LEARNING REPRESENTATIONS OF INTERMITTENT TEMPORAL LATENT PROCESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Identifying time-delayed temporal latent process is crucial for understanding temporal dynamics and enabling downstream reasoning. Although recent methods have made remarkable progress in this field, they cannot address the dynamics in which the influence of some latent factors on both the subsequent latent states and the observed data can become inactive or irrelevant at different time steps. Therefore, we introduce intermittent temporal latent processes, where: (1) any subset of latent factors may be missing during nonlinear data generation at any time step, and (2) the active latent factors at each step are unknown. This framework encompasses both nonstationary and stationary transitions, accommodating changing or consistent active factors over time. Our work shows that under certain assumptions, the latent variables are block-wise identifiable. With further conditional independence assumption, each latent variable can even be recovered up to component-wise transformations. Using this identification theory, we propose an unsupervised approach, `InterLatent` , to reliably uncover the representations of the intermittent temporal latent process. The experiments on both synthetic and real-world datasets verify our theoretical claims.

1 INTRODUCTION

Learning meaningful representations from sequential data remains a fundamental challenge across various fields. Time series data, such as financial markets and climate observations, are ubiquitous and exhibit high nonlinearity Berzuini et al. (2012); Ghysels et al. (2016). This inspires a extensive line of works to temporal latent representation learning Yao et al. (2022b;a); Chen et al. (2024), upon the recent advancement in nonlinear ICA Khemakhem et al. (2020); Zhang et al.; Kong et al. (2022); Zheng et al. (2022); Li et al. (2023); von Kügelgen et al. (2024); Ng et al. (2023); Zheng & Zhang (2024); Morioka & Hyvarinen (2024); Yao et al. (2024); Zheng et al. (2024); Kong et al. (2024); Lachapelle et al. (2024a). However, many real-world systems exhibit latent time-delayed dynamics where the influence of certain latent factors on both subsequent latent states and observed data can be inactive or irrelevant at specific time steps. Consider, for example, a complex manufacturing process: various machine components contribute to the final product quality at different stages, with some components becoming temporarily inactive or irrelevant during certain production phases. Current works may struggle to capture these intermittent influences, potentially missing crucial aspects of the underlying dynamics. This highlights the need for a more flexible and robust framework to identify such temporal processes with intermittence of latent variables.

In this work, we investigate the identification of representations of intermittent temporal latent processes. Two key properties characterize the intermittence of a temporal latent process: (1) any subset of latent factors can be missing during the nonlinear time-delayed data generation at any time step, and (2) the specific set of active latent factors at a time step is unknown. Figure 1 takes an example of data generating mechanism of an intermittent latent temporal process to illustrate its concept. In the transition mechanism (top of Figure 1b), we see the zero entries in Jacobians indicating that not all latent variables influence each other's transitions. Similarly, in the generating mechanism (bottom of Figure 1b), the sparse Jacobians show that not all latent variables contribute to every observed variable. We define the "support" as the set of active latent factors at each time step, for both the transition and generating mechanisms. "Missingness" occurs when a latent factor is absent from the support, having no influence on the subsequent latent state or the observed data.

The intermittent nature of these processes presents two significant challenges for representation learning: (1) The supports for both transition and generating mechanisms are unknown, requiring methods to adapt to the data generated by only active latent factors at each time step. (2) The interactions between intermittently active latent variables may be intricate and time-varying in both mechanisms, necessitating the models that can capture the possible various support and missingness. The existing literature has yet to fully address these challenges. Wiedemer et al. (2024) relies on compositional mixing functions and requires supervision on the latent variables. Lachapelle et al. (2023); Fumero et al. (2023); Xu et al. (2024) are restricted to linear or piecewise linear settings. CaRiNG Chen et al. (2024) tackles missingness only within the mixing function by leveraging historical information during the unmixing process.

In contrast to previous works, we present identification guarantees for uncovering intermittent temporal latent processes. Our theoretical analysis begins with establishing block-wise identifiability under assumptions of the sufficient variability of transitions of latent variables (Theorem 1). Notably, this outcome holds regardless of whether latent variables are within the support or missingness. Building on this foundation, we further prove component-wise identifiability for latent variables within the support in Theorem 2, given an additional assumption of independence of latent variables conditioning on previous time steps. Moreover, our identifiability results are able to handle both nonstationary and stationary temporal latent process, allowing for changing or consistent active factors over time without compromising identifiability guarantees. These theoretical contributions establish, to the best of our knowledge, one of the first general frameworks for uncovering latent variables in intermittent temporal processes with appropriate identifiability guarantees.

Leveraging these theoretical insights, we introduce a novel unsupervised method that extends the Sequential Variational Autoencoder Li & Mandt (2018). Our method, **`InterLatent`** , accommodates supports and missingnesses through sparsity regularizations on both mixing and transition functions, enabling it to model complex interactions between intermittently active latent variables and handle sparse, temporally variable latent spaces. We evaluate our approach on synthetic and real-world datasets, demonstrating its effectiveness in uncovering complex hidden temporal processes, as well as validating the proposed identifiability theory.

2 PROBLEM SETTING

Given a temporal sequence ranging from $t = 1$ to $t = T$, let $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\}$ denote the K-dimensional observations. At each time step t, N latent causal variables $\mathbf{z}_t = \{z_t^1, \dots, z_t^N\}$ generate $\mathbf{x}_t \in \mathbb{R}^K$. We formalize the data generating process as follows:

$$\mathbf{x}_t = g(\mathbf{z}_t), \quad \mathbf{z}_t^n = f_n(\text{Pa}(\mathbf{z}_t^n), \epsilon_t^n) \quad \text{for} \quad n \in [1, N] \tag{1}$$

Here, g is assumed to be an injective, nonlinear, non-parametric mixing function: $\mathbb{R}^N \to \mathbb{R}^K$. In this work, we work on the undercomplete case, where $K \geq N$ to ensure the injectivity of g. f_n denotes the nonlinear, nonparametric time-delayed transition function for the n-th latent variable. $\text{Pa}(z_t^n)$ represents the parent nodes of $\mathbf{z}_t^n$ from previous time steps. Without loss of generality, we assume a time lag of 1 in Eq. equation 1, i.e., $\text{Pa}(\mathbf{z}_t^n) \subset \mathbf{z}_{t-1}$. The general case of multiple lags and sequence lengths is discussed in Appendix B.1. ϵ_t^n is the noise term, sampled independently for each $\mathbf{z}_t^n$ from a standard normal distribution $\mathcal{N}(0, 1)$.

We are now ready to introduce the intermittent temporal latent process upon the concept of missingness of latent variables. In particular, not all components of $\mathbf{z}_t$ participate in the data generating process at a time step. Formally, there exists a $u \in [1, N]$ such that the u-th row of Jacobian of the transition function f^u, denoted as $J_{f,t}^{u,:}$, and the u-th column of the Jacobian of the mixing function, $J_{g,t}^{:,u}$, are zero. This implies that when z_t^u is missing, it neither receives influence from $\mathbf{z}_{t-1}$ nor exerts influence on $\mathbf{z}_{t+1}$ or $\mathbf{x}_t$ in the data generation process. Figure 1 illustrates this concept, where $\mathbf{z}_2^1$ and $\mathbf{z}_3^2$ are examples of such missing latent variables.

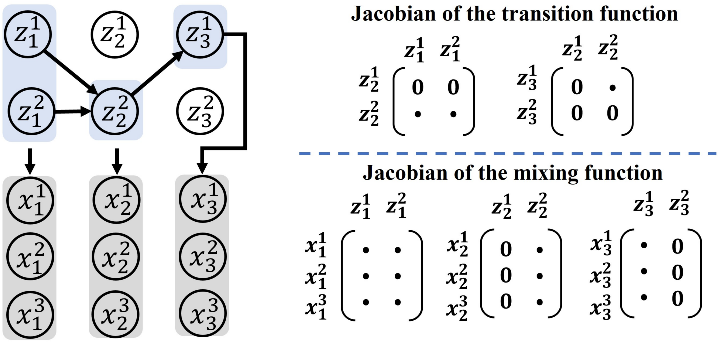
The "non-missing" indices of $\mathbf{z}_t$ help to define the support of the data generating process in Eq. 1 by:

$$\mathbf{s}_t := \{i \in [1, N] \mid \exists \mathbf{z}_{t-1} \text{ and } \mathbf{z}_t, J_{g,t}^{:,i}(\mathbf{z}_t) \neq \mathbf{0} \wedge J_{f,t}^{i,:}(\mathbf{z}_{t-1}) \neq \mathbf{0} \wedge J_{f,t+1}^{:,i}(\mathbf{z}_t) \neq \mathbf{0}\} \tag{2}$$

A similar formulation can define the missingness of $\mathbf{z}_t$ by $\mathbf{s}_t^c$. We assume that $\mathbf{s}_t$ and $\mathbf{s}_t^c$ partition the index set $[1, N]$

$$\mathbf{s}_t^c := \{u \in [1, N] \mid \exists \mathbf{z}_{t-1} \text{ and } \mathbf{z}_t, J_{g,t}^{:,u}(\mathbf{z}_t) = \mathbf{0} \wedge J_{f,t}^{u,:}(\mathbf{z}_{t-1}) = \mathbf{0} \wedge J_{f,t+1}^{:,u}(\mathbf{z}_t) = \mathbf{0}\} \tag{3}$$

Equations 2 and 3 sets the stage for identifying $\mathbf{z}_t$ by characterizing a sparse support at t of both transition and mixing functions. Specifically, there may exist latent variables $\{\mathbf{z}_t^u | u \in \mathbf{s}_t^c\}$ that do not participate in the data generating process described in Eq. 1. The zero entries in the Jacobian matrices of both the transition and mixing functions, as illustrated in Figure 1(b), provide a clear visual representation of this sparsity. Let $d_t = |\mathbf{s}_t|$ denote the cardinality of $\mathbf{s}_t$, and consequently, $|\mathbf{s}_t^c| = d_t^c = N - d_t$. In our analysis, we assume both $\mathbf{s}_t$ and $\mathbf{s}_t^c$ are non-empty for all time steps t. The case where $\mathbf{s}_t^c = \emptyset$ can be considered a special instance of the intermittent temporal latent process. In order to introduce our identification resutls, we define the observational equivalence next.



(a) Data generating process

(b) Jacobian structures of transition and mixing functions

Figure 1: Data generations of Intermittent Temporal Latent Process and its Jacobian Structures for a three-step sequence, e.g., $\mathbf{z}_3^1$ means the latent variable $\mathbf{z}^1$ at $t = 3$. **(a)** illustrates the connections between time steps and how $\mathbf{z}_t$ generates $\mathbf{x}_t$ in intermittent temporal latent process. (b) Jacobian structures reveals the definition of support and missingness in Eq. 2 and Eq. 3 by • and 0, respectively.

Definition 1 **(Observational Equivalence)**: *Given a sequence of observed variables* $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ *for* $t = 1$ *to* T*, let the true temporally causal latent process be specified by* $f, g, p(\epsilon)$ *as in Eq. equation 1. A learned generative model* $\hat{f}, \hat{g}, p(\hat{\epsilon})$ *is observationally equivalent to the ground truth if the model distribution matches the data distribution everywhere:*

$$p_{\hat{f},\hat{g},p(\hat{\epsilon})}(\mathbf{x}_{1:T}) = p_{f,g,p(\epsilon)}(\mathbf{x}_{1:T}) \tag{4}$$

Both the mixing function and transition functions can be recovered (up to certain indeterminacies) once z_t is identified as we assume the injectivity of g and no latent causal confounders, respectively.

Suppose there exists an invertible mapping h, such that $\hat{\mathbf{z}}_t = h(\mathbf{z}_t)$. We further provide the definition of block-wise identifiability and component-wise idenfiability in the following

Definition 2 **(Block-wise Identifiability)**: *h is considered block-wise identifiable if, given a block of the true latent variables* $\mathbf{z}_t^{\mathcal{B}}$*, there exists an unique partitioning* $\mathcal{B}'$ *of* $\hat{\mathbf{z}}_t$ *that matches* $\mathbf{z}_t^{\mathcal{B}}$ *up to a permutation* π*, such that* $\hat{\mathbf{z}}_t^{\mathcal{B}'} = h^{\mathcal{B}}(\pi(\mathbf{z}_t^{\mathcal{B}}))$*, where* $h^{\mathcal{B}}$ *is invertible.*

Definition 3 **(Component-wise Identifiability)**: *For an individual component of the latent variables* $\mathbf{z}_t^n$*, there exists a unique component* n' *of* $\hat{\mathbf{z}}_t$ *matches* $\mathbf{z}_t^n$ *up to a permutation* π*, such that* $\hat{\mathbf{z}}_t^{n'} = h^n(\pi(\mathbf{z}_t^n))$*, where* h^n *is invertible.* $\mathbf{z}_t^n$ *is component-wise identifiable.*

3 IDENTIFIABILITY THEORY

This section presents our identifiability results. We first leverage the assumptions of sufficient variability of temporal data and support sparsity to establish block-wise identifiability, as detailed in Theorem 1. Building upon this foundation, we then demonstrate component-wise identifiability of latent variables by exploiting conditional independence assumptions, as formalized in Theorem 2.

3.1 BLOCK-WISE IDENTIFIABILITY

Eq. 2 and 3 enable the partitioning of $\mathbf{z}_t$ into two subsets: $\{\mathbf{z}_t^i | i \in \mathbf{s}_t\}$ and $\{\mathbf{z}_t^u | u \in \mathbf{s}_t^c\}$. We now present our findings on the block-wise identifiability of these subsets:

Theorem 1 *For the observations $\mathbf{x}_t \in \mathbb{R}^K$ and estimated latent variables $\hat{\mathbf{z}}_t \in \mathbb{R}^N$, suppose that there exist functions $\hat{g}$ and $\hat{f}$ satisfying observational equivalence in Eq. 4 If the following assumptions and regularization hold:*

- *i (Smoothness and positivity): The probability density function of the latent causal variables, $p(\mathbf{z}_t)$, has positive measure in the space of $\mathbf{z}_t$ and is twice continuously differentiable.*
- *ii (Path connected): For any $\mathbf{z}^0, \mathbf{z}^1 \in \mathcal{Z}$, there is a continous function $\phi : [0,1] \rightarrow \mathcal{Z}$, s.t. $\phi(0) = \mathbf{z}^0$ and $\phi(1) = \mathbf{z}^1$.*
- *iii (Sufficient variability of $\mathbf{z}_t$ and $\hat{\mathbf{z}}_t$): Let $q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \log p(\mathbf{z}_t|\mathbf{z}_{t-1})$, as well as $q(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}) = \log p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1})$, and $\mathcal{H}_{\mathbf{z}_t,\mathbf{z}_{t-1}} q(\mathbf{z}_t|\mathbf{z}_{t-1})$ denotes the Hessian matrix of $q(\mathbf{z}_t|\mathbf{z}_{t-1})$ w.r.t. $\mathbf{z}_t$ and $\mathbf{z}_{t-1}$. Suppose $G^{\mathbf{z}_t} \in \{0,1\}^{N\times N}$ as a binary adjacency matrix that indicates the existence of the transitions from $\mathbf{z}_{t-1}$ and $\mathbf{z}_t$, where $G^{\mathbf{z}_t}_{i1,i2} = 1$ means that there exists a transition from $\mathbf{z}_{t-1}^{i1}$ to $\mathbf{z}_t^{i2}$. We assume that:*

  $$\mathit{span}\{\mathcal{H}_{\mathbf{z}_t,\mathbf{z}_{t-1}} q(\mathbf{z}_t|\mathbf{z}_{t-1})\}_{j=1}^{d_t} = \mathbb{R}^{d_t\times d_t}_{G^{\mathbf{z}_t}}$$

  and

  $$\mathit{span}\{\mathcal{H}_{\hat{\mathbf{z}}_t,\hat{\mathbf{z}}_{t-1}} q(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1})\}_{j=1}^{\hat{d}_t} = \mathbb{R}^{\hat{d}_t\times \hat{d}_t}_{\hat{G}^{\hat{\mathbf{z}}_t}}.$$

- *iv (Support sparsity regularization): For any time step t, $\mathbf{s}_t$ is not an empty set, $\hat{d}_t \leq d_t$*

There exists a permutation σ, such that

$$\hat{\mathbf{s}}_t = \sigma(\mathbf{s}_t) \quad \mathit{and} \quad \hat{\mathbf{s}}_t^c = \sigma(\mathbf{s}_t^c).$$

In other words, $\forall i$, $\{\mathbf{z}_t^i | i \in \mathbf{s}_t\}$ and $\forall u$, $\{\mathbf{z}_t^u | u \in \mathbf{s}_t^c\}$ are block-wise identifiable.

Proof sketch The complete proof is provided in the Appendix B.1. Here, we outline the key steps: First, let $p(\mathbf{z}_t|\mathbf{z}_{t-1})$ be the ground-truth transition pdf and $p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1})$ be the estimated transition pdf. Define $q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \log p(\mathbf{z}_t|\mathbf{z}_{t-1})$ and $q(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}) = \log p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1})$. Using h defined in $\hat{\mathbf{z}}_t = h(\mathbf{z}_t)$ helps us derive: $\mathcal{H}_{\hat{\mathbf{z}}_t,\hat{\mathbf{z}}_{t-1}} q(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}) = (J_{h^{-1}}(\hat{\mathbf{z}}_t))^\intercal \mathcal{H}_{\mathbf{z}_t,\mathbf{z}_{t-1}} q(\mathbf{z}_t|\mathbf{z}_{t-1}) J_{h^{-1}}(\hat{\mathbf{z}}_{t-1})$, where $\mathcal{H}$ denotes the Hessian matrix and $J_{h^{-1}}$ is the inverse Jacobian of h. We further leverage the sufficient variability assumption to establish a connection between the support sets $\mathbf{s}_t$ and $\hat{\mathbf{s}}_t$, as well as their complements $\mathbf{s}_t^c$ and $\hat{\mathbf{s}}_t^c$, respectively. By incorporating the support sparsity regularization $\hat{d}_t \leq d_t$, we conclude the block-wise identifiability of both $\forall i$, $\{\mathbf{z}_t^i | i \in \mathbf{s}_t\}$ and $\forall u$, $\{\mathbf{z}_t^u | u \in \mathbf{s}_t^c\}$.

Remarks Assumptions *i* and *ii* have been commonly adopted for the identification theory Chen et al. (2024); Lachapelle et al. (2024b). These assumptions provide the foundations to present Theorem 1, which concerns the transitions from $\mathbf{z}_{t-1}$ to $\mathbf{z}_t$ over the space $\mathcal{Z}$.

Recall that this work does not require all components of $\mathbf{z}_t$ to actively participate in the data generation process. The crux of our identification approach lies in formalizing the relationships between the support sets $\mathbf{s}_t$ and $\hat{\mathbf{s}}_t$, as well as their complements $\mathbf{s}_t^c$ and $\hat{\mathbf{s}}_t^c$. To this end, we further introduce the sufficient variability in assumption *iii* to ensure the span of the Hessian matrices of the log-transition probabilities covers the full space of $\forall i$, $\{\mathbf{z}_t^i | i \in \mathbf{s}_t\}$. We can thus establish our main result on partial identifiability by leveraging support sparsity regularization to reach our conclusion of the block-wise identifiability.

Notably, we do not assume the invariance of the support set $\mathbf{s}_t$ over time. Regardless of whether $\mathbf{s}_t$ changes or remains constant, we demonstrate in Section 3.2 that for all i, $\{\mathbf{z}_t^i | i \in \mathbf{s}_t\}$ can be recovered up to a component-wise invertible transformation and a permutation.

3.2 COMPONENT-WISE IDENTIFIABILITY

In this section, we exploit the conditional independence assumption to establish the component-wise identifiability of $\{\mathbf{z}_t^i | i \in \mathbf{s}_t\}$.

Theorem 2 *Let all assumptions from the Theorem 1 hold. Additionally, suppose the following assumption is exposed to data generating process in Eq. 1 as well:*

i (Conditional independence): At t, we assume that each component of $\mathbf{z}_t$ is conditional independent given the previous latent variables $\mathbf{z}_{t-1}$. For any $i1, i2 \in [N]$:

$$\mathbf{z}_t^{i1} \perp\!\!\!\perp \mathbf{z}_t^{i2} | \mathbf{z}_{t-1} \tag{5}$$

Then $\{\hat{\mathbf{z}}_t^j | j \in \hat{\mathbf{s}}_t\}$ must be a component-wise transformation of a permuted version of true $\{\mathbf{z}_t^i | i \in \mathbf{s}_t\}$.

Proof Sketch: Our main idea rests on proving component-wise identifiability by contradiction. We demonstrate that if component-wise identifiability does not hold, it would violate the conditional independence assumption. More specifically, The proof proceeds as follows: 1. From the Theorem 1, we have: $\mathcal{H}_{\hat{\mathbf{z}}_t,\hat{\mathbf{z}}_{t-1}} q(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}) = (J_{h^{-1}}(\hat{\mathbf{z}}_t))^\intercal \mathcal{H}_{\mathbf{z}_t,\mathbf{z}_{t-1}} q(\mathbf{z}_t|\mathbf{z}_{t-1}) J_{h^{-1}}(\hat{\mathbf{z}}_{t-1})$; 2. the conditional independence of $q(\mathbf{z}_t|\mathbf{z}_{t-1})$ as established in Eq. 5 only hold if $J_h^{-1}(\hat{\mathbf{z}}_t)$ is a diagonal matrix.

Remarks: The conditional independence is widely adopted in identification results for time-series data, as evidenced in recent works Yao et al. (2022b;a); Chen et al. (2024). Our analysis demonstrates that this regularization plays a crucial role in establishing our identification results.

The conclusions derived from Theorem 1 and Theorem 2 are applicable to both stationary and non-stationary processes. We consider the process is nonstationary if the support sets $\mathbf{s}_t$ and $\mathbf{s}_t^c$ vary over time since the transition from $\mathbf{z}_{t-1}$ to $\mathbf{z}_t$ changes as well. Conversely, the process is stationary if these support sets remains unchanged over time. Our framework allows for temporal variation in the support sets $\mathbf{s}_t$ and their complements $\mathbf{s}_t^c$, subject only to the constraint that neither is an empty set at any time point.

Our proposed data generating process encompasses previous models as special cases. For instance, if we remove the intermittent feature described in Eq. 2 and Eq. 3, our model reduces to LEAP Yao et al. (2022b) without the domain index. When handling non-stationary sequences, our identifiability results removes the assumption of known auxiliary variables, which is required by Yao et al. (2022b;a); Chen et al. (2024).

4 INTERLATENT APPROACH

Building upon our identifiability results, we now introduce InterLatent to estimate the latent causal variables. Our approach aims to achieve the observational equivalence by modeling the support sparsity and the conditional independence assumptions for the data generating process in Eq. 1. In general, InterLatent formalizes the probabilistic joint distribution of Eq. 1 as:

$$p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = p_\gamma(\mathbf{x}_1|\mathbf{z}_1) p_\phi(\mathbf{z}_1) \prod_{t=2}^{T} p_\gamma(\mathbf{x}_t|\mathbf{z}_t) p_\phi(\mathbf{z}_t|\mathbf{z}_{t-1}). \tag{6}$$

where γ denotes the parameters for the mixing function g, and ϕ denotes the parameters for the transition function f. To learn $\mathbf{z}_t$ from the observations $\mathbf{x}_t$, we also introduce the encoder $q_\omega(\mathbf{z}_t|\mathbf{x}_t)$ with parameters ω. We build our approach upon Sequential Variational Auto-Encoders Li & Mandt (2018). Figure 2 illustrates the overall framework of **`InterLatent`** . In what follows, we introduce each part of our network individually.

4.1 NETWORK DESIGN

Eq. 6 suggests that the architecture of InterLatent comprises of three key components. The encoder acquires latent causal representations by inferring $q_\omega(\hat{z}_t|x_t)$ from observations. These learned latent variables are then used by the step-to-step decoder $p\gamma(\hat{x}_t|\hat{z}_t)$ to reconstruct the observations, implementing the mixing function g in Eq. 1. To learn the latent variables, we constrain them through the KL divergence between their posterior distribution and a prior distribution, which is estimated using a normalizing flow that converts the prior into Gaussian noise. A detailed exploration of all modules is forthcoming.

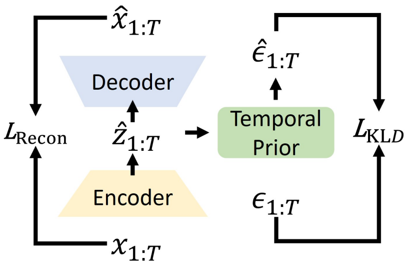


Figure 2: The overall framework of **`InterLatent`** consists of: (1) an encoder that maps observations $\mathbf{x}_t$ to latent variables $\hat{\mathbf{z}}_t$ $(t \in [1, T])$, (2) a decoder that reconstructs observations $\hat{\mathbf{x}}_t$ $(t \in [1, T])$ from $\mathbf{z}_t$, and (3) a temporal prior estimation module that models the transition dynamics between latent states. We train InterLatent by L_{Recon} along L_{KLD}. $\hat{\epsilon}_t$ $(t \in [1, T])$ denotes the estimation of the true noise terms ϵ_t $(t \in [1, T])$.

Encoder $q_\omega(\hat{\mathbf{z}}_t|\mathbf{x}_t)$: We assume $\hat{\mathbf{z}}_t$ is independent of $\hat{\mathbf{z}}_{t'}$ conditioning on $\mathbf{x}$, where $t \neq t'$. Therefore, the decomposition of joint probability distribution of the posterior is $q_\omega(\hat{\mathbf{z}}|\mathbf{x}) = \prod_{t=1}^{T} q_\omega(\hat{\mathbf{z}}_t|\mathbf{x}_t)$. We choose to approximate q by an isotropic Gaussian characterized by mean μ_t and covariance σ_t. To learn the posterior we use an encoder composed of an MLP followed by leaky ReLU activation:

$$\hat{\mathbf{z}}_t \sim \mathcal{N}(\mu_t, \sigma_t), \mu_t, \sigma_t = \text{LeakyReLU}(\text{MLP}(\mathbf{x}_t)). \tag{7}$$

Temporal Prior Estimation $p_\phi(\mathbf{z}_t|\mathbf{z}_{t-1})$: To enforce the conditional independence assumption in Eq. 5, we minimize the KL divergence between the posterior distribution and a prior distribution. This approach encourages the posterior to adopt the independence property as well, such that $\hat{\mathbf{z}}_t|\mathbf{x}_t$ are mutually independent. To address the challenges of directly estimating the arbitrary density function $p_\phi(\mathbf{z}_t|\mathbf{z}_{t-1})$, we introduce a transition prior module based on normalizing flows. This design represents the prior as a Gaussian distribution transformed by the Jacobian of the transition function, enabling efficient computing. Formally, $\forall j, \{\hat{\mathbf{z}}_t^j | j \in \hat{\mathbf{s}}_t\}$, we formulate the prior module as $\hat{\epsilon}_t^j = \hat{f}_j^{-1}(\hat{\mathbf{z}}_t^j|\hat{\mathbf{z}}_{t-1})$. This computation meets the requirement that f_n to be invertible. Then the prior distribution of the j-th dimension of the temporal dynamics, $\hat{\mathbf{z}}_t^j$, can be computed as $p_\phi(\hat{\epsilon}_t^j)|\frac{\partial \hat{f}_j^{\;-1}}{\partial \hat{\mathbf{z}}_t^j}| = p_\phi(\hat{f}_j^{-1}(\hat{\mathbf{z}}_t^j|\hat{\mathbf{z}}_{t-1}))|\frac{\partial \hat{f}_j^{\;-1}}{\partial \hat{\mathbf{z}}_t^j}|$.

In addition, for any v, such that $\{\hat{\mathbf{z}}_t^v | v \in \hat{\mathbf{s}}_t^c\}$, we evaluate that $\hat{\epsilon}_t^v = \hat{f}_v^{-1}(\hat{\mathbf{z}}_t^v)$. The prior distribution is calculated by $p_\phi(\hat{\epsilon}_t^v)|\frac{\partial \hat{f}_v^{\;-1}}{\partial \hat{\mathbf{z}}_t^v}| = p_\phi(\hat{f}_v^{-1}(\hat{\mathbf{z}}_t^v))|\frac{\partial \hat{f}_v^{\;-1}}{\partial \hat{\mathbf{z}}_t^v}|$ as $\hat{\mathbf{z}}_t^v$ is independent of $\hat{\mathbf{z}}_{t-1}$. Combing together, the total prior distribution is:

$$p_\phi(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}) = \prod_{n=1}^{N} p_\phi(\hat{\epsilon}_t^n)|\frac{\partial \hat{f}_n^{-1}}{\partial \hat{\mathbf{z}}_t^n}| \tag{8}$$

The flow model f in Eq. 8 is built with the MLP layers. For more details on the derivations of prior estimation, please refer to Appendix C.1.

Decoder $p_\gamma(\hat{\mathbf{x}}_t|\hat{\mathbf{z}}_t)$: The decoder pairs with our encoder to generate an reconstruct of the observation $\hat{\mathbf{x}}_t$ from the estimated latent variables $\hat{\mathbf{z}}_t$, which consists of a stacked MLP followed by leaky ReLU activation:

$$\hat{\mathbf{x}}_t = \text{LeakyReLU}(\text{MLP}(\hat{\mathbf{z}}_t)). \tag{9}$$

4.2 LEARNING OBJECTIVE

In this work, we extend our learning objective from Sequential Variational Autoencoder Li & Mandt (2018) with a modified ELBO. In general, the ELBO implements the observational equivalence requirement from Definition 1, which ensures our learned model matches the data-generating distribution. We formulate the entire ELBO objective in the following:

$$\begin{aligned}\mathcal{L}_{\text{ELBO}} = &\underbrace{\sum_{t=1}^{T} \mathbb{E}_{\hat{\mathbf{z}}_t \sim q_\omega} \log p_\gamma(\hat{\mathbf{x}}_t|\hat{\mathbf{z}}_t)}_{\mathcal{L}_{\text{Recon}}} - \underbrace{\sum_{t=1}^{T} \beta \mathbb{E}_{\hat{\mathbf{z}}_t \sim q_\omega} \big(\log q\left(\hat{\mathbf{z}}_t|\mathbf{x}_t\right) - \log p_\phi(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1})\big)}_{\mathcal{L}_{\text{KLD}}} \\ &\underbrace{+ \sum_{t=1}^{T} (|J_{\hat{g},t}|_{2,1} + |J_{\hat{f},t}|_{1,1}) + \sum_{t=2}^{T} |J_{\hat{f},t}|_{2,1}}_{\text{Sparsity Regularization}}\end{aligned} \tag{10}$$

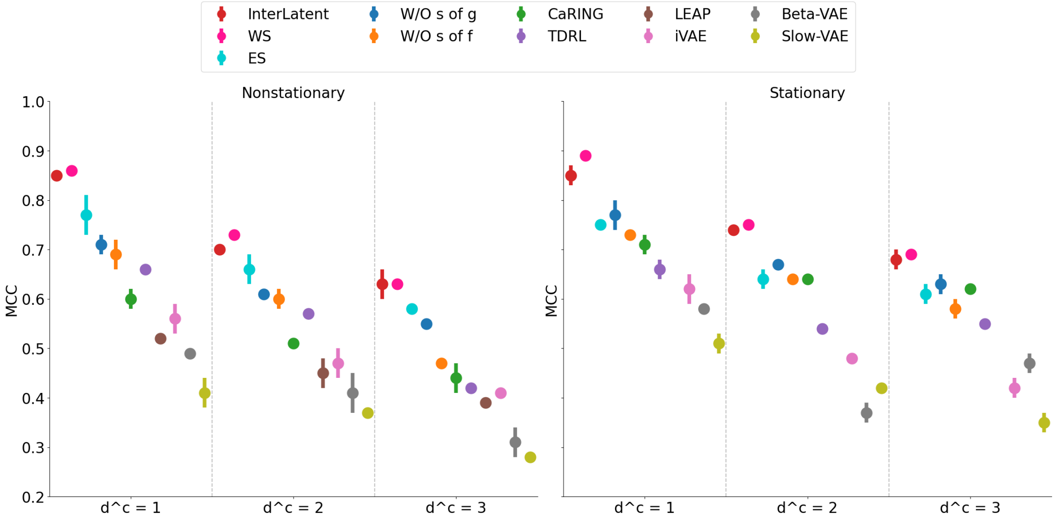


Figure 3: Mean Correlation Coefficient (MCC) scores for various methods for both ”Nonstationary” and ”Stationary” settings. d^c means the size of $\mathbf{s}_t^c$ in a sequence. Higher MCC scores indicate better performance in identifying latent variables.

where β is the hyperparameter to balance the two losses. The reconstruction loss $\mathcal{L}_{\text{Recon}}$ minimizes the discrepancy between $\mathbf{x}_t$ and $\hat{\mathbf{x}}_t$ using mean-squared error.

The KL divergence loss $\mathcal{L}_{\text{KLD}}$ serves dual theoretical purposes. It enforces the conditional independence assumption from Theorem 2 through the factorized prior $p_\phi(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1})$ by Eq. 8, while simultaneously satisfying the sufficient variability assumption from Theorem 1 by encouraging diverse transitions in the latent space. When computing $\mathcal{L}_{\text{KLD}}$, we follow Yao et al. (2022a); Chen et al. (2024) employ a sampling method since the prior distribution lacks an explicit form.

The sparsity regularization implements the support sparsity of intermittent sequences through three terms. The L1 norm on decoder Jacobian columns $J_{\hat{g},t}|_{2,1}$ enforces sparse mixing patterns. The L1 norm on transition Jacobian rows $||J_{\hat{f},t}|_{1,1}$ ensures sparse transitions from $\mathbf{z}_{t-1}$ to $\mathbf{z}_t$. The L1 norm on transition Jacobian columns $|J_{\hat{f},t}|_{2,1}$ maintains consistent sparsity structure. Following standard practice, we use these L1 norms to approximate the L0 norm for differentiability.

5 EXPERIMENTS

5.1 SYNTHETIC EXPERIMENTS

Experimental Setup To evaluate `InterLatent` ability to learn causal processes and identify latent variables in non-invertible scenarios, we conduct simulation experiments using random causal structures with specified sample and variable sizes. We generate synthetic dataset satisfying the identifiability assumptions outlined in Theorem 1 and 2 (details in Appendix D.1), considering both nonstationary ($\mathbf{s}_t$ varying across the sequence) and stationary ($\mathbf{s}_t$ constant throughout) settings.

For each setting, we generate three scenarios of sequences, resulting in six scenarios in total. Each scenario has a particular value of d_t^c. This design allows us to assess the performance of `InterLatent` under different complexities of missingness. The Mean Correlation Coefficient (MCC) serves as our evaluation metric, measuring latent factor recovery by computing absolute correlation coefficients between ground-truth and estimated latent variables. MCC scores range from 0 to 1, with higher values indicating better identifiability.

Results Figure 3 summarizes our main results on our simulations. We evaluate `InterLatent` against several state-of-the-art approaches in identifying time-series causal variables and representation learning, such as LEAP Yao et al. (2022b), TDRL Yao et al. (2022a) and CaRiNG Chen et al. (2024). Additionally, we include classic representation learning approaches, such as BetaVAE Higgins et al. (2016), i-VAE Khemakhem et al. (2020) and SlowVAE Klindt et al. (2020).

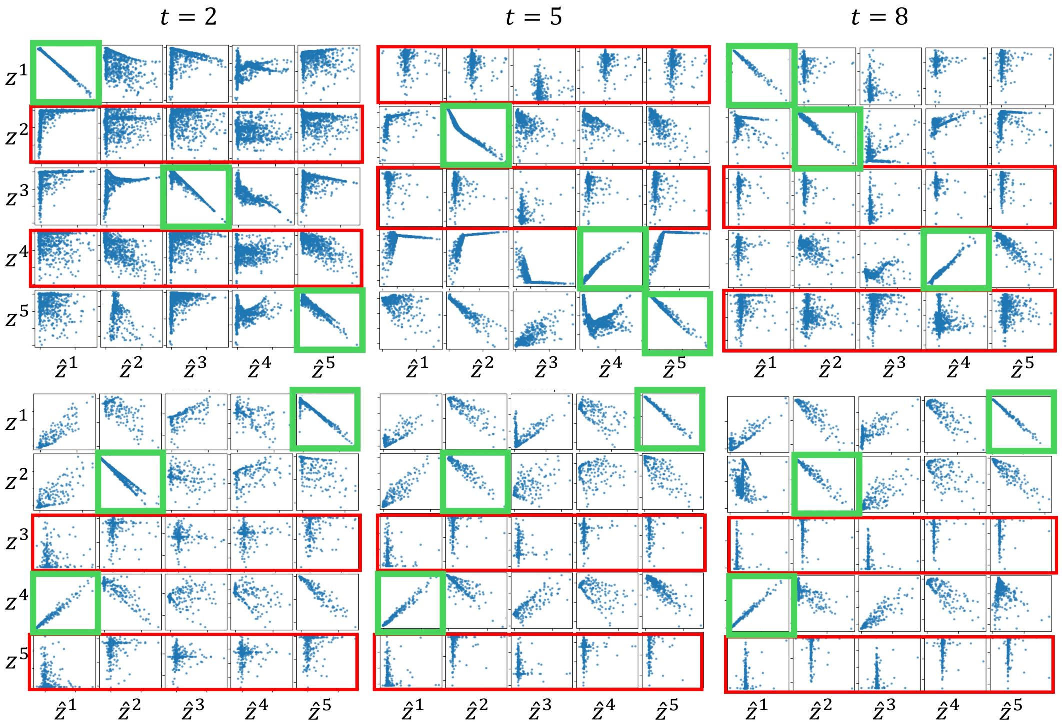


Figure 4: Visualization of the correlations between $\mathbf{z}_t$ and $\hat{\mathbf{z}}_t$ at time steps $t = 2, 5,$ and 8. The top row represents a scatter plot on a nonstationary sequence, while the bottom row depicts a scatter plot on a stationary sequence. The red bounding boxes depicts the missing part of $\mathbf{z}_t$, i.e, $\{\mathbf{z}^u | u \in \mathbf{s}_t^c\}$. The green bounding boxes highlight the latent variables that are component-wise identified for $\{\mathbf{z}_t^i | i \in \mathbf{s}_t\}$. The results confirm that `InterLatent` successfully identifies $\{\mathbf{z}_t^i | i \in \mathbf{s}_t\}$ in both nonstationary and stationary sequences. Also, we can observe that $\{\mathbf{z}^u | u \in \mathbf{s}_t^c\}$ is distinguishable from $\{\mathbf{z}_t^i | i \in \mathbf{s}_t\}$.

The results from Figure 3 demonstrate that `InterLatent` consistently achieves higher Mean Correlation Coefficient (MCC) across both nonstationary and stationary scenarios. For instance, in the nonstationary sequence with $d_t^c = 1$, `InterLatent` outperforms all other methods by a substantial margin, exceeding 0.1 in MCC. We attribute the superior performance of `InterLatent` to its capability of handling missingness in $\mathbf{z}_t$, a feature not present in the comparative methods. This key distinction enables our approach to more accurately capture the temporal dynamics of the latent variables. Figure 4 visualizes the disentanglement between the true latent variable and the estimations at different time steps from a sequence.

Ablation Study and Discussions To elucidate the key assumptions of our data generating process in Eq. 1, we further conduct ablation study focusing on the impact of sparse support. We introduce four baselines: (1) "W/O s of f", which removes sparsity regularization on transition functions; (2) "W/O s of g", which removes sparsity regularization on mixing functions; (3) "WS", a weakly supervised variant drops all sparse regularizations as having access to $\mathbf{s}_t$ and $\mathbf{s}_t^c$ during training; and (4) "ES", which estimates $\mathbf{s}_t$ and $\mathbf{s}_t^c$ using the gumbel softmax trick following the literature of structure learning Brouillard et al. (2020); Lorch et al. (2021).

We summarize our experimental results in Figure 3. InterLatent obtains the scores on par with "WS" baseline. This speaks the effectiveness of using the sparsity regularization terms against using $\mathbf{s}_t$ and $\mathbf{s}_t^c$ for g and f directly. Also, InterLatent advances "ES" baseline across all the settings. The "W/O S of f" baseline, which assumes $\mathbf{s}_t^c = \emptyset$, yields a significantly lower Mean Correlation Coefficient (MCC) compared to `InterLatent` approach. Similarly, "W/O S of g" fails to achieve competitive results due to its disregard for missingness in the mixing function. These outcomes confirm that without accounting for missing components, the baselines are unable to adequately model our simulated data. Experiments on various N can be found in Appendix D.4.

5.2 REAL-WORLD EXPERIMENTS

Task setup To evaluate our proposed identification theories in complex real-world scenarios, we apply them to the task of Group Activity Recognition (GAR) using the Volleyball dataset Ibrahim et al. (2016). GAR aims to categorize the activity for an individual frames in multi-actor scenes, aligning well with our scenario of intermittent temporal latent processes. This is because not all actors participate in every activity, reflecting real-world dynamics where some may be occluded or out of view in fast-evolving sporting scenarios. In our implementation, each actor at a given time point is modeled as a specific component of the latent variables, with occluded or out-of-view actors treated as ”missing” in the activity representations. This setup provides a solid testbed for our identification theory, allowing us to assess its robustness and effectiveness in handling real-world complexities such as partial observations and dynamic participant involvement.

Let $\mathbf{x} = \{\mathbf{x}_t^n\}_{t=1,n=1}^{T,N}$ denote a video consisting of T-frame observations and N agents. For each time step t and agent n, there exists a latent variable $\mathbf{z}_t^n \in \{\mathbf{z}_t^n\}_{t=1,n=1}^{T,N}$ that generates $\mathbf{x}_t^n$ according to Equation 1.

We take inspirations from the two-phase training pipeline from Li et al. (2024) to modify our training objective. First, we train `InterLatent` using the objective function defined in Eq. 10. Subsequently, a classifier $\hat{c}$ predicts the one-hot activity label $\hat{y}$ from the learned sequence of latent representations $\hat{\mathbf{z}}_{1:T}$ using an MLP: $\hat{y} = \text{MLP}(\text{Concat}(\hat{\mathbf{z}}_{1:T}))$. The classifier is trained using a cross-entropy loss with a L1 regularization on its Jacobian:

$$\mathcal{L}_{\text{cls}}^{CE} = -\mathbb{E}_{\hat{y}}\big(\text{one-hot}(y) \cdot \log(\text{softmax}(\hat{y}))\big) + |J_{\hat{c}}|_{2,1} \tag{11}$$

where one-hot(y) denotes the one-hot embedding of the true activity label. More data preprocessing details can be found in Appendix D.2.

Data and Comparing Methods The Volleyball dataset Ibrahim et al. (2016) contains 55 video recordings of volleyball games and is split into 3493 training clips and 1337 testing clips. The center frame of each clip is annotated with one group activity label out of eight labels (i.e. right set, right spike, right pass, right winpoint, left set, left spike, left pass, and left winpoint).

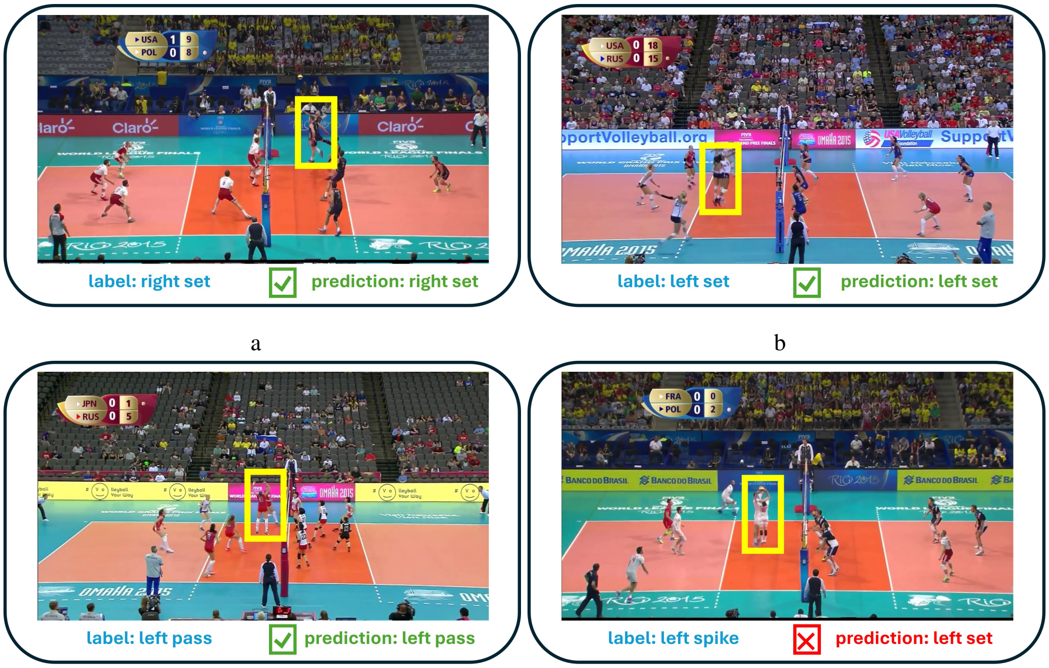
The comparing methods include the state-of-the-art methods on GAR task, such as SAM Yan et al. (2020), AT Gavrilyuk et al. (2020), ASACRF Pramono et al. (2020), DIN Yuan et al. (2021), DFGAR Kim et al. (2022), HiGCIN Yan et al. (2023), PAP Nakatani et al. (2024), and BiCausal Zhang et al. (2024). We also benchmark against TDRL Yao et al. (2022a) and CaRiNG Chen et al. (2024) to evaluate the efficacy of identifying the intermittent temporal latent process. For fair comparisons, `InterLatent` adopts the ResNet-18 backbone He et al. (2016) and weakly supervised setting from Yan et al. (2020) for feature extractions from the RGB frames, which is also commonly utilized by other approaches.

Methods	MCA
SACRF Pramono et al. (2020)	83.3
AT Gavrilyuk et al. (2020)	84.3
SAM Yan et al. (2020)	86.3
DIN Yuan et al. (2021)	86.5
DFGAR Kim et al. (2022)	90.5
HiGCIN Yan et al. (2023)	91.4
PAP Nakatani et al. (2024)	91.8
Dual-AI Han et al. (2022)	93.2
BiCausal Zhang et al. (2024)	93.4
TDRL Yao et al. (2022a)	92.9
CaRiNG Chen et al. (2024)	94.0
`InterLatent`	**95.7**

Table 1: **Comparison with the state-of-the-art methods on Volleball dataset**

Results and Discussions Table 1 presents a comparison of Multi-class Classification Accuracy (MCA) on the Volleyball dataset. Notably, `InterLatent` demonstrates superior performance with respect to those do not consider the missingness in both transition and mixing functions, i.e., CaRiNG and TDRL. For example, `InterLatent` achieves the highest accuracy of 95.7, significantly surpassing the previous best result of 94.0 obtained by CaRiNG. `InterLatent` also outperforms the state-of-the-art approaches on GAR, such as Dual-AI and BiCausal by significant margins of 2.5 and 2.3 points, respectively.

Figure 5 illustrates the visual examples of activity classification outcomes produced by `InterLatent` . The model demonstrates robust performance in handling occlusion-induced missingness, accurately categorizing activities in challenging scenarios. Figure 5a∼ 5c showcase successful classifications of “right set”, “left set” and “left pass” respectively, despite partial occlusions



a b

c d

Figure 5: The visual examples of **`InterLatent`** on Volleyball dataset. Highlighted frames show the annotated activity, with yellow bounding boxes indicating occluded actors. **`InterLatent`** correctly predicts the three activities but misclassifies a video of left spike" as left set". Note that the spike activity is performed by a actor that is severely occluded. This implies the misclassification may stem from that the label itself is not grounded by the true process.

of key players. We also present a failure case in Figure 5d, where **`InterLatent`** misclassifies a "left spike" as a "left set." However, this misclassification stems from that the label itself is not grounded by the true process, since the spike activity is performed by a player that is severely occluded.

6 CONCLUSION

We establish a set of novel identifiability results for intermittent latent temporal processes, extending the identifiability theory to scenarios where latent factors may be missing or inactive at different time steps. Specifically, we prove block-wise identifiability under assumptions on support sparsity, and further demonstrate component-wise identifiability within the support given conditional independence assumption. These results hold for both nonstationary and stationary transitions, accommodating a wide range of real-world temporal dynamics. Our theoretical findings are validated through experiments on both synthetic and real-world datasets, demonstrating the practical applicability of our approach. Future work could explore the application of this framework to related tasks such as temporal disentanglement, transfer learning in time series data, and causal discovery in dynamic systems. The identifiability guarantees we've established for intermittent latent temporal processes strengthen our ability to uncover hidden truths in diverse real-world settings, potentially impacting fields ranging from neuroscience to economics. While we have demonstrated the effectiveness of our approach on visual group activity recognition task, the lack of other applications is a limitation of this work.

REFERENCES

Elad Ben Avraham, Roei Herzig, Karttikeya Mangalam, Amir Bar, Anna Rohrbach, Leonid Karlinsky, Trevor Darrell, and Amir Globerson. Bringing image scene structure to video via frame-clip consistency of object tokens. *Advances in Neural Information Processing Systems*, 35:26839–26855, 2022.

Carlo Berzuini, Philip Dawid, and Luisa Bernardinell. *Causality: Statistical perspectives and applications*. John Wiley & Sons, 2012.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877, 2020.

Guangyi Chen, Yifan Shen, Zhenhao Chen, Xiangchen Song, Yuewen Sun, Weiran Yao, Xiao Liu, and Kun Zhang. Caring: Learning temporal causal representation under non-invertible generation process. *arXiv preprint arXiv:2401.14535*, 2024.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

Martin Danelljan, Gustav Häger, Fahad Khan, and Michael Felsberg. Accurate scale estimation for robust visual tracking. In *British machine vision conference, Nottingham, September 1-5, 2014*. Bmva Press, 2014.

Marco Fumero, Florian Wenzel, Luca Zancato, Alessandro Achille, Emanuele Rodolà, Stefano Soatto, Bernhard Schölkopf, and Francesco Locatello. Leveraging sparse and shared feature activations for disentangled representation learning. *Advances in Neural Information Processing Systems*, 36:27682–27698, 2023.

Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees GM Snoek. Actor-transformers for group activity recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 839–848, 2020.

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

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pp. 5842–5850, 2017.

Mingfei Han, David Junhao Zhang, Yali Wang, Rui Yan, Lina Yao, Xiaojun Chang, and Yu Qiao. Dual-ai: Dual-path actor interaction learning for group activity recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2990–2999, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2016.

Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. 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.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pp. 2207–2217. PMLR, 2020.

Dongkeun Kim, Jinsung Lee, Minsu Cho, and Suha Kwak. Detector-free weakly supervised group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20083–20093, 2022.

Manjin Kim, Paul Hongsuck Seo, Cordelia Schmid, and Minsu Cho. Learning correlation structures for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18941–18951, 2024.

David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*, 2020.

Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial disentanglement for domain adaptation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11455–11472. PMLR, 17–23 Jul 2022.

Lingjing Kong, Guangyi Chen, Biwei Huang, Eric P Xing, Yuejie Chi, and Kun Zhang. Learning discrete concepts in latent hierarchical models. *arXiv preprint arXiv:2406.00519*, 2024.

Sébastien Lachapelle, Tristan Deleu, Divyat Mahajan, Ioannis Mitliagkas, Yoshua Bengio, Simon Lacoste-Julien, and Quentin Bertrand. Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning. In *International Conference on Machine Learning*, pp. 18171–18206. PMLR, 2023.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. *arXiv preprint arXiv:2401.04890*, 2024a.

Sébastien Lachapelle, Divyat Mahajan, Ioannis Mitliagkas, and Simon Lacoste-Julien. Additive decoders for latent variables identification and cartesian-product extrapolation. *Advances in Neural Information Processing Systems*, 36, 2024b.

Dongho Lee, Jongseo Lee, and Jinwoo Choi. Cast: cross-attention in space and time for video action recognition. *Advances in Neural Information Processing Systems*, 36, 2024.

Yingzhen Li and Stephan Mandt. Disentangled sequential autoencoder. *arXiv preprint arXiv:1803.02991*, 2018.

Yuke Li, Guangyi Chen, Ben Abramowitz, Stefano Anzellotti, and Donglai Wei. Learning causal domain-invariant temporal dynamics for few-shot action recognition. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=LvuuYqU0BW`.

Zijian Li, Ruichu Cai, Guangyi Chen, Boyang Sun, Zhifeng Hao, and Kun Zhang. Subspace identification for multi-source domain adaptation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=BACQLWQW8u`.

Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause. Dibs: Differentiable bayesian structure learning. In *NeurIPS*, pp. 24111–24123, 2021. URL `https://proceedings.neurips.cc/paper/2021/hash/ca6ab34959489659f8c3776aaf1f8efd-Abstract.html`.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Hiroshi Morioka and Aapo Hyvarinen. Causal representation learning made identifiable by grouping of observational variables. In *Forty-first International Conference on Machine Learning*, 2024.

Chihiro Nakatani, Hiroaki Kawashima, and Norimichi Ukita. Learning group activity features through person attribute prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18233–18242, 2024.

Ignavier Ng, Yujia Zheng, Xinshuai Dong, and Kun Zhang. On the identifiability of sparse ica without assuming non-gaussianity. *Advances in Neural Information Processing Systems*, 36:47960–47990, 2023.

Rizard Renanda Adhi Pramono, Yie Tarng Chen, and Wen Hsien Fang. Empowering relational network by self-attention augmented conditional random fields for group activity recognition. In *European Conference on Computer Vision*, pp. 71–90. Springer, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

Xiangchen Song, Weiran Yao, Yewen Fan, Xinshuai Dong, Guangyi Chen, Juan Carlos Niebles, Eric Xing, and Kun Zhang. Temporally disentangled representation learning under unknown nonstationarity. *NeurIPS*, 2023.

Xiangchen Song, Zijian Li, Guangyi Chen, Yujia Zheng, Yewen Fan, Xinshuai Dong, and Kun Zhang. Causal temporal representation learning with nonstationary sparse transition. *arXiv preprint arXiv:2409.03142*, 2024.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36, 2024.

Thaddäus Wiedemer, Prasanna Mayilvahanan, Matthias Bethge, and Wieland Brendel. Compositional generalization from first principles. *Advances in Neural Information Processing Systems*, 36, 2024.

Danru Xu, Dingling Yao, Sébastien Lachapelle, Perouz Taslakian, Julius von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. *arXiv preprint arXiv:2403.08335*, 2024.

Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Social adaptive module for weakly-supervised group activity recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pp. 208–224. Springer, 2020.

Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Higcin: Hierarchical graph-based cross inference network for group activity recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 45(06):6955–6968, 2023.

Dingling Yao, Danru Xu, Sebastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *The Twelfth International Conference on Learning Representations*, 2024.

Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. *Advances in Neural Information Processing Systems*, 35:26492–26503, 2022a.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. In *International Conference on Learning Representations*, 2022b.

Hangjie Yuan, Dong Ni, and Mang Wang. Spatio-temporal dynamic inference network for group activity recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7476–7485, 2021.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. In *Forty-first International Conference on Machine Learning*.

Youliang Zhang, Wenxuan Liu, Danni Xu, Zhuo Zhou, and Zheng Wang. Bi-causal: Group activity recognition via bidirectional causality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1450–1459, 2024.

Yujia Zheng and Kun Zhang. Generalizing nonlinear ica beyond structural sparsity. *Advances in Neural Information Processing Systems*, 36, 2024.

Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ica: Sparsity and beyond. *Advances in neural information processing systems*, 35:16411–16422, 2022.

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, 25(60):1–8, 2024. URL `http://jmlr.org/papers/v25/23-0970.html`.

A NOTIONS

Table of notions

Data and estimations			
$\mathbf{x}_t \in \mathbb{R}^K$	Observations	$\hat{\mathbf{x}}_t \in \mathbb{R}^K$	Reconstructions
$\mathbf{z}_t \in \mathbb{R}^N$	Latent variables	$\hat{\mathbf{z}}_t \in \mathbb{R}^N$	Latent variable estimations
ϵ_t	True noise term	$\hat{\epsilon}_t$	Estimated noise term
$\mathbf{s}_t$	The support of $\mathbf{z}_t$	$\hat{\mathbf{s}}_t$	The support of $\hat{\mathbf{z}}_t$
$\mathbf{s}_t^c$	The missingness of $\mathbf{z}_t$	$\hat{\mathbf{s}}_t^c$	The missingness of $\hat{\mathbf{z}}_t$
d_t	The cardinality of $\mathbf{s}_t$	$\hat{d}_t$	The cardinality of $\hat{\mathbf{s}}_t$
d_t^c	The cardinality of $\mathbf{s}_t^c$	$\hat{d}_t^c$	The cardinality of $\hat{\mathbf{s}}_t^c$
Indices			
$t \in [1, T]$	Time step	$n \in [N]$	The indices of $\mathbf{z}_t$
$i \in \mathbf{s}_t$	The indices of $\mathbf{z}_t$ within $\mathbf{s}_t$	$j \in \hat{\mathbf{s}}_t$	The indices of $\hat{\mathbf{z}}_t$ within $\hat{\mathbf{s}}_t$
$u \in \mathbf{s}_t^c$	The indices of $\mathbf{z}_t$ within $\mathbf{s}_t^c$	$v \in \hat{\mathbf{s}}_t^c$	The indices of $\hat{\mathbf{z}}_t$ within $\hat{\mathbf{s}}_t^c$
Ground-truth $\&$ Learned model			
g	True mixing function	$\hat{g}$	Learned mixing function
f	True transition function	$\hat{f}$	Learned transition function
J_g	Jacobian of g	$J_{\hat{g}}$	Jacobian of $\hat{g}$
J_f	Jacobian of f	$J_{\hat{f}}$	Jacobian of $\hat{f}$
Optimizations			
ϕ	True parameters of f	$\hat{\phi}$	Learned parameters of $\hat{f}$
γ	True parameters of g	$\hat{\gamma}$	Learned parameters of $\hat{g}$
ω	True parameters of encoder	$\hat{\omega}$	Learned parameters of encoder
$\lvert * \rvert_{2,1}$	L1 norm on columns of $*$	$\lvert * \rvert_{1,1}$	L1 norm on rows of $*$

B PROOF OF THEOREM1 AND THEOREM2

In this section, we provide proof of our identifiability results in Theorem1 and Theorem2. To this end, we take insipirations from Lemma B.1 of Lachapelle et al. (2023) to present a Lemma that is throughout our proof.

Lemma 1 (Invertible matrix contains a permutation) Let $L \in \mathbb{R}^{N\times N}$ be an invertible matrix. Then, there exists a permutation σ such that $L^{n,\sigma(n)} \neq 0$ for all n, In other words, $\mathcal{P}^\intercal \subset L$ where $\mathcal{P}$ is the permutation matrix associated with σ, i.e. $\mathcal{P}e^n = e^{\sigma(n)}$.

Proof: Since the matrix L is invertible, its determinant is nonzero. We can obtain the following with the assistance of Leibniz formula:

$$|\det L| = \sum_{\sigma\in\mathcal{S}} \text{sign}(\sigma) \prod_i L^{n,\sigma(n)} \neq 0 \tag{12}$$

where $\mathcal{S}$ denotes a set of permutation. Eq. 12 suggests that at least one term of the sum is non-zero, meaning there exists $\sigma \in \mathcal{S}$, such that $\forall n, L^{n,\sigma(n)} \neq 0$.

B.1 PROOF OF THEOREM 1

Theorem 1 is where most of the theoretical contribution of this work lies. Let us recall Theorem 1:

Theorem 1 (Block-wise identifiability): *For the observations $\mathbf{x}_t \in \mathbb{R}^K$ and estimated latent variables $\hat{\mathbf{z}}_t \in \mathbb{R}^N$, suppose there exist functions $\hat{g}$ satisfying observational equivalence in Eq. 4 If the following assumptions and regularization hold:*

- *i (Smoothness and positivity): The probability density function of the latent causal variables, $p(\mathbf{z}_t)$, has positive measure in the space of $\mathbf{z}_t$ and is twice continuously differentiable.*
- *ii (Path connected): For any $\mathbf{z}^0, \mathbf{z}^1 \in \mathcal{Z}$, there is a continous function $\phi : [0,1] \to \mathcal{Z}$, s.t. $\phi(0) = \mathbf{z}^0$ and $\phi(1) = \mathbf{z}^1$.*
- *iii (Sufficient variability of $\mathbf{z}_t$ and $\hat{\mathbf{z}}_t$): Let $q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \log p(\mathbf{z}_t|\mathbf{z}_{t-1})$, as well as $q(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}) = \log p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1})$, and $\mathcal{H}_{\mathbf{z}_t,\mathbf{z}_{t-1}} q(\mathbf{z}_t|\mathbf{z}_{t-1})$ denotes the Hessian matrix of $q(\mathbf{z}_t|\mathbf{z}_{t-1})$ w.r.t. $\mathbf{z}_t$ and $\mathbf{z}_{t-1}$. Suppose $G^{\mathbf{z}_t} \in \{0,1\}^{N\times N}$ as a binary adjacency matrix that indicates the existence of the transitions from $\mathbf{z}_{t-1}$ and $\mathbf{z}_t$. $G^{\mathbf{z}_t}_{i1,i2} = 1$ means that there exists a transition from $\mathbf{z}^{i1}_{t-1}$ to $\mathbf{z}^{i2}_t$. We assume that:*

  $$span\{\mathcal{H}_{\mathbf{z}_t,\mathbf{z}_{t-1}} q(\mathbf{z}_t|\mathbf{z}_{t-1})\}_{i=1}^{d_t} = \mathbb{R}^{d_t\times d_t}_{G^{\mathbf{z}_t}}$$

  and

  $$span\{\mathcal{H}_{\hat{\mathbf{z}}_t,\hat{\mathbf{z}}_{t-1}} q(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1})\}_{j=1}^{\hat{d}_t} = \mathbb{R}^{\hat{d}_t\times \hat{d}_t}_{\hat{G}^{\hat{\mathbf{z}}_t}}$$

- *iv (Support sparsity regularization): For any time step t, $\mathbf{s}_t$ is not an empty set, $\hat{d}_t \leq d_t$*

There exists a permutation σ, such that

$$\hat{\mathbf{s}}_t = \sigma(\mathbf{s}_t) \quad and \quad \hat{\mathbf{s}}^c_t = \sigma(\mathbf{s}^c_t)$$

In other words, both $\forall i$, $\{\mathbf{z}^i_t | i \in \mathbf{s}_t\}$ and $\forall u$, $\{\mathbf{z}^u_t | u \in \mathbf{s}^c_t\}$ are block-wise identifiable.

Proof: Taking inspirations from Zheng & Zhang (2024), applying the chain rule to our definition in Definition 1 leads to:

$$\mathbf{x}_t = \hat{\mathbf{x}}_t \implies g(\mathbf{z}_t) = g(h^{-1}(\hat{\mathbf{z}}_t)) = \hat{g}(\hat{\mathbf{z}}_t) \implies J_g(\mathbf{z}_t)\cdot J_{h^{-1}}(\hat{\mathbf{z}}_t) = J_{\hat{g}}(\hat{\mathbf{z}}_t), \tag{13}$$

where J_g, $J_{h^{-1}}$, and $J_{\hat{g}}$ denote the Jacobian matrices of g, h^{-1}, and $\hat{g}$, respectively. Eq. 13 provides a rigorous definition of observational equivalence in the context of intermittent temporal latent processes, establishing the relationship between the true and learned models through their distributions and Jacobians.

We follow Yao et al. (2022b;a); Song et al. (2023); Chen et al. (2024); Song et al. (2024) to connect $J_{h^{-1}}(\hat{\mathbf{z}}_t)$ in Eq. 13 with the transition probability density function $p(\mathbf{z}_t|\mathbf{z}_{t-1})$ as we work on identification over the time-series data. Given the fact that $p(\mathbf{z}_t|\mathbf{z}_{t-1}) = p(\mathbf{z}_t|g(\mathbf{z}_{t-1})) = p(\mathbf{z}_t|\mathbf{x}_{t-1})$ as well as $p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}) = p(\hat{\mathbf{z}}_t|g(\hat{\mathbf{z}}_{t-1})) = p(\mathbf{z}_t|\mathbf{x}_{t-1})$, we are able to map $(\mathbf{x}_{t-1}, \mathbf{z}_t)$ to $(\mathbf{x}_{t-1}, \hat{\mathbf{z}}_t)$ with the jaocbian $\begin{pmatrix} I & 0 \\ 0 & J_h(\mathbf{z}_t) \end{pmatrix}$:

$$p(\mathbf{z}_t|\mathbf{x}_{t-1}) = p(\hat{\mathbf{z}}_t|\hat{\mathbf{x}}_{t-1})|\det J_h(\mathbf{z}_t)| \Rightarrow p(\mathbf{z}_t|\mathbf{z}_{t-1}) = p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1})|det J_h(\mathbf{z}_t)|, \tag{14}$$

where h is an invertible mapping, such that $\hat{\mathbf{z}}_t = h(\mathbf{z}_t)$. $|\det J_h(\mathbf{z}_t)|$ denotes the determinant of $h(\mathbf{z}_t)$.

Taking the logarithm on both sides of Eq. 14, we have:

$$\log p(\mathbf{z}_t|\mathbf{z}_{t-1}) - \log|\det J_h(\mathbf{z}_t)| = \log p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}). \tag{15}$$

We replace $q(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}) = \log p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}))$ as well as $q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \log p(\mathbf{z}_t|\mathbf{z}_{t-1})$, and calculate the Hessian with $\hat{\mathbf{z}}_t$ and $\hat{\mathbf{z}}_{t-1}$ on both sides of Eq. 15 using change-of-variable and chain rule:

$$\mathcal{H}_{\hat{\mathbf{z}}_t,\hat{\mathbf{z}}_{t-1}} q(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}) = (J_{h^{-1}}(\hat{\mathbf{z}}_t))^\intercal \mathcal{H}_{\mathbf{z}_t,\mathbf{z}_{t-1}} q(\mathbf{z}_t|\mathbf{z}_{t-1}) J_{h^{-1}}(\hat{\mathbf{z}}_{t-1}). \tag{16}$$

We can rewrite Eq. 16 based on *assumption iii* by:

$$span\{\mathcal{H}_{\hat{\mathbf{z}}_t,\hat{\mathbf{z}}_{t-1}} q(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1})\}_{j=1}^{\hat{d}_t} = (J_{h^{-1}}(\hat{\mathbf{z}}_t))^\intercal span\{\mathcal{H}_{\mathbf{z}_t,\mathbf{z}_{t-1}} q(\mathbf{z}_t|\mathbf{z}_{t-1})\}_{i=1}^{d_t} J_{h^{-1}}(\hat{\mathbf{z}}_{t-1}), \tag{17}$$

where $\circ$ denotes the Hadmard product.

For any $i1, i2 \in \mathbf{s}_t, (i1 \neq i2)$ we can obtain the standrad one-hot basis vector e_{i1} and e_{i2} of Hessian matrix, such that $e_{i1}(e_{i2})^\intercal \in \mathbb{R}^{d_t \times d_t}_{G^{\mathbf{z}_t}}$. Eq. 17 indicates the existence of a permutation matrix $\mathcal{P}$ associated with the permutation σ of $J_{h^{-1}}$ (Lemma 1), such that:

$$\mathcal{P}^\intercal e_{i1} \circ (e_{i2})^\intercal \mathcal{P} = e_{\sigma(i1)}(e_{\sigma(i2)})^\intercal \subseteq \mathbb{R}^{\hat{d}_t \times \hat{d}_t}_{\hat{G}^{\hat{z}_t},} \tag{18}$$

which implies:

$$(\sigma(i1), \sigma(i2)) \in \mathcal{H}_{\hat{\mathbf{z}}_t, \hat{\mathbf{z}}_{t-1}} q(\hat{\mathbf{z}}_t | \hat{\mathbf{z}}_{t-1}). \tag{19}$$

The support sparsity constraint suggests that:

$$\hat{d}_t = |\mathcal{H}_{\hat{\mathbf{z}}_t, \hat{\mathbf{z}}_{t-1}} q(\hat{\mathbf{z}}_t | \hat{\mathbf{z}}_{t-1})|_{1,0} \leq |\mathcal{H}_{\mathbf{z}_t, \mathbf{z}_{t-1}} q(\mathbf{z}_t | \mathbf{z}_{t-1})|_{1,0} = d_t$$

where $| * |_{1,0}$ denotes the ℓ_0-norm of the rows of matrix $*$.

Combining this with Equation 19, we can conclude that:

$$\hat{s}_t = \sigma(s_t). \tag{20}$$

If this were not the case, there would exist a pair $(i_1', i_2') \in G^{\mathbf{z}_t}$, where $i_1' \neq i_2'$, that contradicts Equation 19.

Using a similar strategy, combing Lemma 1 and Eq. 20, we can also conclude that, $\forall u \in \mathbf{s}_t^c, \sigma(u) \subset \hat{s}_t^c$. This leads to:

$$\hat{\mathbf{s}}_t^c = \sigma(\mathbf{s}_t^c). \tag{21}$$

If Eq. 21 is unstatisfied, h cannot be invertible, which contradicts with h being a invertible mapping.

Eq. 20 and Eq. 21 suggest that, $\forall i \in \mathbf{s}_t$ and $v \in \hat{\mathbf{s}}_t^c$:

$$\frac{\partial \mathbf{z}_t^i}{\partial \hat{\mathbf{z}}_t^v} = 0 \tag{22}$$

Thus, we reach the conclusion of block-wise identification.

Generalize from $\mathbf{z}_{t-1}$ to $\mathbf{z}_{<t}$: Notably we can easily generalize Theorem 1 by replacing $\mathbf{z}_{t-1}$ with $\mathbf{z}_{<t}$, and $\hat{\mathbf{z}}_{t-1}$ with $\hat{\mathbf{z}}_{<t}$ from Eq. 14 to Eq. 17, respectively. In other words, we can extend the conditional probability density function into a non-markov setting. Accordingly, $G^{\mathbf{z}_t} \in \{0,1\}^{d_t \times d_t(t-1)}$ and $G^{\hat{\mathbf{z}}_t} \in \{0,1\}^{\hat{d}_t \times \hat{d}_t(t-1)}$, if both $\mathbf{z}_{<t}$ and $\hat{\mathbf{z}}_{<t}$ start from $t = 1$ in Eq. 14.

B.2 PROOF OF THEOREM 2

Theorem 1 allows us to further explore the identifiability $\forall i, \{\mathbf{z}_t^i | i \in \mathbf{s}_t\}$. In what follows, we provide the proof of Theorem 2 in details.

Theorem 2 (Component-wise identifiability $\forall i, \{\mathbf{z}_t^i | i \in \mathbf{s}_t\}$): *Let all assumptions from the Theorem hold. Additionally, suppose the following assumption is true for data generating process in Eq. 1 as well:*

i (Conditional independence): At t, we assume that each component of $\mathbf{z}_t$ is conditional independent given the previous latent variables $\mathbf{z}_{t-1}$. For any $i1, i2 \in [N]$:

$$\mathbf{z}_t^{i1} \perp\!\!\!\perp \mathbf{z}_t^{i2} | \mathbf{z}_{t-1}$$

Then for $\{\hat{\mathbf{z}}_t^j | j \in \hat{\mathbf{s}}_t\}$ must be a component-wise transformation of a permuted version of true $\{\mathbf{z}_t^i | i \in \mathbf{s}_t\}$.

Proof: Following previous works Zheng & Zhang (2024), our goal can be rewritten as demonstrating that $J_{h^{-1}}(\hat{\mathbf{z}}_t) = \mathcal{D}(\hat{\mathbf{z}}_t)\mathcal{P}$, where $\mathcal{D}$ denotes an diagonal matrix. $\mathcal{P}$ is a permutation matrix that is defined in Lemma 1, and has been proven in Theorem 1. If $J_{h^{-1}}(\hat{\mathbf{z}}_t) \neq \mathcal{D}(\hat{\mathbf{z}}_t)\mathcal{P}$, there must exist $i1$

and $i2$ ($i1 \neq i2$), such that $j1, j2 \in J^{:,i1}_{g^{-1}}$, and $j2 \in J^{:,i2}_{h^{-1}}$. $J^{:,i1}_{h^{-1}}$ is the $i1$-th column of $J_{h^{-1}}$, which corresponds to $\mathbf{z}^{i1}_t$. Similarly, $J^{:,i2}_{h^{-1}}$ corresponds to $\mathbf{z}^{i2}_t$. Given Eq. 16, we can obtain:

$$\hat{\mathbf{z}}^{j1}_t, \hat{\mathbf{z}}^{j2}_t \in J^{:,i1}_{h^{-1}}(\hat{\mathbf{z}}_t)^\intercal \mathcal{H}_{\mathbf{z}_t,\mathbf{z}_{t-1}} q(\mathbf{z}_t|\mathbf{z}_{t-1}) J_{h^{-1}}(\hat{\mathbf{z}}_{t-1}). \tag{23}$$

Also,

$$\hat{\mathbf{z}}^{j2}_t \in J^{:,i2}_{h^{-1}}(\hat{\mathbf{z}}_t)^\intercal \mathcal{H}_{\mathbf{z}_t,\mathbf{z}_{t-1}} q(\mathbf{z}_t|\mathbf{z}_{t-1}) J_{h^{-1}}(\hat{\mathbf{z}}_{t-1}). \tag{24}$$

Therefore, $\mathbf{z}^{i1}_t$ and $\mathbf{z}^{i2}_t$ are dependent as given they are both dependent on $\hat{\mathbf{z}}^{j2}_t$. This contradicts with our conditional independence assumption.

B.3 EXTENSION OF TEMPORAL SUPPORT SPARSITY

Proosition 1(Identifiability under temporal support sparsity): *In addition to the assumptions of Theorem 1, if the following assumption and regularization hold:*

- *i (positivity and independence of the support): For any time step t, there exists the probability density function of the support, $p(\mathbf{s}_t)$, has positive measure in the space of $\mathbf{s}_t$. The support at any time step t is independent of the supports at other time steps, thus can be factorized by $p(\mathbf{s}_{1:T}) = \prod_{t=1}^{T} \mathbf{s}_t$*
- *ii (temporal support sparsity regularization): For any time step t, $\mathbf{s}_t$ is not an empty set, $\mathbb{E}(\hat{d}_{1:T}) \leq \mathbb{E}(d_{1:T})$*

There exists a permutation σ, such that

$$\hat{\mathbf{s}}_t = \sigma(\mathbf{s}_t) \quad \textit{and} \quad \hat{\mathbf{s}}^c_t = \sigma(\mathbf{s}^c_t)$$

In other words, both $\forall i$, $\{\mathbf{z}^i_t | i \in \mathbf{s}_t\}$ and $\forall u$, $\{\mathbf{z}^u_t | u \in \mathbf{s}^c_t\}$ are block-wise identifiable.

Proof:

We start from Eq. 19 in Theorem 1 to prove the Theorem 3. Let $\mathcal{H}_t = \mathcal{H}_{\mathbf{z}_t,\mathbf{z}_{t-1}} q(\mathbf{z}_t|\mathbf{z}_{t-1})$, and $\hat{\mathcal{H}}_t = \mathcal{H}_{\mathbf{z}_t,\mathbf{z}_{t-1}} q(\mathbf{z}_t|\mathbf{z}_{t-1})$, the expected sparsity constraint can be reformulated by:

$$\begin{aligned}
\mathbb{E}(d_{1:T}) = \mathbb{E}|\mathcal{H}_{1:T}|_{1,0} = \mathbb{E}_{p(\mathbf{s}_{1:T})}\mathbb{E}(\sum_{n=1}^{N} \mathbb{1}(\mathcal{H}^{n,:}_t \neq 0)|\mathbf{s}_t) \\
= \mathbb{E}_{p(\mathbf{s}_{1:T})} \sum_{n=1}^{N} \mathbb{E}(\mathbb{1}(\mathcal{H}^{n,:}_t \neq 0)|\mathbf{s}_t) \\
= \mathbb{E}_{p(\mathbf{s}_{1:T})}(\sum_{n=1}^{N} \mathbb{P}_{\mathcal{H}_t|\mathbf{s}_t}(\mathcal{H}^{n,:}_t \neq 0))
\end{aligned} \tag{25}$$

where $\mathbb{1}(*)$ denotes the indicator function of $*$, $\mathbb{P}_{\mathcal{H}|\mathbf{s}_t}$ denotes the

Let $\mathcal{J} = J_{h^{-1}}(\hat{\mathbf{z}}_{t-1})$, and $\mathcal{J}^{-1} = J_{h^{-1}}(\hat{\mathbf{z}}_t)$. We can perform the similar steps to obtain:

$$\begin{aligned}
\mathbb{E}(\hat{d}_{1:T}) = \mathbb{E}|\hat{\mathcal{H}}_{1:T}|_{1,0} = \mathbb{E}_{p(\mathbf{s}_{1:T})}\mathbb{E}(\sum_{n=1}^{N} \mathbb{1}(\mathcal{J}^{-1}\mathcal{H}^{n,:}_t \mathcal{J} \neq 0)|\mathbf{s}_t) \\
= \mathbb{E}_{p(\mathbf{s}_{1:T})} \sum_{n=1}^{N} \mathbb{E}(\mathbb{1}(\mathcal{J}^{-1}\mathcal{H}^{n,:}_t \mathcal{J} \neq 0)|\mathbf{s}_t) \\
= \mathbb{E}_{p(\mathbf{s}_{1:T})}(\sum_{n=1}^{N} \mathbb{P}_{\mathcal{H}_t|\mathbf{s}_t}(\mathcal{J}^{-1}\mathcal{H}^{n,:}_t \mathcal{J} \neq 0))
\end{aligned} \tag{26}$$

The temporal support sparsity constraint suggests that: $\mathbb{E}(\hat{d}_{1:T}) \leq \mathbb{E}(d_{1:T})$, which leads to

$$\begin{aligned} &\mathbb{E}_{p(\mathbf{s}_{1:T})}(\sum_{n=1}^{N} \mathbb{P}_{\mathcal{H}_t|\mathbf{s}_t}(\mathcal{J}^{-1}\mathcal{H}^{n,:}\mathcal{J} \neq 0)) - \mathbb{E}_{p(\mathbf{s}_{1:T})}(\sum_{n=1}^{N} \mathbb{P}_{\mathcal{H}_t|\mathbf{s}_t}(\mathcal{H}^{n,:} \neq 0)) \leq 0 \\ &= \mathbb{E}_{p(\mathbf{s}_{1:T})}(\sum_{n=1}^{N}(\mathbb{P}_{\mathcal{H}_t|\mathbf{s}_t}(\mathcal{J}^{-1}\mathcal{H}_t^{n,:}\mathcal{J}) - \mathbb{P}_{\mathcal{H}_t|\mathbf{s}_t}(\mathcal{H}_t^{n,:})) \leq 0 \end{aligned} \tag{27}$$

Eq. 19 suggests that $\forall n \in [1, N], \exists \sigma(n), s.t. \mathbb{P}_{\mathcal{H}_t|\mathbf{s}_t}(\mathcal{H}_t^{n,:}) = (\mathbb{P}_{\mathcal{H}_t|\mathbf{s}_t}(\mathcal{J}^{-1}\mathcal{H}_t^{\sigma(n),:}\mathcal{J})$, the L.H.S. of Eq. 27 is a sum of non-negative terms which is itself non-positive. This means that every term in the sum is zero. The rest of the proof remains the same with Theorem 1 to obtain the block-wise identifiability. Moreover, if the conditional independence assumption in Theorem 2 holds, we can further obtain the component-wise identifiability.

B.4 ESTIMATING $\mathbf{s}_t$ AND MODIFIED ELBO

The temporal support sparsity in Theorem 3 requires to obtain $p(\mathbf{s}_{1:T})$ for the identifiability results. In order to allow for gradient-based optimization of $\hat{\mathbf{s}}_{1:T}$, we take inspirations from the structure learning Brouillard et al. (2020); Lorch et al. (2021) to treat $\hat{\mathbf{s}}_t \cup \hat{\mathbf{s}}_t^c = \mathcal{S}$ as a $1 \times N$ vector. Each entries of this vector is a independent Bernoulli random variable with probability of success $\sigma(\alpha_n)$, where σ is the sigmoid function and α_n is a parameter learned using the Gumbel-Softmax trick. Accordingly, our ELBO needs to be modified as following:

$$\mathcal{L}_{\text{ELBO}} = \mathcal{L}_{\text{Recon}} + \mathcal{L}_{\text{KLD}}, \text{subject to } \mathbb{E}_{\mathcal{S}\sim\sigma(\alpha)}|\mathcal{S}| \leq \beta \tag{28}$$

where β is an hyperparameter (which should be set ideally to the true d_t) and $\mathcal{S} \sim \sigma(\alpha)$ means that each entry of $\mathcal{S}$ are independent and distributed according to $\sigma(\alpha)$. Comparing to Eq. 10, Eq. 28 drops the sparsity regularization terms as we use Gumbel-Softmax instead.

C IMPLEMENTATION DETAILS

C.1 PRIOR LIKELIHOOD DERIVATION

Consider a paradigmatic instance of latent causal processes. In this case, we are concerned with two time-delayed latent variables, namely, $\mathbf{z}_t = [\mathbf{z}_t^1, \mathbf{z}_t^2]$. We set time lag is defined as 1 for simplicity. This implies that each latent variable, $\mathbf{z}_t^n$, is formulated as $\mathbf{z}_t^n = f_n(\text{Pa}(\mathbf{z}_t^n), \epsilon_t^n)$, where $\text{Pa}(\mathbf{z}_t^n) \subset \mathbf{z}_{t-1}$ is the parent of $\mathbf{z}_t^n$. The noise terms, ϵ_t^n, are mutually independent. To represent this latent process more succinctly, we introduce a transformation map, denoted as f. It's worth noting that in this context, we employ an overloaded notation; specifically, the symbol f serves dual purposes, representing both transition functions and the transformation map.

$$\begin{bmatrix} \mathbf{z}_{t-1}^1 \\ \mathbf{z}_{t-1}^2 \\ \mathbf{z}_t^1 \\ \mathbf{z}_t^2 \end{bmatrix} = \mathbf{f}\left(\begin{bmatrix} \mathbf{z}_{t-1}^1 \\ \mathbf{z}_{t-1}^2 \\ \epsilon_t^1 \\ \epsilon_t^2 \end{bmatrix}\right) \tag{29}$$

By leveraging the change of variables formula on the map $\mathbf{f}$, we can evaluate the joint distribution of the latent variables $p(\mathbf{z}_{t-1}^1, \mathbf{z}_{t-1}^2, \mathbf{z}_t^1, \mathbf{z}_t^2)$ as:

$$p(\mathbf{z}_{t-1}^1, \mathbf{z}_{t-1}^2, \mathbf{z}_t^1, \mathbf{z}_t^2) = p(\mathbf{z}_{t-1}^1, \mathbf{z}_{t-1}^2, \epsilon_t^1, \epsilon_t^2) / \left|\det J_f\right|, \tag{30}$$

where J_f is the Jacobian matrix of the map $\mathbf{f}$, which is naturally a low-triangular matrix:

$$J_f = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{\partial \mathbf{z}_t^1}{\partial \mathbf{z}_{t-1}^1} & \frac{\partial \mathbf{z}_t^1}{\partial \mathbf{z}_{t-1}^2} & \frac{\partial \mathbf{z}_t^1}{\partial \epsilon_t^1} & 0 \\ \frac{\partial \mathbf{z}_t^2}{\partial \mathbf{z}_{t-1}^1} & \frac{\partial \mathbf{z}_t^2}{\partial \mathbf{z}_{t-1}^2} & 0 & \frac{\partial \mathbf{z}_t^2}{\partial \epsilon_t^2} \end{bmatrix}.$$

Given that this Jacobian is triangular, we can efficiently compute its determinant as $\prod_n \frac{\partial \mathbf{z}_t^n}{\partial \epsilon_t^n}$. Furthermore, because the noise terms are mutually independent, and hence $\epsilon_t^1 \perp \epsilon_t^2$, and $\epsilon_t \perp \mathbf{z}_{t-1}$, we can write Eq. 30 as:

$$\begin{aligned} p(\mathbf{z}_{t-1}^1, \mathbf{z}_{t-1}^2, \mathbf{z}_t^1, \mathbf{z}_t^2) &= p(\mathbf{z}_{t-1}^1, \mathbf{z}_{t-1}^2) \times p(\epsilon_t^1, \epsilon_t^2) / \left|\det J_f\right| \quad (\text{because } \epsilon_t \perp \mathbf{z}_{t-1}) \\ &= p(\mathbf{z}_{t-1}^1, \mathbf{z}_{t-1}^2) \times \prod_i p(\epsilon_t^i) / \left|\det J_f\right| \quad (\text{because } \epsilon_t^1 \perp \epsilon_t^2). \end{aligned} \tag{31}$$

Let $\{f_n^{-1}\}_{n=1,2,3...}$ be a set of learned inverse dynamics transition functions that take the estimated latent causal variables in the dynamics subspace and lagged latent variables, and output the noise terms, i.e., $\epsilon_t^n = f_n^{-1}\left(\mathbf{z}_t^n, \text{Pa}(\mathbf{z}_t^n)\right)$. By eliminating the marginals of the lagged latent variable $p(\mathbf{z}_{t-1}^1, \mathbf{z}_{t-1}^2)$ on both sides, we derive the total transition prior likelihood as:

$$p(\mathbf{z}_t^1, \mathbf{z}_t^2 | \mathbf{z}_{t-1}^1, \mathbf{z}_{t-1}^2) = \prod_n p(\epsilon_t^i) / \left|\det J_f\right| = \prod_n p(f_n^{-1}\left(\mathbf{z}_t^n, \text{Pa}(\mathbf{z}_t^n)\right)) \times \left|\det J_f^{-1}\right| \tag{32}$$

in which, $\forall i, \{\mathbf{z}_t^i | i \in \mathbf{s}_t\}$, the prior likelihood is:

$$p(\mathbf{z}_t^i | \mathbf{z}_{t-1}) = \prod_i p(\epsilon_t^i) / \left|\det J_f\right| = \prod_i p(f_i^{-1}\left(\mathbf{z}_t^i, \text{Pa}(\mathbf{z}_t^i)\right)) \times \left|\det J_f^{-1}\right|. \tag{33}$$

Then, $\forall u, \{\mathbf{z}_t^u | u \in \mathbf{s}_t^m\}$, given $\text{Pa}(\mathbf{z}_t^u) = \emptyset$, the prior likelihood is:

$$p(\mathbf{z}_t^u) = \prod_u p(\epsilon_t^u) / \left|\det J_f\right| = \prod_u p(f_u^{-1}(\mathbf{z}_t^u)) \times \left|\det J_f^{-1}\right|. \tag{34}$$

C.2 NETWORK ARCHITECTURES

Configuration	Description	Output dimensions
Encoder		
Input: concat($\mathbf{x}_{1:T}$)		BS $\times T \times K$
Dense	128 neurons, LeakyReLU	BS $\times T \times 128$
Dense	128 neurons, LeakyReLU	BS $\times T \times 128$
Dense	Temporal embeddings	BS $\times T \times 2N$
Bottleneck	Compute mean and variance of posterior	μ, σ
Reparameterization	Sequential sampling	$\hat{\mathbf{z}}_{1:T}$
Decoder		
Input: $\hat{\mathbf{z}}_{1:T}$		BS $\times T \times N$
Dense	128 neurons, LeakyReLU	BS $\times T \times 128$
Dense	128 neurons, LeakyReLU	BS $\times T \times 128$
Dense	input embeddings	BS $\times T \times K$
Temporal prior module		
Input	$\hat{\mathbf{z}}_{1:T}$	BS $\times T \times N$
InverseTransition	$\hat{\epsilon}_t$	BS $\times T \times N$
JacobianCompute	$\log\left|\det J_f\right|$	BS

Table 2: The details of our network architectures for `InterLatent` , where BS means batch size.

Table 2 summarizes the network architectures of `InterLatent` .

C.3 TRAINING DETAILS

Simulation Experiments

We implemented our models using PyTorch 1.11.0. For optimization, we employed the AdamW optimizer Loshchilov & Hutter (2019), which has been shown to improve generalization performance in deep learning models. The hyperparameters were set as follows: learning rate of 1e-3 and mini-batch size of 64. To ensure robustness and statistical significance, we trained each model under 10 different random seeds and report the overall performance as mean ± standard deviation across these runs. The loss function balances reconstruction error and KL-divergence, with the latter weighted by $\beta = 0.02$. This choice of β was determined through preliminary experiments to achieve an optimal trade-off between reconstruction quality and latent space regularity. All experiments were conducted on a single NVIDIA GeForce RTX 2080 Ti GPU with 11GB meory.

Real-World Experiments

We employ the AdamW optimizer with cosine annealing for training our network. The initial learning rate is set to 2e-3, with a weight decay of 1e-2 to mitigate overfitting. For all video sequences in Volleyball dataset, we uniformly sample $T = 10$ frames as input. The ELBO loss is computed with a β value of 0.02. We utilize a batch size of 128, which we found to provide a good trade-off between computational efficiency and optimization stability. The network is implemented using PyTorch [2], leveraging its dynamic computational graph and GPU acceleration capabilities. Training is conducted for 80 epochs on a multi-GPU setup consisting of four NVIDIA GeForce RTX 2080 Ti GPUs, providing a total of 44GB of meory.

D ADDITIONAL EXPERIMENTS DETAILS

D.1 SYNTHETIC DATA GENERATION PROCESS

Our approach generates six distinct scenarios of sequences, encompassing both stationary and non-stationary settings with varying degrees of missingness. Each sequence consists of 9 time steps, with latent variables $\mathbf{z}_t \in \mathbb{R}^5$ and observations $\mathbf{x}_t \in \mathbb{R}^5$. Missingness is introduced by selecting a constant value $d^c \in \{1, 2, 3\}$ for each sequence, representing the number of missing dimensions throughout that sequence. The set of missing dimensions, $\mathbf{s}_t^c$, is then determined based on d_t^c. We generate six scenarios in total: (1). non-stationary sequences with $d_t^c = 1$; (2). non-stationary sequences with $d_t^c = 2$; (3). non-stationary sequences with $d_t^c = 3$; (4). stationary sequences with $d_t^c = 1$; (5). stationary sequences with $d_t^c = 2$; (6). stationary sequences with $d_t^c = 3$. In non-stationary sequences, $\mathbf{s}_t^c$ varies every 3 time steps, while in stationary sequences, it remains fixed throughout.

For each scenario, the data generation process begins with 10,000 initial states drawn from $\mathbf{z}_0 \sim U(0, 1)$. From $t = 1$ to $t = 9$, $\mathbf{z}_t$ within $\mathbf{s}_t$ is generated using a nonlinear function f with non-additive, zero-biased Gaussian noise ϵ_t^i, where ($\sigma = 0.1$): $\forall i \in \mathbf{s}_t, \mathbf{z}_t^i = f_i(\mathbf{z}_{t-1}^{\{i'\}_{i'=1}^{d_t}}, \epsilon_t^i)$, where $\mathbf{z}_{t-1}^{\{i'\}_{i'=1}^{d_t}}$ is the set of $\mathbf{z}_{t-1}^{i'}$ within $\mathbf{s}_{t-1}$. The missing dimensions are set as $\forall u \in \mathbf{s}_t^c, \mathbf{z}_t^u = f_u(\epsilon_t)$. Observations are then generated using a mixing function g that only considers $\mathbf{z}_t$ within $\mathbf{s}_t$: $\mathbf{x}_t = g(\mathbf{z}_t^{\{i\}_{i=1}^{d_t}})$, where $\mathbf{z}_t^{\{i\}_{i=1}^{d_t}}$ is the set of $\mathbf{z}_t^i$ within $\mathbf{s}_t$.

D.2 ADDITOINAL DETAILS OF THE VOLLEYBALL DATASET

Our preprocessing and feature extraction pipeline builds upon the procedure from Yan et al. (2023). We leverage a pretrained Faster R-CNN model Ren et al. (2016) implemented via the MMDetection toolbox Chen et al. (2019) to detect potential persons in each frame. These detections are then tracked across frames using the method proposed by Danelljan et al. (2014). For feature extraction, we utilize ResNet-18 He et al. (2016). We apply RoIAlign He et al. (2017) with a crop size of 5×5. The resulting features are embedded into a $K = 1024$ vector. we select the top $N = 20$ person proposals based on detection confidence scores.

D.3 ADDITIONAL EXPERIMENTS ON VARIOUS d_t

In this section, we aim to understand the robustness of `InterLatent` under the settings with changing d_t. In particular, we synthesize another 10,000 sequence with identical procedure in pre-

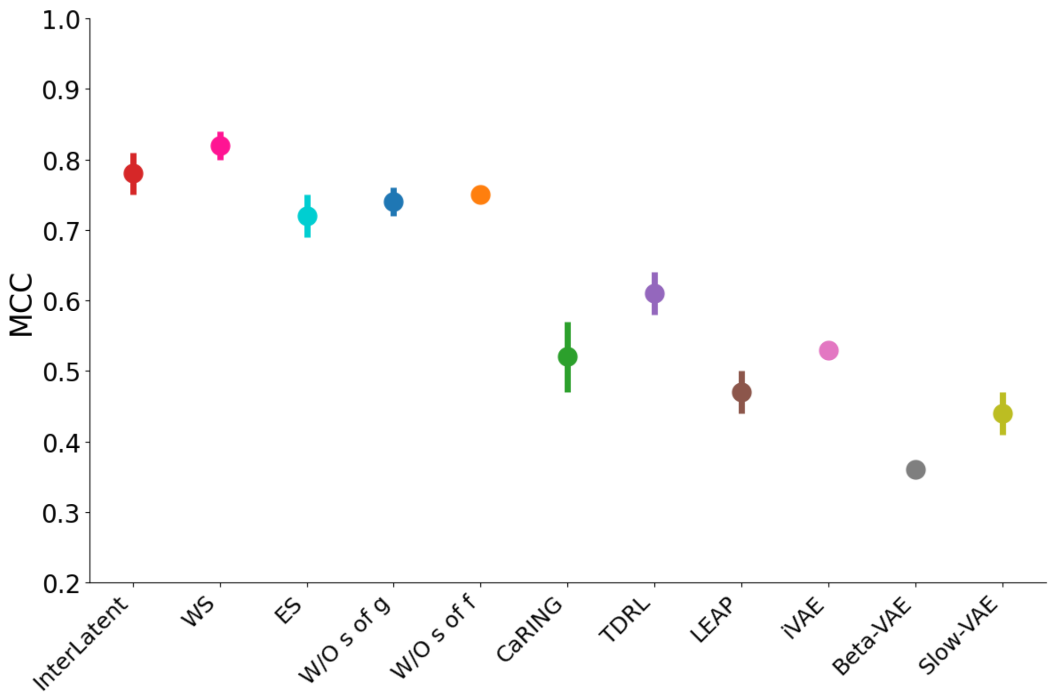


Figure 6: Mean Correlation Coefficient (MCC) scores for varying d_t. Higher MCC scores indicate better performance in identifying latent variables.

vious section. The only difference is that we introduce the missingness by choosing $d_t = 1, \ (t \in [1,5])$ and $d_{5:9} = 2, \ (t \in [5,9])$.

Figure 6 shows the outcomes of our experiments on the varying d_t. Despite WS having access to ground truth values of $\mathbf{s}_t$ and $\mathbf{s}_t^c$ across all time steps, InterLatent achieves comparable performance, ranking second best among all comparisons. This demonstrates the effectiveness of applying sparsity regularization to the Jacobians of both functions f and g in Eq. 10.

D.4 ADDITIONAL EXPERIMENTS ON VARIOUS N

To demonstrate our method's scalability, we extended our non-stationary sequence experiments to higher dimensions. We examined scenarios with latent dimensions $N \in \{8, 12\}$ and corresponding observation dimensions $K \in \{10, 16\}$, while maintaining $d_t^c = 1$.

	$N = 8$	$N = 12$
CaRiNG	0.574 ± 0.03	0.491 ± 0.02
`InterLatent`	**0.818 ± 0.01**	**0.655 ± 0.02**

Table 3: Ablation study results on the various N

Table 1 summarizes our ablation study findings. Our proposed model, **`InterLatent`** , consistently outperforms CaRiN across different values of N, the dimensionality of the latent variables. While we observe a general decline in performance as N increases and the complexity of data generating spikes, **`InterLatent`** demonstrates a bettercapability in identifying intermittent temporal latent processes compared to CaRiNG.

D.5 ADDITIONAL EXPERIMENTS ON ACTION RECOGNITION

We also test the effectiveness of InterLatent on action recognition task by using Something Something V2 dataset Goyal et al. (2017). Something-Something v2 (SSv2) is a dataset containing 174 action categories of common human-object interactions. It includes 220,847 videos, with 168,913 in the training set, 24,777 in the validation set and 27,157 in the test set. In each video sequence, there might be occlusion between human and object. Thus, this dataset plays a solid ground for our experiments. InterLatent adopts ViT-B/16 Radford et al. (2021) pretrained on as the backbone to obtain x_t. Regarding the hyperparameters, we set $N = 12$ in Eq. 1. Also, we use the same two-phase training strategy.

To evaluate the efficacy of identifying intermittent temporal latent processes, we benchmark InterLatent against both causal representation learning methods (TDRL Yao et al. (2022a), CaRiNG Chen et al. (2024)) and state-of-the-art action recognition approaches (SViT Ben Avraham et al. (2022), VideoMAE Tong et al. (2022), CAST Lee et al. (2024), StructVit Kim et al. (2024)). The Top-1 accuracy results from Table 4 demonstrate that InterLatent outperforms all competing methods, validating its effectiveness.

	Top-1
SViT	65.8
VideoMAE	70.8
CAST	71.6
StructVit	71.5
TDRL	71.4
CaRiNG	72.0
`InterLatent`	**72.7**

Table 4: The Top-1 results on SSv2