direct observation, if the components of $\hat{\mathbf{z}}_t$ are mutually independent given $\hat{\mathbf{z}}_{t-1:t-r-1}$, then for any distinct $i \neq j$, $\hat{z}_{it}$ and $\hat{z}_{jt}$ are conditionally independent given $(\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}, \hat{z}_{jt}\}) \cup \hat{\mathbf{z}}_{t-1:t-r-1}$. This mutual independence of the components of $\hat{\mathbf{z}}_t$ based on $\hat{\mathbf{z}}_{t-1:t-r-1}$ implies two things:

- $\hat{z}_{it}$ is independent from $\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}, \hat{z}_{jt}\}$ conditional on $\hat{\mathbf{z}}_{t-1:t-r-1}$. Formally,
$$p(\hat{z}_{it} \,|\, \hat{\mathbf{z}}_{t-1:t-r-1}) = p(\hat{z}_{it} \,|\, (\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}, \hat{z}_{jt}\}) \cup \hat{\mathbf{z}}_{t-1:t-r-1}).$$
- $\hat{z}_{it}$ is independent from $\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}\}$ conditional on $\hat{\mathbf{z}}_{t-1:t-r-1}$. Represented as:
$$p(\hat{z}_{it} \,|\, \hat{\mathbf{z}}_{t-1:t-r-1}) = p(\hat{z}_{it} \,|\, (\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}\}) \cup \hat{\mathbf{z}}_{t-1:t-r-1}).$$

From these two equations, we can derive:

$$p(\hat{z}_{it} \,|\, (\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}\}) \cup \hat{\mathbf{z}}_{t-1:t-r-1}) = p(\hat{z}_{it} \,|\, (\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}, \hat{z}_{jt}\}) \cup \hat{\mathbf{z}}_{t-1:t-r-1}),$$

which yields that $\hat{z}_{it}$ and $\hat{z}_{jt}$ are conditionally independent given $(\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}, \hat{z}_{jt}\}) \cup \hat{\mathbf{z}}_{t-1:t-r-1}$ for $i \neq j$. Leveraging an inherent fact, i.e., if $\hat{z}_{it}$ and $\hat{z}_{jt}$ are conditionally independent given $(\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}, \hat{z}_{jt}\}) \cup \hat{\mathbf{z}}_{t-1:t-r-1}$, the subsequent equation arises:

$$\frac{\partial^2 \log p(\hat{\mathbf{z}}_t, \hat{\mathbf{z}}_{t-1:t-r-1})}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} = 0,$$

assuming the cross second-order derivative exists.

Given that $p(\hat{\mathbf{z}}_t, \hat{\mathbf{z}}_{t-1:t-r-1}) = p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1:t-r-1}) p(\hat{\mathbf{z}}_{t-1:t-r-1})$ and $p(\hat{\mathbf{z}}_{t-1:t-r-1})$ remains independent of $\hat{z}_{it}$ or $\hat{z}_{jt}$, the above equality is equivalent to

$$\frac{\partial^2 \log p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1:t-r-1})}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} = 0. \tag{A9}$$

Referencing Eq A8, it gets expressed as:

$$\log p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1:t-r-1}) = \log p(\mathbf{z}_t \,|\, \mathbf{z}_{t-1:t-r-1}) + \log |\mathbf{H}_t| = \sum_{k=1}^{n} \eta_{kt} + \log |\mathbf{H}_t|. \tag{A10}$$

The partial derivative w.r.t. $\hat{z}_{it}$ is presented below:

$$\begin{aligned} \frac{\partial \log p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1:t-r-1})}{\partial \hat{z}_{it}} &= \sum_{k=1}^{n} \frac{\partial \eta_{kt}}{\partial z_{kt}} \cdot \frac{\partial z_{kt}}{\partial \hat{z}_{it}} + \frac{\partial \log |\mathbf{H}_t|}{\partial \hat{z}_{it}} \\ &= \sum_{k=1}^{n} \frac{\partial \eta_{kt}}{\partial z_{kt}} \cdot \mathbf{H}_{kit} + \frac{\partial \log |\mathbf{H}_t|}{\partial \hat{z}_{it}}. \end{aligned}$$

The second-order cross derivative can be depicted as:

$$\frac{\partial^2 \log p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1:t-r-1})}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} = \sum_{k=1}^{n} \Big(\frac{\partial^2 \eta_{kt}}{\partial z_{kt}^2} \cdot \mathbf{H}_{kit} \mathbf{H}_{kjt} + \frac{\partial \eta_{kt}}{\partial z_{kt}} \cdot \frac{\partial \mathbf{H}_{kit}}{\partial \hat{z}_{jt}} \Big) + \frac{\partial^2 \log |\mathbf{H}_t|}{\partial \hat{z}_{it} \partial \hat{z}_{jt}}. \tag{A11}$$

According to Eq A9, the right-hand side of the presented equation consistently equals 0. Therefore, for each index l ranging from 1 to n, and every associated value of $z_{l,t-r-1}$, its partial derivative with respect to $z_{l,t-r-1}$ remains 0. That is,

$$\sum_{k=1}^{n} \Big(\frac{\partial^3 \eta_{kt}}{\partial z_{kt}^2 \partial z_{l,t-r-1}} \cdot \mathbf{H}_{kit} \mathbf{H}_{kjt} + \frac{\partial^2 \eta_{kt}}{\partial z_{kt} \partial z_{l,t-r-1}} \cdot \frac{\partial \mathbf{H}_{kit}}{\partial \hat{z}_{jt}} \Big) \equiv 0, \tag{A12}$$

where we leveraged the fact that entries of $\mathbf{H}_t$ do not depend on $z_{l,t-r-1}$. Considering any given value of $\mathbf{z}_t$, there exists at least $2n$ different values of $\mathbf{v}_{lt}$ such that they are linearly independent. To make the above equation hold true, one has to set $\mathbf{H}_{kit}\mathbf{H}_{kjt} = 0$ or $i \neq j$. In other words, each row of $\mathbf{H}_t$ consists of at most a single non-zero entry, and $\mathbf{z}_t$ must be a component-wise transformation of a permuted version of $\hat{\mathbf{z}}_t$. □

Note that in the proof of Theorem A1, we require the transition lag τ to be larger than the mixing lag $r = 1$. When a mixing lag exists, the guarantee of identifiability requires dynamic information from a more previous time step. As long as this inequality $\tau > r$ is satisfied, the parameters τ can be extended to arbitrary numbers following a similar modification in Appendix A1.4.

A1.2. Discussion for the sufficiency assumption.

This assumption describes the changability of latent variables. Taking the video understanding as an example, the latent variables may represent the concepts. The linear independence of the latent variables means that there exists a characteristic of the concept that cannot be linearly represented by others. To further illustrate the sufficiency assumption, we give 2 examples (Yao et al., 2022a) to show when and when not the sufficiency assumption holds.

One possible distribution that breaks this assumption is the additive Gaussian noise. Denote $\mathbf{z}_h$ as historical parents. Let $z_{kt} = q_k(\mathbf{z}_h) + \epsilon_{kt}$ where $\epsilon_{kt} \sim N(0, 1)$. In this case, we have $\eta_{kt} = \log P(z_{kt}|\mathbf{z}_h) = -\log\sqrt{2\pi} - \frac{(z_{kt}-q_k(\mathbf{z}_h))^2}{2}$, and $\frac{\partial^2 \log P(z_{kt}|\mathbf{z}_h)}{\partial^2 z_{kt}} = 0$, which will violate the assumption.

On the opposite, if ϵ_{kt} subjects a zero mean generalized normal distribution: $P(\epsilon_{kt}) \propto e^{-\lambda|\epsilon_{kt}|^\beta}$ with $\lambda > 0$ and $\beta > 2$ and $\beta \neq 3$. Let $z_{kt} = q_k(\mathbf{z}_h) + \epsilon_{kt}$ in which q is a linear function. If for each z_{kt} there exists at least one k' such that $c_{kk'} = \frac{\partial z_{kt}}{\partial z_{k',t-1}} \neq 0$, the sufficiency assumption must hold.

In this case, we have

$$\frac{\partial^3 \eta_{kt}}{\partial^2 z_{kt} \partial z_{k',t-1}} = -\lambda \operatorname{sgn}(\epsilon_{kt})\beta(\beta-1)(\beta-2)|\epsilon_{kt}|^{\beta-3} c_{kk'} \tag{A13}$$

and

$$\frac{\partial^2 \eta_{kt}}{\partial z_{kt} \partial z_{k',t-1}} = -\lambda\beta(\beta-1)|\epsilon_{kt}|^{\beta-2} c_{kk'}. \tag{A14}$$

We know that $|\epsilon_{lt}|^{\beta-2}$ and $|\epsilon_{lt}|^{\beta-2}$ are linearly independent since their ratio $|\epsilon_{lt}|$ is not constant. Besides, $|\epsilon_{lt}|^{\beta-2}$ and $|\epsilon_{lt}|^{\beta-2}$, with $l = 1, 2, \cdots, n$ are $2n$ linearly independent functions because of the different arguments involved. Suppose there exists α_{l1}, α_{l2} for $l = 1, 2, \cdots, n$, such that the weighted sum with regard to $\mathbf{v}_{l,t}$ is zero. Thus, for any k we have

$$\alpha_{k1} c_{kk'} |\epsilon_{kt}|^{\beta-2} + \alpha_{k2} c_{kk'} |\epsilon_{kt}|^{\beta-3} + \sum_{l \neq k} (\alpha_{l1} c_{lk'} |\epsilon_{lt}|^{\beta-2} + \alpha_{l2} c_{lk'} |\epsilon_{lt}|^{\beta-3}) = 0. \tag{A15}$$

Since $|\epsilon_{kt}|^{\beta-2}$ and $|\epsilon_{kt}|^{\beta-3}$ with $l = 1, 2, \cdots, n$ are linearly independent and $c_{kk'} \neq 0$, the make the above equation holds, we have $\alpha_{k1} = \alpha_{k2} = 0$. As this applies to any k, we know that α_{l1} and α_{l2} must be 0, for all $l = 1, 2, \cdots, n$. That is, $\{\mathbf{v}_{lt}\}$ is linearly independent. Thus, the sufficiency assumption holds.

Please note that the sufficiency assumption is crucial to the identifiability theory, yet not that restrictive. Even if it is not completely satisfied, we can still obtain some subspace identifiability (Kong et al., 2022).

A1.3. Discussion for the cross-time disentanglement

This section demonstrates how the entanglements between variables across time steps are prevented. Generally speaking, if the information is lost in the transition from $\mathbf{z}_t$ to $\mathbf{x}_t$, we have to borrow the information from context such as $\mathbf{z}_{t-1}$ to recover it. It is natural to receive information from $\mathbf{z}_{t-1}$ in order to find the best estimator.

Specifically, let us consider a generating process $x_{it} = g_i(\mathbf{z}_t), z_{it} = f_i(\mathbf{z}_{t-1}, \epsilon_{it})$, where $\mathbf{x}_t \in R^d, \mathbf{z}_t \in R^n$. Since x_{it} can be fully charactized by $\mathbf{z}_t$, but not a function of $\mathbf{z}_{t-1}$, we have $\frac{\partial x_{i,t}}{\partial \mathbf{z}_{t-1}} = 0$. For the estimation process $x_{i,t} = \hat{g}_i(\hat{\mathbf{z}}_t)$, we have $\frac{\partial x_{i,t}}{\partial \mathbf{z}_{t-1}} = \frac{\partial x_{i,t}}{\partial \hat{\mathbf{z}}_t} \cdot \frac{\partial \hat{\mathbf{z}}_t}{\partial \mathbf{z}_{t-1}} = 0$ for $i = 1, 2, \cdots, d$. Formally, we have an equation as $\frac{\partial \mathbf{x}_t}{\partial \hat{\mathbf{z}}_t} \cdot \frac{\partial \hat{\mathbf{z}}_t}{\partial \mathbf{z}_{t-1}} = 0$, i.e.,

$$\begin{bmatrix} \frac{\partial x_{1,t}}{\partial \hat{z}_{1t}} & \cdots & \frac{\partial x_{1,t}}{\partial \hat{z}_{nt}} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_{d,t}}{\partial \hat{z}_{1t}} & \cdots & \frac{\partial x_{d,t}}{\partial \hat{z}_{nt}} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial \hat{z}_{1t}}{\partial z_{1,t-1}} & \cdots & \frac{\partial \hat{z}_{1t}}{\partial z_{n,t-1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \hat{z}_{nt}}{\partial z_{1,t-1}} & \cdots & \frac{\partial \hat{z}_{nt}}{\partial z_{n,t-1}} \end{bmatrix} = \mathbf{0}. \tag{A16}$$

If there exists at least n different i such that derivative of x_{it} with respect to $\hat{\mathbf{z}}_t$ as n vector functions $\begin{bmatrix} \frac{\partial x_{i,t}}{\partial \hat{z}_{1t}} & \cdots & \frac{\partial x_{i,t}}{\partial \hat{z}_{nt}} \end{bmatrix}$ are linearly independent, we have

$$\begin{bmatrix} \frac{\partial \hat{z}_{1t}}{\partial z_{1,t-1}} & \cdots & \frac{\partial \hat{z}_{1t}}{\partial z_{n,t-1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \hat{z}_{nt}}{\partial z_{1,t-1}} & \cdots & \frac{\partial \hat{z}_{nt}}{\partial z_{n,t-1}} \end{bmatrix} = \mathbf{0} \tag{A17}$$

holds true. If the rank $R(\frac{\partial \mathbf{x}_t}{\partial \hat{\mathbf{z}}_t})$ of the matrix of $\frac{\partial \mathbf{x}_t}{\partial \hat{\mathbf{z}}_t}$ is less than n, we have that the entanglement can only happened on the rest $n - R(\frac{\partial \mathbf{x}_t}{\partial \hat{\mathbf{z}}_t})$ dimensions.

Let us consider 2 extreme cases. First, if mixing function $\mathbf{g}$ is invertible, we have $R(\frac{\partial \mathbf{x}_t}{\partial \hat{\mathbf{z}}_t}) = n$, and the entanglement between time steps are prevented. As another extreme case, in the NG setting as mentioned in Appendix A2.1, we set one dimension of the latent variable that is totally lost during the mixing process. In this case, we have to use $\mathbf{z}_{t-1}$ for its estimation. Even in this case, the unnecessary entanglements are prevented as well.

A1.4. Extension to Multiple Lags

Multiple Transition Time Lag τ. For the sake of simplicity, we consider only one special case with $\tau = r + 1$ in Theorem A1. Our identifiability theorem can be actually extended to arbitrary lags directly. For any given τ, according to modularity, we have different conclusion at Eq A7 as **LHS** $= P(\mathbf{z}_t|\mathbf{x}_{t-1:t-\mu-r-\tau}) = P(\mathbf{z}_t|\mathbf{z}_{t-1:t-r-\tau})$. Similarity **RHS** $= P(\hat{\mathbf{z}}_t|\mathbf{x}_{t-1:t-\mu-r-\tau}) = P(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1:t-r-\tau})$ holds true as well. In addition, some modifications are needed in sufficiency assumption, i.e., re-define $\eta_{kt} \triangleq \log p(z_{kt}|\mathbf{z}_{t-1:t-r-\tau})$ and there should be at least $2n$ linear independent vectors for $\mathbf{v}$ with regard to $z_{lt'}$ where $l = 1, 2, \cdots, n$ and $t - \tau \leq t' \leq t - r - 1$. No extra changes are needed.

Infinite Mixing Lag r. Theorem A1 can also be easily extended to infinite mixing lag since $\hat{\mathbf{z}}_t = \mathbf{h}(\mathbf{z}_t; \mathbf{x}_{<t})$ still exists when $r \to \infty$, and the theorem still holds true.

A1.5. Continuity for Permutation Invariance

Let us first give an extreme example to illustrate the importance of extra constraints for identifiability when the probability density of $\mathbf{z}_t$ is not non-zero everywhere in $\mathbb{R}^n$. Consider 4 independent random variables u, v, x, y subject to standard normal distribution respectively. Suppose that there exists an invertible function $(x, y) = \mathbf{h}(u, v)$ satisfies

$$\begin{cases} x = \mathbb{I}(x + y > 0) \cdot u + \mathbb{I}(x + y \leq 0) \cdot v \\ y = \mathbb{I}(x + y > 0) \cdot v + \mathbb{I}(x + y \leq 0) \cdot u. \end{cases} \tag{A18}$$

Notice that the Jacobian from (u, v) to (x, y) contains at most one non-zero entry for each column or row. However, the result (x, y) is still entangled, and the identifiability of (u, v) is not achieved. What if now we notate latent variable as $\hat{\mathbf{z}} = (u, v)$, estimated latent variable as $\mathbf{z} = (x, y)$ and the transition process with two mixing functions as $\mathbf{h} = \mathbf{g}^{-1} \circ \hat{\mathbf{g}}$?

In the literature of nonlinear ICA, the gap between $\mathbf{H}_{ij} \cdot \mathbf{H}_{ik} = 0$ when $j \neq k$ and identifiability is ill-discussed. In linear ICA, since the Jacobian is a constant matrix, these two statements are equivalent. Nevertheless, in nonlinear ICA, $\mathbf{H} = \frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}$ is not a constant, but a function of $\hat{\mathbf{z}}$, which may leads to the failure of identifiability as shown in Eq A18.

The counterexamples can still be easily constructed even if function $\mathbf{h}$ is continuous. For brevity, let us denote a segment-wise linear indicator function as $f(u, v) = \min(\max(0, u + v + 0.5), 1)$, and we have $\mathbf{h}$ as

$$\begin{cases} x = f(u, v) \cdot u + (1 - f(u, v)) \cdot v \\ y = f(u, v) \cdot v + (1 - f(u, v)) \cdot u. \end{cases} \tag{A19}$$

When u, v, x, y are independent uniform distributions on $[-2, -1] \cup [1, 2]$, all conditions are still satisfied while the identifiability cannot be achieved.

To fill this gap, we provide two more assumptions. The domain $\hat{\mathcal{Z}}$ of $\hat{\mathbf{z}}$ should be path-connected, i.e., for any $\hat{\mathbf{z}}^{(1)}, \hat{\mathbf{z}}^{(2)} \in \hat{\mathcal{Z}}$, there exists a continuous path connecting $\hat{\mathbf{z}}^{(1)}$ and $\hat{\mathbf{z}}^{(2)}$ with all points of the path in $\hat{\mathcal{Z}}$. In addition, the derivative of function $\mathbf{h}$ is not zero for any value of $\hat{\mathbf{z}} \in \hat{\mathcal{Z}}$

Lemma A1 (Disentanglement with Continuity)**.** *For second-order differentiable invertible function* $\mathbf{h}$ *defined on a path-connected domain* $\hat{\mathcal{Z}} \subseteq \mathbb{R}^n$ *which satisfies* $\mathbf{z} = \mathbf{h}(\hat{\mathbf{z}})$*, suppose the non-degeneracy condition holds. If there exists at most one non-zero entry in each row of the Jacobian matrix* $\mathbf{H} = \frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}$*, the identifiability under Permutation Invariance can be established.*

Proof. For any row i, $\frac{\partial \mathbf{z}_i}{\partial \hat{\mathbf{z}}} = [\frac{\partial \mathbf{z}_i}{\partial \hat{\mathbf{z}}_1}, \frac{\partial \mathbf{z}_i}{\partial \hat{\mathbf{z}}_2}, ..., \frac{\partial \mathbf{z}_i}{\partial \hat{\mathbf{z}}_n}] \in \mathbb{R}^n$ is a n-dimensional variable. Its image is a subspace as $\bigcup_{k=1}^{n} \Big\{ (\frac{\partial \mathbf{z}_i}{\partial \hat{\mathbf{z}}_1}, \frac{\partial \mathbf{z}_i}{\partial \hat{\mathbf{z}}_2}, ..., \frac{\partial \mathbf{z}_i}{\partial \hat{\mathbf{z}}_n}) \in \mathbb{R}^n : \frac{\partial \mathbf{z}_i}{\partial \hat{\mathbf{z}}_j} = 0 \text{ for all } j \neq k, \text{ and } x_k \neq 0 \Big\}$, since there exists at most one non-zero entry in each row of the Jacobian matrix $\mathbf{H} = \frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}$ and the derivative of function $\mathbf{h}$ is not zero for any value, according to the non-degeneracy condition.

We use proof by contradiction. Suppose there exist two different samples $\mathbf{a}, \mathbf{b} \in \mathcal{Z} \subseteq \mathbb{R}^n$ with different non-zero entries $j \neq k$ subjects to

$$\left[\frac{\partial z_i}{\partial \hat{\mathbf{z}}} \bigg|_{\hat{\mathbf{z}}=\mathbf{a}} \right]_j \neq 0, \quad \left[\frac{\partial z_i}{\partial \hat{\mathbf{z}}} \bigg|_{\hat{\mathbf{z}}=\mathbf{b}} \right]_k \neq 0 \tag{A20}$$

where $[\cdot]_j$ refers to the j-th entry of vector. Their values are respectively within $\Big\{ (0, 0, ..., \frac{\partial z_i}{\partial \hat{z}_j}, 0, ..., 0) \in \mathbb{R}^n : \frac{\partial z_i}{\partial \hat{z}_j} \neq 0 \Big\}$ and $\Big\{ (0, 0, ..., \frac{\partial z_i}{\partial \hat{z}_k}, 0, ..., 0) \in \mathbb{R}^n : \frac{\partial z_i}{\partial \hat{z}_k} \neq 0 \Big\}$. Clearly, there is no path from $\frac{\partial z_i}{\partial \hat{\mathbf{z}}}\big|_{\hat{\mathbf{z}}=\mathbf{a}}$ to $\frac{\partial z_i}{\partial \hat{\mathbf{z}}}\big|_{\hat{\mathbf{z}}=\mathbf{b}}$. Since $\mathbf{h}$ is a second-order differentiable invertible function, we have its derivative $\mathbf{h}'$ is also differentiable. Thus, $\hat{\mathcal{Z}} \subseteq \mathbb{R}^n$ is a path-connected domain which denotes that the image of $\frac{\partial z_i}{\partial \hat{\mathbf{z}}}$ is also path-connected. It will be violated that there is no path from $\frac{\partial z_i}{\partial \hat{\mathbf{z}}}\big|_{\hat{\mathbf{z}}=\mathbf{a}}$ to $\frac{\partial z_i}{\partial \hat{\mathbf{z}}}\big|_{\hat{\mathbf{z}}=\mathbf{b}}$ thus the proof is established.

□

When it comes to partially invertible function with regard to side information $\mathbf{c}$, the proof is the same with only a modification on conditions. That is, the path-connected domain assumption is applied to $(\mathbf{z}, \mathbf{c})$, and the infinite differentiability is extended to both $\mathbf{z}$ and $\mathbf{c}$, i.e., $\frac{\partial^2 z_i}{\partial a \partial b}$ for $a, b \in \{z|\mathbf{z}_i\} \times \{c|\mathbf{c}_i\}$ when $a \neq b$ exists.

Let's further review the example we provided earlier. Examples in Eq A18 and Eq A19 respectively demonstrate the scenarios where the assumptions of differentiability and connectivity fail, leading to the breakdown of identifiability.

Lemma A2 (Disentanglement with Continuity under Side Information)**.** *For second-order differentiable invertible function* $\mathbf{h}$ *defined on a path-connected domain* $\hat{\mathcal{Z}} \times \mathcal{C} \subseteq \mathbb{R}^{n+m}$ *which satisfies* $\mathbf{z} = \mathbf{h}(\hat{\mathbf{z}}, \mathbf{c})$*, suppose the non-degeneracy condition holds. If there exists at most one non-zero entry in each row of the Jacobian matrix* $\mathbf{H}(\mathbf{c}) = \frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}$*, the identifiability under Permutation Invariance can be established.*

Proof. Suppose there exist two different samples $\mathbf{a}, \mathbf{b} \in \hat{\mathcal{Z}} \times \mathcal{C} \subseteq \mathbb{R}^n$ with different non-zero entries $j \neq k$ subjects to

$$\left[\frac{\partial z_i}{\partial (\hat{\mathbf{z}}, \mathbf{c})} \bigg|_{(\hat{\mathbf{z}}, \mathbf{c})=\mathbf{a}} \right]_j \neq 0, \quad \left[\frac{\partial z_i}{\partial (\hat{\mathbf{z}}, \mathbf{c})} \bigg|_{(\hat{\mathbf{z}}, \mathbf{c})=\mathbf{b}} \right]_k \neq 0. \tag{A21}$$

Similar to Lemma A1, there exists no path between them because they are blocked in $\hat{\mathcal{Z}}$ alone. In the same way, since $\mathbf{h}$ is a second-order differentiable invertible function, and the non-degeneracy condition holds, the image of $\frac{\partial z_i}{\partial (\hat{\mathbf{z}}, \mathbf{c})}$ is also path-connected. It will be violated and the proof is established.

□

A1.6. Identifiability Benefits from Non-Stationarity

We can further leverage the advantage of non-stationary data for identifiability. We rewrite $\mathbf{v}_{lt}$, which is defined in Eq A3, as $\mathbf{s}_{lt}(u_r)$ in the u_r context as

$$\mathbf{s}_{lt}(u_r) \triangleq \Big(\frac{\partial^2 \eta_{1t}(u_r)}{\partial z_{1t} \partial z_{l,t-r-1}}, ..., \frac{\partial^2 \eta_{nt}(u_r)}{\partial z_{nt} \partial z_{l,t-r-1}}, \frac{\partial^3 \eta_{1t}(u_r)}{\partial z_{1t}^2 \partial z_{l,t-r-1}}, ..., \frac{\partial^3 \eta_{nt}(u_r)}{\partial z_{nt}^2 \partial z_{l,t-r-1}} \Big)^\intercal. \tag{A22}$$

We also consider the version of subtraction $\mathring{\mathbf{s}}_t(u_r)$ from u_r to u_0 without taking the derivative with respect to $z_{l,t-r-1}$ as

$$\begin{aligned}\mathring{\mathbf{s}}_t(u_r) \triangleq \Big(&\frac{\partial^2\eta_{1t}(u_r)}{\partial z_{1t}} - \frac{\partial^2\eta_{1t}(u_0)}{\partial z_{1t}}, ..., \frac{\partial^2\eta_{nt}(u_r)}{\partial z_{nt}} - \frac{\partial^2\eta_{nt}(u_0)}{\partial z_{nt}},\\ &\frac{\partial^3\eta_{1t}(u_r)}{\partial z_{1t}^2} - \frac{\partial^3\eta_{1t}(u_0)}{\partial z_{1t}^2}, ..., \frac{\partial^3\eta_{nt}(u_r)}{\partial z_{nt}^2}^{\intercal} - \frac{\partial^3\eta_{nt}(u_0)}{\partial z_{nt}^2}\Big)^{\intercal}.\end{aligned} \tag{A23}$$

As provided below, in our case, the identifiability of $\mathbf{z}_t$ is guaranteed by the linear independence of the whole function vectors $\mathbf{s}_{lt}(u_r)$ and $\mathring{\mathbf{s}}_t(u_r)$, with $l = 1, 2, ..., n$ and every u_r. This linear independence is generally a much stronger condition. Theorem A1 can be considered as a special case where the number of domains u_r is 1. In this case, only $\mathbf{s}_{lt}(u_0)$ in Eq A22 is utilized but $2n$ values of $z_{l,t-r-1}$ are required. Otherwise, in the nonstationary case, the domain information u_r can increase the changeability of $s_{lt}(u_r)$. Besides, $\mathring{\mathbf{s}}_t(u_r)$ in Eq A23 can also help to find more independent vectors to satisfy the sufficiency assumption.

Corollary A1 (Identifiability under Non-Stationary Process)**.** *Suppose $\mathbf{x}_t = \mathbf{g}(\mathbf{z}_{t:t-r})$, $\mathbf{z}_t = \mathbf{m}(\mathbf{x}_{t:t-\mu})$, and that the conditional distribution $p(z_{kt} \,|\, \mathbf{z}_{t-1:t-r-1}, \mathbf{u})$ may change across $a+1$ values of the auxiliary variable $\mathbf{u}$, denoted by u_0, u_1, ..., u_a. Suppose the components of $\mathbf{z}_t$ are mutually independent conditional on $\mathbf{z}_{t-1:t-r-1}$ with each auxiliary variable. Assume that the components of $\hat{\mathbf{z}}_t$ are also mutually independent conditional on $\hat{\mathbf{z}}_{t-1:t-r-1}$. Suppose the domain is path-connected and $\mathbf{m}, \hat{\mathbf{m}}, \mathbf{g}, \hat{\mathbf{g}}$ are second-order differentiable and their combination subjects to non-degenerate condition. If there exists $2n$ different values of function vectors $\mathbf{s}_{lt}(u_r)$ or $\mathring{\mathbf{s}}_t(u_r)$ and $\mathring{\mathbf{s}}_t(u_r)$, with $l = 1, 2, ..., n$ and every u_r, are linearly independent, then $\hat{\mathbf{z}}_t$ is a permuted invertible component-wise transformation of $\mathbf{z}_t$.*

Proof. For any t we have

$$\begin{aligned}\mathbf{z}_t &= \mathbf{m}(\mathbf{x}_{t:t-\mu})\\ &= \mathbf{m}(\hat{\mathbf{g}}(\hat{\mathbf{z}}_t, \hat{\mathbf{z}}_{t-1:t-r}), \mathbf{x}_{t-1:t-\mu})\\ &= \mathbf{m}(\hat{\mathbf{g}}(\hat{\mathbf{z}}_t, \hat{\mathbf{m}}(\mathbf{x}_{t-1:t-\mu-1}), \cdots, \hat{\mathbf{m}}(\mathbf{x}_{t-r:t-\mu-r})), \mathbf{x}_{t-1:t-\mu}),\end{aligned} \tag{A24}$$

as well as $\hat{\mathbf{z}}_t = \hat{\mathbf{m}}(\mathbf{g}(\mathbf{z}_t, \mathbf{m}(\mathbf{x}_{t-1:t-\mu-1}), \cdots, \mathbf{m}(\mathbf{x}_{t-r:t-\mu-r})), \mathbf{x}_{t-1:t-\mu})$ similarly. Thus, we have an unified partially invertible function $\mathbf{z}_t = \mathbf{h}(\hat{\mathbf{z}}_t | \mathbf{x}_{t-1:t-\mu-r})$ where $\mathbf{h} = \mathbf{m} \circ \hat{\mathbf{g}}$ with Jacobian $\frac{\partial \mathbf{z}_t}{\partial \hat{\mathbf{z}}_t} = \mathbf{H}_t(\hat{\mathbf{z}}_t; \mathbf{x}_{t-1:t-\mu-r})$. Let us consider the mapping from joint distribution $(\hat{\mathbf{z}}_t, \mathbf{x}_{t-1:t-\mu-r-1})$ to $(\mathbf{z}_t, \mathbf{x}_{t-1:t-\mu-r-1})$, i.e.,

$$P(\mathbf{z}_t, \mathbf{x}_{t-1:t-\mu-r-1}) = P(\hat{\mathbf{z}}_t, \mathbf{x}_{t-1:t-\mu-r-1}) \,/\, |\mathbf{J}_t|, \tag{A25}$$

where

$$\mathbf{J}_t = \begin{bmatrix} \frac{\partial \mathbf{z}_t}{\partial \hat{\mathbf{z}}_t} & \mathbf{0} \\ * & \mathbf{I} \end{bmatrix}, \tag{A26}$$

which is a lower triangle matrix, where $\mathbf{I}$ infers eye matrix and $*$ infers any possible matrix. Thus, we have determinant $|\mathbf{J}_t| = |\frac{\partial \mathbf{z}_t}{\partial \hat{\mathbf{z}}_t}| = |\mathbf{H}_t|$. Dividing both sides of Eq A25 by $P(\mathbf{x}_{t-1:t-\mu-r-1}, u_r)$ gives

$$\mathbf{LHS} = P(\mathbf{z}_t | \mathbf{x}_{t-1:t-\mu-r-1}, u_r) = P(\mathbf{z}_t | \mathbf{z}_{t-1:t-r-1}, u_r), \tag{A27}$$

since $\mathbf{z}_t$ and $\mathbf{x}_{t-1:t-\mu-r-1}$ are independent conditioned on $\mathbf{z}_{t-1:t-r-1}$ with any auxiliary variable u_r. Similarly, $\mathbf{RHS} = P(\hat{\mathbf{z}}_t | \mathbf{x}_{t-1:t-\mu-r-1}, u_r) = P(\hat{\mathbf{z}}_t | \hat{\mathbf{z}}_{t-r-1}, u_r)$ holds true as well, which yields to

$$P(\mathbf{z}_t | \mathbf{z}_{t-1:t-r-1}, u_r) = P(\hat{\mathbf{z}}_t | \hat{\mathbf{z}}_{t-1:t-r-1}, u_r) \,/\, |\mathbf{H}_t|. \tag{A28}$$

With conditional independence, we have

$$\frac{\partial^2 \log p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1:t-r-1}, u_r)}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} = 0. \tag{A29}$$

Referencing Eq A28, it gets expressed as:

$$\log p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1:t-r-1}, u_r) = \log p(\mathbf{z}_t \,|\, \mathbf{z}_{t-1:t-r-1}, u_r) + \log |\mathbf{H}_t| = \sum_{k=1}^{n} \eta_{kt}(u_r) + \log |\mathbf{H}_t|. \tag{A30}$$

The second-order derivative is

$$\sum_{k=1}^{n}\Big(\frac{\partial^2\eta_{kt}(u_r)}{\partial z_{kt}^2}\cdot\mathbf{H}_{kit}\mathbf{H}_{kjt}+\frac{\partial\eta_{kt}(u_r)}{\partial z_{kt}}\cdot\frac{\partial\mathbf{H}_{kit}}{\partial\hat{z}_{jt}}\Big)+\frac{\partial^2\log|\mathbf{H}_t|}{\partial\hat{z}_{it}\partial\hat{z}_{jt}}\equiv 0. \tag{A31}$$

The right-hand side of the presented equation consistently equals 0. Therefore, for each index l ranging from 1 to n, and every associated value of $z_{l,t-r-1}$, its partial derivative with respect to $z_{l,t-r-1}$ remains 0. That is,

$$\sum_{k=1}^{n}\Big(\frac{\partial^3\eta_{kt}(u_r)}{\partial z_{kt}^2\partial z_{l,t-r-1}}\cdot\mathbf{H}_{kit}\mathbf{H}_{kjt}+\frac{\partial^2\eta_{kt}(u_r)}{\partial z_{kt}\partial z_{l,t-r-1}}\cdot\frac{\partial\mathbf{H}_{kit}}{\partial\hat{z}_{jt}}\Big)\equiv 0, \tag{A32}$$

where we leveraged the fact that entries of $\mathbf{H}_t$ do not depend on $z_{l,t-r-1}$.

Again start from Eq A31. Using the fact that $\mathbf{H}_t$ is not affected by the auxiliary variable, we can subtract the equation with u_0 from that of u_r. We have

$$0=\sum_{k=1}^{n}\Big(\Big(\frac{\partial^2\eta_{kt}(u_r)}{\partial z_{kt}^2}-\frac{\partial^2\eta_{kt}(u_0)}{\partial z_{kt}^2}\Big)\cdot\mathbf{H}_{kit}\mathbf{H}_{kjt}+\Big(\frac{\partial\eta_{kt}(u_r)}{\partial z_{kt}}-\frac{\partial\eta_{kt}(u_0)}{\partial z_{kt}}\Big)\cdot\frac{\partial\mathbf{H}_{kit}}{\partial\hat{z}_{jt}}\Big). \tag{A33}$$

Considering any given value of $\mathbf{z}_t$, there exists at least $2n$ different values of $\mathbf{s}_{lt}$ or $\mathring{\mathbf{s}}_t$, which corresponds to Eq A32 and Eq A33 respectively, such that they are linearly independent. To make the above equation hold true, one has to set $\mathbf{H}_{kit}\mathbf{H}_{kjt}=0$ or $i\neq j$. In other words, each row of $\mathbf{H}_t$ consists of at most a single non-zero entry, and $\mathbf{z}_t$ must be a component-wise transformation of a permuted version of $\hat{\mathbf{z}}_t$. □

A2. Synthetic experiments

A2.1. Synthetic Dataset Generation

In this section, we give 2 representative simulation settings for **NG** and **NG-TDMP** respectively to reveal the identifiability results. For each synthetic dataset, we set latent space to be 3, i.e., $\mathbf{x}_t\in\mathcal{X}\subseteq\mathbb{R}^3$.

Non-invertible Generation For **NG**, we set the transition lag as $\tau=1$. We first generate $10,000$ data points from a uniform distribution as the initial state $\mathbf{z}_0\sim U(0,1)$. For $t=1,\cdots,9$, each latent variable $\mathbf{z}_t$ will be generated from the proceeding latent variable $\mathbf{z}_{t-1}$ through a nonlinear function $\mathbf{f}$ with a non-additive zero-biased Gaussian noise ϵ_t ($\sigma=0.1$), i.e., $\mathbf{z}_t=\mathbf{f}(\mathbf{z}_t,\epsilon_t)$. To introduce the non-invertibility, the mixing function $\mathbf{g}$ leverages only the first two entries of the latent variables to generate the 2-d observation $\mathbf{z}_t=\mathbf{g}(x_{1,t},x_{2,t})\in\mathcal{Z}\subseteq\mathbb{R}^2$.

Time-Delayed Mixing Process For **UG-TDMP**, we set the transition lag as $\tau=1$ and mixing lag $r=2$. Similar to the Non-invertible Generation scenario, we generate the initial states from a uniform distribution and the subsequent latent variables following a nonlinear transition function. The noise is also introduced in a nonlinear Gaussian ($\sigma=0.1$) way. The mixing process is a nonlinear function with regard to $\mathbf{z}_t$ plus a side information from previous steps $\mathbf{z}_{t-1:t-2}$, i.e.,

$$\mathbf{x}_t=A_{3\times3}\cdot\sigma\big(B_{3\times3}\cdot\sigma(C_{3\times3}\cdot\mathbf{z}_t)\big)+\begin{bmatrix}0\\0\\D_{3\times1}\mathbf{z}_{t-1}+E_{3\times1}\mathbf{z}_{t-2}\end{bmatrix}, \tag{A34}$$

where σ refers to the ReLU function and the capital characters refer to matrices. Note that we make two modifications to show the advantage of **CaRiNG**. The reason we consider larger mixing lag is that it is a much more difficult scenario to handle, with more distribution from the mixing process and less dynamic information from transition. We run experiments in both scenarios with different transition and mixing lag. Besides, we also find out that even without time-lagged latent variables in the decoder, it leads to a smaller model that is more stable and easy to train. Refer to Table A1 for a detailed ablation study.

setting	$\tau = 1, r = 2$	$\tau = 2, r = 1$
CaRiNG	0.9436	0.9131
CaRiNG (lagged decoder)	0.9250	0.9220
TDRL	0.8947	0.7519

Table A1: Ablation study on different settings for **UG-TDMP**. (a) The second column is a more difficult scenario compared to the first, where the performance of **CaRiNG** remains good while that of baseline decreases significantly. (b) Omit the time-lagged latent variables in the decoder will not damage the performance much, but one can enjoy the benefits from a much simpler model.

Post-processing Precedure During the generating process, we did not explicitly enforce the data to meet the constraint $\mathbf{z}_t = \mathbf{m}(\mathbf{x}_{t:t-\mu})$. On the contrary, we implement a checker to filter the data that is qualified. To be more precise, we do linear regression from $\mathbf{x}_{t:t-\mu}$ to $\mathbf{z}_t$ to figure out how much information of latent variables can be recovered from observation series in the best case. We choose the smallest μ when the amount of information that can be recovered is acceptable. We set $\mu = 2$ for **UG** and $\mu = 4$ for **UG-TDMP**.

A2.2. Implementation Details

Network Architecture To implement the Sequence-to-Step encoder, we leverage the *torch.unfold* to generate the nesting observations. Let us denote $\mathbf{x}_t^{(\mu)} = [\mathbf{x}_t, \cdots, \mathbf{x}_{t-\mu}]$ as inputs. For the time steps that do not exist, we simply pad them with zero. Refer to Table A2 for detailed network architecture.

Training Details The models were implemented in PyTorch 1.11.0. An AdamW optimizer is used for training this network. We set the learning rate as 0.001 and the mini-batch size as 64. We train each model under four random seeds $(770, 771, 772, 773)$ and report the overall performance with mean and standard deviation across different random seeds.

Table A2: Architecture details. BS: batch size, T: length of time series, i_dim: input dimension, o_dim: output dimension, z_dim: latent dimension, LeakyReLU: Leaky Rectified Linear Unit.

Configuration	**Description**	**Output**
1. Sequence-to-Step Encoder	Encoder for Synthetic Data	
Input: $\mathbf{x}_{1:T}^{(\mu)}$	Observed time series	BS $\times$ T $\times$ i_dim
Dense	128 neurons, LeakyReLU	BS $\times$ T $\times$ 128
Dense	128 neurons, LeakyReLU	BS $\times$ T $\times$ 128
Dense	128 neurons, LeakyReLU	BS $\times$ T $\times$ 128
Dense	Temporal embeddings	BS $\times$ T $\times$ z_dim
2. Step-to-Step Decoder	Decoder for Synthetic Data	
Input: $\hat{\mathbf{z}}_{1:T}$	Sampled latent variables	BS $\times$ T $\times$ z_dim
Dense	128 neurons, LeakyReLU	BS $\times$ T $\times$ 128
Dense	128 neurons, LeakyReLU	BS $\times$ T $\times$ 128
Dense	i_dim neurons, reconstructed $\hat{\mathbf{x}}_{1:T}$	BS $\times$ T $\times$ o_dim
3. Factorized Inference Network	Bidirectional Inference Network	
Input	Sequential embeddings	BS $\times$ T $\times$ z_dim
Bottleneck	Compute mean and variance of posterior	$\mu_{1:T}, \sigma_{1:T}$
Reparameterization	Sequential sampling	$\hat{\mathbf{z}}_{1:T}$
4. Modular Prior	Nonlinear Transition Prior Network	
Input	Sampled latent variable sequence $\hat{\mathbf{z}}_{1:T}$	BS $\times$ T $\times$ z_dim
InverseTransition	Compute estimated residuals $\hat{\epsilon}_{it}$	BS $\times$ T $\times$ z_dim
JacobianCompute	Compute $\log\left(\lvert\det\left(\mathbf{J}\right)\rvert\right)$	BS

Table A3: **MCC scores of synthetic datasets with higher dimension.**

Dimension	CaRiNG	TDRL
6	0.9199	0.6329
12	0.9366	0.6155
18	0.7175	0.5265

A2.3. Exploration on higher dimension.

To demonstrate the scalability of our method, we have included experiments with higher dimensions. We keep the experimental setup consistent with NG and set the dimensions of latent variables to be $6, 12, 18$ and that of observation to be $4, 8, 12$, respectively. The transition function is a permutation function with a shift of $2, 4, 6$ dimensions respectively. As shown in Table A3, CaRiNG can achieve a consistent improvement over the baseline TDRL when using various dimensions. When the dimension is too high, although the performance of both CaRiNG and TDRL drops because of the complexity of the model, we demonstrate that CaRiNG still benefits from contextual information. This indicates that CaRiNG is scalable and robust to the dimensionality of the latent variables.

A2.4. Model Selection with Varying μ

In this subsection, we will discuss the preliminary experiment that was instrumental in the model selection process for our application in the NG-TDMP settings. The experiment focused on evaluating the performance of the model with varying lengths of time lag μ.

Our findings indicate that an increase in μ does not always correlate with enhanced model performance. We observed that the effectiveness of each latent variable diminishes as the time lag μ increases. In practical applications, this motivates a strategy of model selection where an appropriate value of μ is chosen based on the model's performance. The following table summarizes our experimental results:

μ	3	4	5
Accuracy (%)	0.88	0.92	0.92

Table A4: Impact of varying μ on model performance in NG-TDMP settings.

These results suggest that while a larger μ might imply a more extensive recovery of context information, it can also introduce inefficiencies in information recovery, potentially adding noise and impeding model training.

A3. Real-world Experiments on TrafficQA

A3.1. Implementation Details

We choose HCRN (Le et al., 2020) (without classification head) as the encoder backbone of **CaRiNG** on the real-world dataset: SUTD-TrafficQA. Given that HCRN is an encoder that calculates the cross attention between visual input and text input sequentially, we apply a decoder, which shares the same structure as the Step-to-Step Decoder shown in Table A2 to reconstruct the visual feature embedded with the temporal information. As it goes to transition prior, we use the Modular Prior shown in Table A2. This encoder-decoder structure can guide the model to learn the hidden representation with identifiable guarantees under the non-invertible generation process.

A3.2. More Qualitative Results

As shown in Figure A1, we provide some positive examples and also fail cases to analyze our model. From the top two examples, we can find that our method can solve the occlusions well. From the bottom right one, we find that our model can solve the blurred situation. However, when the alignment between visual and textual domains is difficult. The model may fail.

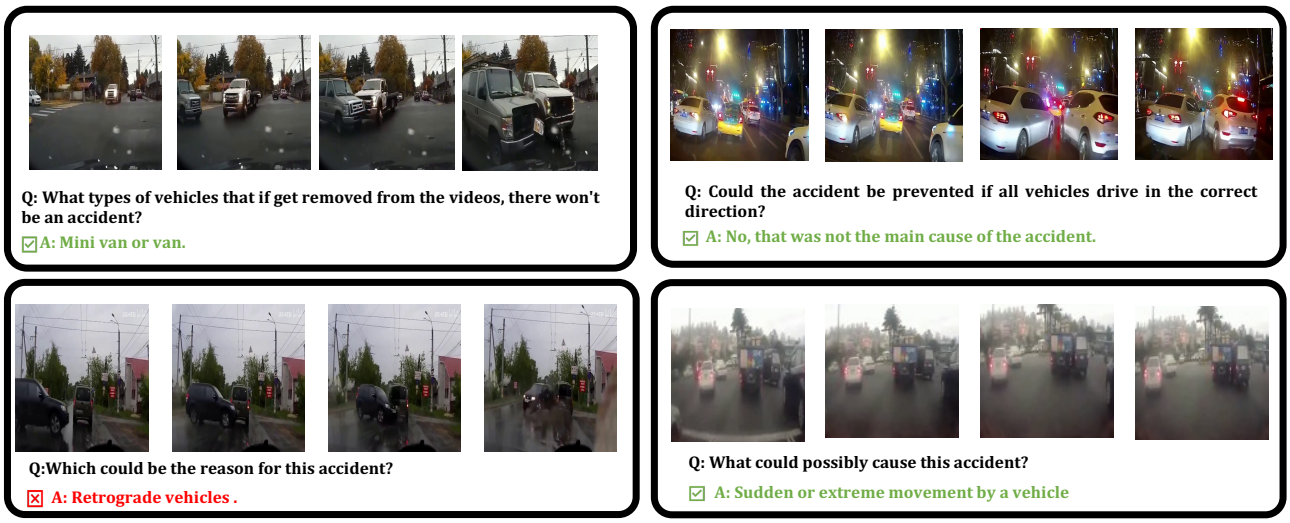


Figure A1: **Qualitative results on SUTD-TrafficQA dataset.** We provide some positive examples and also fail cases to analyze our model.

A3.3. Computation Cost Comparison

We provide the comparisons between the computational cost of the **CaRiNG** model and HCRN to analyze our efficiency. As shown in Table A5, we provide a detailed comparison of the number of parameters, training time, and inference efficiency. It is important to note that while the **CaRiNG** model requires a longer training time due to the application of normalizing flow for calculating the Jacobian matrix, its inference efficiency remains on par with HCRN, as the normalizing flow is utilized only for calculating KL loss and not during inference.

Method	HCRN	CaRiNG
Number of Parameters	42,278,786	43,721,954
Training Time per Epoch	6min 54s/epoch	13min 26s/epoch
Inference Time per Epoch	49s/epoch	49s/epoch

Table A5: Comparative Analysis of HCRN and **CaRiNG** Models

This analysis clearly demonstrates that the increased training time for the **CaRiNG** model is offset by its comparable inference efficiency, highlighting its practical applicability in scenarios where inference time is critical.

A3.4. Evaluation of Identifiability in the QA Benchmark

In the context of real-world applications, particularly in scenarios lacking ground truth for rigorous metrics like MCC, alternative evaluation strategies become essential. we leverage proxy metrics to assess the performance of the proposed algorithm, focusing on two pivotal aspects: disentanglement and reconstruction ability of the learned representations. Intuitively, as delineated in Theorem A1 and detailed in Section 4, a representation can be considered identifiable if it possesses the dual capability of fully reconstructing the observation while also achieving disentanglement. Thus, as a supplement to the accuracy we used before, we benchmark disentanglement and reconstruction ability as side evidence to support that the improvement is caused by better identifiability.

We use the ELBO loss as a proxy metric to evaluate the identifiability. Figure A2 illustrates our method's performance compared to the baseline TDRL method. The results clearly show that our approach exhibits superior disentanglement and reconstruction abilities. This evidence suggests that the advantage of our proposed algorithm is primarily attributed to its enhanced identifiability and effective disentanglement of data representations.

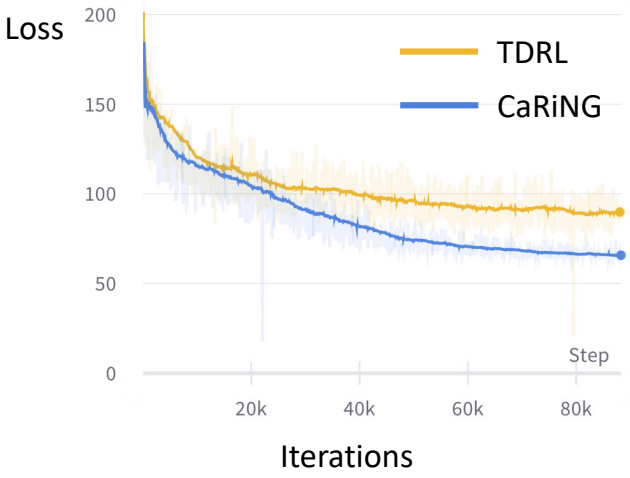


Figure A2: Comparative analysis of disentanglement and reconstruction abilities of different methods.

τ	1	2	3
Accuracy (%)	41.22	41.23	41.27

Table A6: Parameter analysis results of τ on model accuracy in the SUTD-TrafficQA dataset.

A3.5. Parameter analysis on τ

In this section, we present the results of our parameter analysis conducted on the SUTD-TrafficQA dataset, focusing on the impact of varying the time lag τ. The study aimed to assess the robustness of our model to changes in the time lag parameter. As the table below illustrates, the model demonstrates consistent accuracy across different values of τ, indicating robustness to the variation in time lag.

A4. Real-world Experiments on the Volleyball Dataset

A4.1. Dataset

The volleyball dateset (Ibrahim et al., 2016) is a video action recognition dataset with 4,830 clips from 55 videos. There are 8 group activity labels, including 4 main activities (set, spike, pass, win-point) that are divided into two subgroups, left and right. Two formats for inputs are provided: RGB videos and keypoints time series. In our setting, we simply use key points as the input. We utilized the 'original' split of the Volleyball dataset in which all videos were randomly assigned, consisting of 39 training videos and 16 testing videos. We adopt this dataset due to the complex occlusion in the sports which is aligned with our non-invertible generation setting.

A4.2. Implementation Details

The method is implemented using a VAE network. Specifically, the Sequence-to-Step Encoder processes the data by first flattening the features from all time steps. Then, following (Zhou et al., 2022), we apply a Composer to incorporate the interactions with fine-grained information. Subsequently, we aggregate the contextual information through an MLP, mapping from a space of $\mathbb{R}^{T\times d}$ to $\mathbb{R}^d$. The Step-to-Step Decoder is also an MLP network mapping from $\mathbb{R}^d$ to $\mathbb{R}^d$. We adopt the same Modular Prior network as Table A2. For the implementation of TDRL, the only difference is the removal of temporal dependencies during the encoding process of the model (don't aggregate the contextual information).

A4.3. Results and Analysis

As shown in Table A7, we observe that **CaRiNG** achieve consistent performance improvement on both person and group activity accuracy. It indicates that the temporal context is useful in the temporal dynamic modeling. Though the goal of

this task is the group activity recognition, we found that the person activity accuracy achieves more improvement. It is not surprising since our method ensures better disentanglement and identification of the latent variables of the group activity, i.e., containing the information of persons.

Method	CaRiNG	TDRL
Group Activity Top1 Accuracy(%)	93.044	92.895
Group Activity Top3 Accuracy(%)	99.028	98.280
Person Activity Top1 Accuracy(%)	74.551	73.286
Person Activity Top3 Accuracy(%)	98.087	96.634

Table A7: Model accuracy in the Volleyball dataset.

A5. Related Work

A5.1. Causal Discovery with Latent Variables

Some studies have aimed to discover causally related latent variables, such as (Silva et al., 2006; Kummerfeld & Ramsey, 2016; Huang et al., 2022) leverage the vanishing Tetrad conditions (Spearman, 1928) or rank constraints to identify latent variables in linear-Gaussian models, and (Shimizu et al., 2009; Cai & Xie, 2019; Xie et al., 2020; 2022) draw upon non-Gaussianity in their analysis for linear, non-Gaussian scenarios. Furthermore, some methods aim to find the structure beyond the latent variables, resulting in the hierarchical structure. Some hierarchical model-based approaches assume tree-like configurations, such as (Pearl, 1988; Zhang, 2004; Choi et al., 2011; Drton et al., 2017), while the other methods assume a broader hierarchical structure (Xie et al., 2022; Huang et al., 2022). However, these methods remain confined to linear frameworks and face escalating challenges with intricate datasets, such as videos.

A5.2. Nonlinear ICA for Time Series Data

Nonlinear ICA represents an alternative methodology to identify latent causal variables within time series data. Such methods leverage auxiliary data—like class labels and domain indices—and impose independence constraints to facilitate the identifiability of latent variables. To illustrate: Time-contrastive learning (TCL (Hyvarinen & Morioka, 2016)) adopts the independent sources premise and capitalizes on the variability in variance across different data segments. Furthermore, Permutation-based contrastive (PCL (Hyvarinen & Morioka, 2017)) puts forth a learning paradigm that distinguishes genuine independent sources from their permuted counterparts. Furthermore, i-VAE (Khemakhem et al., 2020) utilizes deep neural networks, VAEs, to closely approximate the joint distribution encompassing observed and auxiliary non-stationary regimes. Recent work, exemplified by LEAP (Yao et al., 2022b), has tackled both stationary and non-stationary scenarios in tandem. In the stationary context, LEAP postulates a linear non-Gaussian generative process. For the non-stationary context, it assumes a nonlinear generative process, gaining leverage from auxiliary variables. Advancing beyond LEAP, TDRL (Yao et al., 2022a) initially extends the linear non-Gaussian generative assumption to a nonlinear formulation for stationary scenarios. Subsequently, it broadens the non-stationary framework to accommodate structural shifts, global alterations, and combinations thereof. Additionally, CITRIS (Lippe et al., 2022b;a) champions the use of intervention target data to precisely identify scalar and multi-dimensional latent causal factors. However, a common thread across these methodologies is the presumption of an invertible generative process, a stance that often deviates from the realities of actual data. Besides, (Hartford et al., 2022) demonstrates that under a non-invertible scenario without extra information, identifiability can be only achieved in a subspace where bijective mapping exists. Their work provides additional support for the importance of addressing non-invertibility.

A5.3. Temporal modeling

Sequential Variational Autoencoders have gained significant popularity for their applications in temporal modeling, including generation, representation, and prediction. Variational RNN (Chung et al., 2015) introduces the Variational Autoencoders into Recurrent Neural Networks, enabling variational inference on time series data. SRNN (Fraccaro et al., 2016) further utilizes the concept of SSM (State Space Model) for temporal modeling. In addition, SKD (Berman et al., 2022) utilizes a structured Koopman autoencoder to achieve multifactor sequential disentanglement. However, none of these methods incorporates a transition function for capturing the temporal dynamics of multivariate data. By integrating a transition

function with independent noise through normalizing flow (Rezende & Mohamed, 2015; Ziegler & Rush, 2019), our model can effectively track and represent the causal relations of latent variables over time. Such enhancement positions **CaRiNG** as a method focused on learning causal representations with clear identifiability guarantees, marking a departure from the generation-centric objectives commonly seen in traditional VAE-based methods.