LEARNING TEMPORALLY CAUSAL LATENT PROCESSES FROM GENERAL TEMPORAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Our goal is to recover time-delayed latent causal variables and identify their relations from measured temporal data. Estimating causally related latent variables from observations is particularly challenging as the latent variables are not uniquely recoverable in the most general case. In this work, we consider both a nonparametric, nonstationary setting and a parametetric setting for the latent processes and propose two provable conditions under which temporally causal latent processes can be identified from their nonlinear mixtures. We propose LEAP, a theoretically-grounded architecture that extends Variational Autoencoders (VAEs) by enforcing our conditions through proper constraints in causal process prior. Experimental result on various data sets demonstrate that temporally causal latent processes are reliably identified from observed variables under different dependency structures and that our approach considerably outperforms baselines that do not leverage history or nonstationarity information. This is one of the first works that successfully recover time-delayed latent processes from nonlinear mixtures without using sparsity or minimality assumptions.

1 INTRODUCTION

Causal discovery seeks to identify the underlying structure of the data generation process by exploiting an appropriate class of assumptions (Spirtes et al., 1993; Pearl, 2000). Despite its success in certain domains, most existing work either focuses on estimating the causal relations between observed variables (Spirtes & Glymour, 1991; Chickering, 2002; Shimizu et al., 2006), or starts from the premise that causal variables are given beforehand (Spirtes et al., 2013). Real-world observations (e.g., image pixels, sensor measurements, etc.), however, are not structured into causal variables to begin with. Estimating latent causal variable graphs from observations is particularly challenging as the latent variables, even with independent factors of variation (Locatello et al., 2019), are not identifiable or "uniquely" recoverable in the most general case (Hyvärinen & Pajunen, 1999). There exist several pieces of work aiming to uncover causally related latent variables. For instance, by exploiting the vanishing Tetrad conditions (Spearman, 1928) one is able to identify latent variables in linear-Gaussian models (Silva et al., 2006), and the so-called Generalized Independent Noise (GIN) condition was proposed to estimate linear, non-Gaussian latent variable causal graph (Xie et al., 2020). However, these approaches are constrained within linear relations, need sparsity or minimality assumptions, and require a relatively large number of measured variables as children of the latent variables. The work of (Bengio et al., 2019; Ke et al., 2019) used "quick adaptation" as training signals for learning latent structure but the identifiability results were not theoretically established.

Recent advances in the theory of nonlinear ICA have proven strong identifiability results (Hyvarinen & Morioka, 2016; 2017; Hyvarinen et al., 2019; Khemakhem et al., 2020; Sorrenson et al., 2020) by exploiting certain side information in additional to independence – it is assumed that generative latent factors $\mathbf{z}$ are conditionally independent given auxiliary variables $\mathbf{u}$ (e.g., time index, class label, etc.) which augment observation data $\mathbf{x}$. (we use latent factors and latent processes interchangeably.) Deep generative models fit with such tuples $(\mathbf{x}, \mathbf{u})$ are identifiable in function space and can recover *independent factors* up to a trivial transformation of original latent variables. Although temporal structure is widely used for nonlinear ICA, most existing work that establishes identifiability results considers only independent sources and assumes linear transitions. However, these unrealistic assumptions severely distort the results if the real latent factors have causal relations in between,

Table 1: Desirable attributes of existing theories. A green check denotes that a method has a desirable attribute, whereas a red cross denotes the opposite. † indicates an approach we implemented.

Approach	Temporal Dataset	Causally-related Latents	Nonparametric expression	Stationary Process
TCL (Hyvarinen & Morioka, 2016)	✓	✗	✗	✗
PCL (Hyvarinen & Morioka, 2017)	✓	✗	✓	✓
GCL (Hyvarinen et al., 2019)	✗	✗	✗	✗
iVAE (Khemakhem et al., 2020)	✗	✗	✗	✗
HM-NLICA (Hälvä & Hyvarinen, 2020)	✓	✗	✓	✗
CausalVAE (Yang et al., 2021)	✗	✓	✗	✗
LEAP (Theorem 1) †	✓	✓	✓	✗
LEAP (Theorem 2) †	✓	✓	✗	✓

or if the relations are nonlinear. It is not clear yet how the temporal structure may help in learning *causally-related latent factors*, together with their causal structure, from temporal observation data.

In this paper, we focus on the scenario where observed data variables $\mathbf{x}_t$ do not have direct causal edges but are generated by latent processes or confounders that have time-delayed causal relations in between. That is, the observed data $\mathbf{x}_t$ are unknown, arbitrary (but invertible) nonlinear mixtures of the underlying sources: $\mathbf{x}_t = g(\mathbf{z}_t)$. Our first goal is then to understand under what conditions the latent temporally causal processes can be identified. Inspired by real situations, we consider both a nonparametric, nonstationary setting and a parametetric setting for the latent processes. In the nonparametric setting, the generating process of each latent causal factors z_{it} is characterized by nonparametric assignment $z_{it} = f_i(\mathbf{Pa}(z_{it}), \epsilon_{it})$, in which the parents of z_{it}, together with noise term ϵ_{it}, generate z_{it} via function f_i, whose form is not constrained, with some time delay. In the parametric setting, the time-delayed causal influences among latent factors follow follow a certain parametric form. In both setting, we establish the identifiability of the latent factors and their causal influences, rendering them recoverable from the measured data.

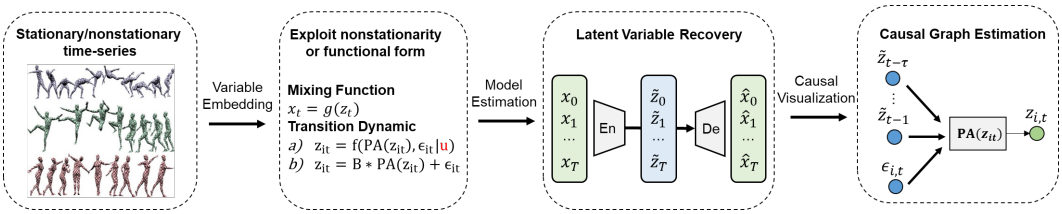

Figure 1: **Our approach:** we leverage nonstationarity in process noise or functional and distributional forms of temporal statistics to identify temporally causal latent processes from observation.

Our second goal is to develop a theoretically-grounded training framework that enforces the assumed conditions through proper constraints. We propose Latent tEmporally cAusal Processes estimation (**LEAP**), a novel architecture that extends Variational Autoencoders (VAEs) with a learned causal process prior network that enforces the Independent Noise (IN) condition and models nonstationarity through flow-based estimators. We evaluate LEAP on a number of synthetic and real-world datasets including video and motion capture data with the properties required by our conditions. Experiments demonstrate that temporally causal latent processes are reliably identified from observed variables under different dependency structure and our approach considerably outperforms those existing methods without leveraging history or nonstationarity information.

The closest work to ours includes (Klindt et al., 2020; Hyvarinen & Morioka, 2017), which require the underlying sources to be jointly independent for identifiability. Our work extends the theories to the discovery of conditional independent sources with time-delayed causal relations in between by leveraging non-stationarity, or functional and distribution forms of temporal statistics. To the best of our knowledge, this is one of the first works that successfully recover time-delayed latent processes from their nonlinear mixtures without using sparsity or minimality assumptions. The proposed framework may serve as a promising tool for creating versatile models robust to domain shifts and may be extended for more general conditions with changing causal relations over time.

2 IDENTIFIABILITY THEORY

2.1 IDENTIFIABILITY PROPERTY

We summarize the recent literature related to our work from four perspectives and compared them with our work in Table 1. Detailed discussions are given in Appendix D. Following prior work, we define identifiability in representation function space.

Definition 1 (Componentwise identifiability) *Let $\mathbf{x}_t$ be a sequence of observed variables generated by the true temporally causal latent processes with (f, p_ϵ) and nonlinear mixing function g, given in the introduction. A learned generative model $(\hat{g}, \hat{f}, \hat{p}_\epsilon)$ is observationally equivalent to (g, f, p_ϵ) if the joint distribution $p_{\hat{g},\hat{f},\hat{p}_\epsilon}(\mathbf{x}_t)$ matches $p_{g,f,p_\epsilon}(\mathbf{x}_t)$ everywhere. We say latent causal processes are identifiable if observational equivalence can always lead to identifiability of latent variables up to permutation π and component-wise invertible transformation T.*

$$p_{\hat{g},\hat{f},\hat{p}_\epsilon}(\mathbf{x}_t) = p_{g,f,p_\epsilon}(\mathbf{x}_t) \Rightarrow \hat{g} = g \circ T \circ \pi. \tag{1}$$

Once latent causal processes are identifiable up to componentwise transformations, latent causal relations are also identifiable because conditional independence relations fully characterize time-delayed causal relations in a causally sufficient system. Note that invertible componentwise transformations on latent causal processes does not change their conditional independence relations.

2.2 OUR PROPOSED CONDITIONS

We consider two novel conditions that ensure the identifiability of temporally causal latent processes, using (1) nonstationarity or (2) functional and distributional forms of transition distribution, respectively, and the identifiability of the latent processes is established in the following two theorems. Formal proof is provided in Appendix A.

Theorem 1 (Nonparametric processes) *Assume nonparametric processes in Equation 2, where the transition function f_i can be any third-order differentiable functions and mixing function g is injective and differentiable almost everywhere. $\mathbf{Pa}(z_{it})$ denotes the set of (time-delayed) parent nodes of z_{it}.*

$$\underbrace{\mathbf{x}_t = g(\mathbf{z}_t)}_{\textit{Nonlinear mixing}}, \quad \underbrace{z_{it} = f_i\left(\{z_{j,t-\tau} | z_{j,t-\tau} \in \mathbf{Pa}(z_{it})\}, \epsilon_{it}\right)}_{\textit{Nonparametric transition}} \; with \; \underbrace{\epsilon_{it} \sim p_{\epsilon_i|u}}_{\textit{Nonstationary noise}} \tag{2}$$

Assume:

1. *(Independent Noise): The noise terms ϵ_{it} are jointly independent (i.e., spatially and temporally independent), and also independent from $\mathbf{Pa}(z_{it})$;*
2. *(Nonstationary Noise): Noise distribution $p_{\epsilon_i|u}$ is modulated (in any way) by the observed categorical auxiliary variables $\mathbf{u}$;*
3. *(Sufficient Variability): For any $\mathbf{z}_t \in \mathbb{R}^n$ there exist $2n+1$ values for $\mathbf{u}$, denoted by $\mathbf{u}_j$ such that $2n$ vectors $\mathbf{w} \in \mathbb{R}^{2n}$ given by: $[\mathbf{w}(\mathbf{z}_t, \mathbf{u}_1) - \mathbf{w}(\mathbf{z}_t, \mathbf{u}_0), \ldots, \mathbf{w}(\mathbf{z}_t, \mathbf{u}_{2n}) - \mathbf{w}(\mathbf{z}_t, \mathbf{u}_0)]$ are linearly independent with $\mathbf{w}(\mathbf{z}_t, \mathbf{u})$ defined below where q_i is log density and $\mathbf{z}_{Hx} = \{\mathbf{z}_{t-\tau}\}$ denotes history information up to maximum time lag L:*

$$\mathbf{w}(\mathbf{z}_t, \mathbf{u}) \triangleq \left(\frac{\partial q_1(z_{1t}|\mathbf{z}_{Hx}, \mathbf{u})}{\partial z_{1t}}, \ldots, \frac{\partial q_n(z_{nt}|\mathbf{z}_{Hx}, \mathbf{u})}{\partial z_{nt}}, \frac{\partial^2 q_1(z_{1t}|\mathbf{z}_{Hx}, \mathbf{u})}{\partial z_{1t}^2}, \ldots, \frac{\partial^2 q_n(z_{nt}|\mathbf{z}_{Hx}, \mathbf{u})}{\partial z_{nt}^2} \right). \tag{3}$$

Then the identifiability property of temporally causal latent processes is ensured.

Theorem 2 (Parametric processes) *Assume the vector autoregressive process in Equation 4 where the state transition functions are linear and additive and mixing function g is injective and differentiable almost everywhere. Let $B_\tau \in \mathbb{R}^{n \times n}$ be the state transition matrix at lag τ. The process noise*

ϵ_t is assumed to be stationary and both spatially and temporally independent.

$$\underbrace{\mathbf{x}_t = g(\mathbf{z}_t)}_{\textit{Nonlinear mixing}}, \quad \underbrace{\mathbf{z}_t = \sum_{\tau=1}^{L} \mathbf{B}_\tau \mathbf{z}_{t-\tau} + \epsilon_t}_{\textit{Linear additive transition}} \ \textit{with} \ \underbrace{\epsilon_{it} \sim p_{\epsilon_i}}_{\textit{Independent noise}} \tag{4}$$

Assume:

1. *(Generalized Laplacian Noise): Process noise $\epsilon_i \sim p_{\epsilon_i}$ is jointly independent and conforms to generalized Laplacian distribution $p_{\epsilon_i} = \frac{\alpha_i \lambda_i}{2\Gamma(1/\alpha_i)} \exp\left(-\lambda_i |\epsilon_i|^{\alpha_i}\right)$ with $\alpha < 2$;*
2. *(Nonsingular State Transitions): At least one state transition matrix $\mathbf{B}_\tau$ is of full rank.*

Then the identifiability property of temporally causal latent processes is ensured.

3 LEAP: Latent Temporally Causal Processes Estimation

To validate our proof, we propose LEAP – a Latent tEmporally cAusal Processes estimation framework built upon the framework of VAEs while enforcing the conditions in Section 2.2 as constraints for identification of the latent causal processes. The model architecture is shown in Figure 2.

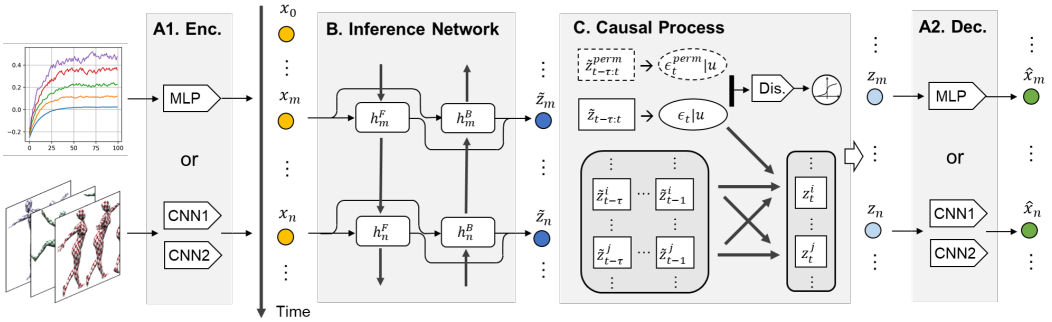

Figure 2: **Model overview**. LEAP consists of (A) encoders and decoders for specific data types; (B) a recurrent inference network that approximates the posteriors of latent causal variables, and (C) a causal process prior network that models nonstationary latent causal processes with IN constraints.

3.1 Causal Process Prior Network

We model latent causal processes in the learned prior network. To enforce the IN condition, latent transition priors $p(z_{it}|\mathbf{Pa}(z_{it}))$ are reparameterized into factorized noise distributions using the change of variable formula. The nonstationary noise distributions are learned by flow-based estimators and independence constraints are enforced through contrastive approach. Finally, pruning techniques based on masked inputs and soft-thresholding are used to produce interpretable, sparse causal relations.

3.1.1 Transition Prior Modeling

Nonparametric and parametric transition priors are modeled. The IN conditions are used to reparameterize transition priors into factorized noise distributions.

Nonparametric Transition Let $\{h_i\}$ be a set of learned inverse causal transition functions that take the estimated latent causal variables and output the noise terms, i.e., $\hat{\epsilon}_{it} = h_i\left(\hat{z}_{it}, \{\hat{\mathbf{z}}_{t-\tau}\}\right)$. We model each output component ϵ_{it} with a separate Multi-Layer Perceptron (MLP) network, so that we can easily disentangle the effects from inputs to outputs. Design transformation $\mathbf{A} \to \mathbf{B}$ with low-triangular Jacobian as follows:

$$\underbrace{\left[\mathbf{z}_{t-L}, \ldots, \mathbf{z}_{t-1}, \mathbf{z}_t\right]^\top}_{\mathbf{A}} \Rightarrow \underbrace{\left[\mathbf{z}_{t-L}, \ldots, \mathbf{z}_{t-1}, \epsilon_t\right]^\top}_{\mathbf{B}} \quad \textit{with} \quad \mathbf{J}_{\mathbf{A}\to\mathbf{B}} = \begin{pmatrix} \mathbb{I}_{nL} & 0 \\ * & \operatorname{diag}\left(\frac{\partial h_i}{\partial z_{it}}\right) \end{pmatrix}. \tag{5}$$

By applying IN condition, the joint distribution of latent causal variables can be written as:

$$\log p(\mathbf{A}) = \log p(\mathbf{B}) + \log\left(\left|\det\left(\mathbf{J}_{\mathbf{A}\to\mathbf{B}}\right)\right|\right) \tag{6}$$

$$= \underbrace{\log p\left(\mathbf{z}_{t-L}, \ldots, \mathbf{z}_{t-1}\right) + \sum_{i=1}^{n} \log p(\epsilon_i|u)}_{\text{IN Condition}} + \log\left(\left|\det\left(\mathbf{J}_{\mathbf{A}\to\mathbf{B}}\right)\right|\right). \tag{7}$$

The transition prior $p\left(\mathbf{z}_t|\{\mathbf{z}_{t-\tau}\}_{\tau=1}^{L}\right)$ can thus be evaluated using factorized noise distributions by cancelling out the marginals of time-delayed causal variables on both sides of Equation 7. Given that this Jacobian is triangular, we can efficiently compute its determinant as $\prod_i \frac{\partial h_i}{\partial z_{it}}$.

$$\log p\left(\mathbf{z}_t|\{\mathbf{z}_{t-\tau}\}_{\tau=1}^{L}\right) = \sum_{i=1}^{n} \log p(\epsilon_i|u) + \log\left(\left|\det\left(\mathbf{J}_{A\to B}\right)\right|\right) \tag{8}$$

Parametric Transition A group of learned linear layers $\{\mathbf{h}_\tau\}$ is used to model the inverse transition functions: $\hat{\epsilon}_{it} = \hat{\mathbf{z}}_t - \sum_{\tau=1}^{L} \mathbf{h}_\tau \hat{\mathbf{z}}_{t-\tau}$. Because of additive noise, the transition priors can be directly written in terms of factorized noise distributions.

$$\log p\left(\mathbf{z}_t|\{\mathbf{z}_{t-\tau}\}_{\tau=1}^{L}\right) = \sum_{i=1}^{n} \log p(\epsilon_i) \tag{9}$$

3.1.2 Nonstationary Noise Estimation

A flow-based density estimator is used to fit the residuals (e.g., non-Gaussian noises) and score the likelihood in Equation 8 and 9. Independence of residuals is enforced inside the density estimator.

Flow-based Noise Estimation We apply the componentwise neural spline flow model (Dolatabadi et al., 2020) to fit the estimated noise terms. The distribution of each noise component $\epsilon_{it} \sim p_{\epsilon_i|u}$ is modeled separately by transforming standard normal noises through linear rational splines $s_{i,u}$. To model nonstationarity, we keep a copy of spline flows $\{s_{i,u}\}$ for each nonstationary regime u and trigger it when data falls into this category: $p(\epsilon_i|u) = p_{\mathcal{N}(0,1)}\left(s_{i,u}(\epsilon_i)\right)$. Stationary sources is considered as a special case where the nonstationary context $u = 1$ for all data samples.

Independence by Contrastive Learning We further force the estimated noises $\{\epsilon_{it}\}$ to be spatially and temporally independent with regarding to (i, t) by contrastive approach. Similar to FactorVAE (Kim & Mnih, 2018), we train a discriminator $\mathcal{D}(\{\epsilon_{it}\})$ together with the latent variable model, to distinguish positive samples which are the estimated noise terms, against negative samples $\{\epsilon_{it}^{\text{perm}}\}$ which are random permutations of the noise terms across the batch for each noise dimension (i, t). The total correlation (TC) of noise terms can be estimated by the density-ratio trick in Equation 10 and is added to the loss function to enforce joint independence of noise terms:

$$\mathcal{L}_{TC} = \mathbb{E}_{\{\epsilon_{it}\}\sim(q(\mathbf{z}_t),h_i)} \log \frac{\mathcal{D}(\{\epsilon_{it}\})}{1 - \mathcal{D}(\{\epsilon_{it}\})} \tag{10}$$

3.1.3 Structure Estimation

We recover causal structure in latent processes by a combination of masked inputs and pruning approaches. Note that our identifiability results do not rely on sparsity in latent processes. The pruning is used only for visualizing causal relations when underlying causal processes are nonlinear.

Masked Input and Regularization For sparse latent process, each MLP of the transition model f_i has a learned N-dimensional soft mask vector $\sigma(\gamma_i) \in [0, 1]$ with the j-th input of the MLP being multiplied by $\sigma(\gamma_{ij})$. A fixed L_1 penalty is added on mask to slightly control false discoveries during training: $\mathcal{L}_{\text{Mask}} = \|\sigma(\gamma_{ij})\|_1$. For linear transition, the penalty is added on the transition matrices since the weights directly decide whether an edge exists.

Pruning For nonlinear relations, we use LassoNet (Lemhadri et al., 2021) as a post-processing step to remove spurious edges. This approach prunes input nodes by jointly passing the residual layer and the first hidden layer through a hierarchical soft-thresholding optimizer. We fit the model on a subset of recovered latent causal variables. Though the true causal relations may not follow the causal additive assumption, the pruning step usually performs well.

3.2 INFERENCE NETWORK

Bidirectional GRU is used to infer latent variables with both past and future information. We approximate the posterior $q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$ with isotropic Gaussian, using mean and variance terms given by the inference network. The KL divergence is $\mathcal{L}_{\text{KLD}} = D_{KL}(q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})||p(\mathbf{z}_{1:T}))$ and is estimated via sampling approach because the distribution form of prior is not specified but rather learned by causal process network.

3.3 ENCODER AND DECODER

The reconstruction loss is $\mathcal{L}_{\text{Recon}} = p_{\text{reconstruct}}(\mathbf{x}_t|\hat{\mathbf{x}}_t)$, where $\hat{\mathbf{x}}_t$ is the output of decoder and $p_{\text{reconstruct}}$ is the decoder distribution. The architecture details can be found in Appendix C. For synthetic and point cloud data, we use MLP with LeakyReLU units as encoder and decoder. For video datasets with single objects (e.g., KittiMask), vanilla CNNs are used. For videos with multiple objects, we apply a disentangled design (Kulkarni et al., 2019) with two separate CNNs, one for extracting visual features and the other for locating object locations with spatial softmax units. The decoder retrieves object features using object locations and reconstructs the scene.

3.4 OPTIMIZATION

We train the VAE and noise discriminator jointly. The VAE parameters are updated using the augmented ELBO loss in Equation 11. The discriminator is trained to classify between residuals from $q(\{\epsilon_{it}\})$ and $q(\{\epsilon_{it}^{\text{perm}}\})$ in Equation 12, thus approximating the density ratio for estimating $\mathcal{L}_{\text{TC}}$:

$$\mathcal{L}_{\text{ELBO}} = \frac{1}{N}\sum_{i\in N} \mathcal{L}_{\text{Recon}} + \beta\mathcal{L}_{\text{KLD}} + \gamma\mathcal{L}_{\text{Mask}} + \sigma\mathcal{L}_{\text{TC}}, \tag{11}$$

$$\mathcal{L}_{\mathcal{D}} = \frac{1}{2N}\sum_{i\in N}\left[\log \mathcal{D}(\{\epsilon_{it}\}) + \sum_{i\in N'} \log\left(1 - \mathcal{D}(\{\epsilon_{it}^{\text{perm}}\})\right)\right]. \tag{12}$$

4 EXPERIMENTS

We comparatively evaluate LEAP on a number of temporal datasets with the required assumptions satisfied or violated. We aim to answer the following questions:

1. Can LEAP reliably learn temporally causal latent processes from scratch, under the proposed conditions? What are the contributions of each module in the architecture?
2. Is history/nonstationary necessary for identifiability of latent causal variables?
3. How do common assumptions in nonlinear ICA, i.e., independence sources or linear relations assumptions distort identifiability if the real latent factors have time-delayed causal relations in between, or if the latent processes are nonlinear?
4. Can LEAP generalize when some critical assumptions (e.g., conditional independence, stationary causal relations, etc.) in the proposed conditions are violated?

To measure the identifiability of latent causal variables, we compute **Mean Correlation Coefficient** (MCC) on the validation dataset, a standard metric in the ICA literature for continuous variables. MCC reaches `1` when latent variables are perfectly identifiable up to permutation and component-wise invertible transformation in the noiseless case. To evaluate the recovery performance on causal relations, we use different approaches for (1) linear and (2) nonlinear transitions: the entries of estimated state transition matrices are compared with the ground-truth ones after permutation, signs and scaling are adjusted; the estimated causal skeleton is compared with the true data generating structure.

4.1 SYNTHETIC EXPERIMENTS

We first design synthetic datasets with the properties required by nonparametric **(NP)** and parametric **(VAR)** conditions in Section 2.2. We set latent size $n = 8$. The lag number of the process is set to $L = 2$. The mixing function g is a random three-layer MLP with LeakyReLU units. We provided four kinds of data generation processes for nonparametric conditions, parametric conditions and two generalized conditions separately. The details are provided in Appendix B.1.

Baselines and Ablation We experimented with three kinds of nonlinear ICA baselines: (1) BetaVAE (Higgins et al., 2016) and FactorVAE (Kim & Mnih, 2018) which ignores both history and nonstationarity information; (2) IVAE (Khemakhem et al., 2020) and TCL (Hyvarinen & Morioka, 2016) which exploits nonstationarity to establish identifiability, and (3) SlowVAE (Klindt et al., 2020) and PCL (Hyvarinen & Morioka, 2017) which exploits temporal constraints but assumes independent sources. Model variants are built to disentangle the contributions of each module. As in Table 2, we start with BetaVAE and add our proposed modules successively and do not change any training settings. Finally (4) , we fit our LEAP that use linear transitions (LEAP-VAR) with NP data to show if linear relation assumptions distort identifiability. The details are in Appendix C.

Main Results Figure 3 gives the results on NP datasets. The latent processes are successfully recovered, as indicated by (a) high MCC for casually-related factors and (b) recovery of time-delayed causal relations. The latent causal variables are estimated up to permutation and componentwise invertible transformation (c). The comparisons with baselines are shown in (d). In general, the baselines that do not exploit history or nonstationarity cannot identify the latent processes. SlowVAE and PCL distort the results due to independent sources assumptions. Interestingly, LEAP-VAR which uses linear causal transitions gains partial identifiability on NP datasets. This might promote the usage of the linear component of the transition signals to guild the learning of latent causal variables.

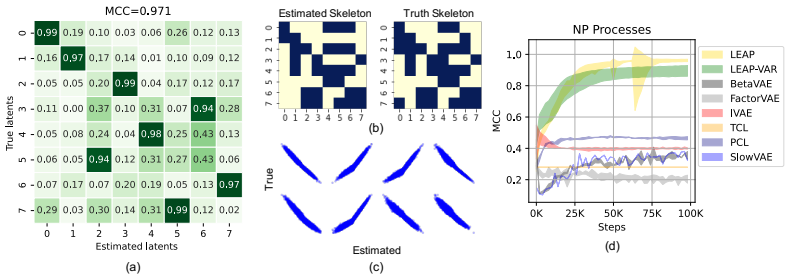

Figure 3: Results for synthetic nonparametric processes (NP) datasets: (a) MCC for causally-related factors; (b) recovered causal skeletons; (c) scatterplots between estimated and true factors, and (d) MCC trajectories comparisons between LEAP and baselines.

The results for VAR datasets are in Figure 4. Similarly, the latent processes are recovered, as indicated by (a) high MCC and (b) recovery of state transition matrices. The latent causal variables are estimated up to permutation and scaling (c). The baselines without using history or assume independent sources again fail to recover the latent processes (d). We disentangle the contributions of different components of LEAP in Table 2. Causal process prior and nonstationary flow significantly improve identifiability. Noise discriminator further increases MCC and reduces variance. Note that we use a dense NP dataset with input masks removed for ablation to validate that our proposed conditions do not rely on sparse causal structure for latent causal discovery.

Table 2: Contribution of each module to MCC.

Modules	**NP (Dense)**	**VAR**
Baseline (β-VAE)	0.446 ± 0.004	0.495 ± 0.007
+ Causal Transition Prior	0.721 ± 0.121	0.752 ± 0.035
+ Nonstationary Flow	0.939 ± 0.008	0.935 ± 0.014
+ Noise Discriminator	0.983 ± 0.002	0.978 ± 0.004

Towards Generalized Conditions We verify the robustness of our framework by experimenting with two VAR datasets that clearly violate our assumptions: (1) with causal relations changing over time, and (2) with instantaneous causal relations. We fit LEAP on these datasets without any modification. The results are in Figure 5. Our framework can gain partial identifiability under (1) but violating conditional independence assumption (2) distorts the results. Our causal process prior

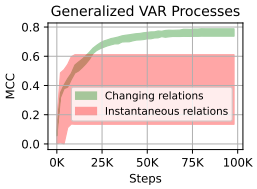

Figure 5: MCC for generalized conditions.

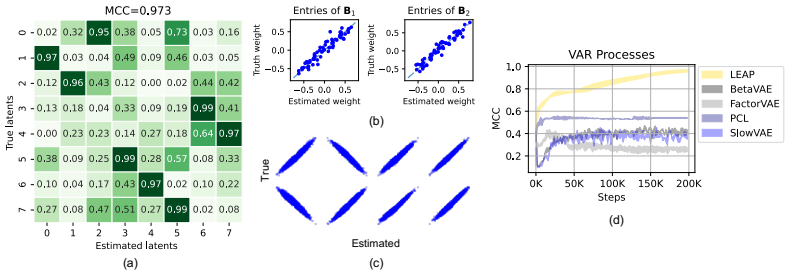

Figure 4: Results for synthetic parametric processes (VAR) datasets: (a) MCC for causally-related factors; (b) scatterplots of the entries of $\mathbf{B}_\tau$; (c) scatterplots between estimated and true factors, and (4) MCC trajectories comparisons between LEAP and baselines.

network, although designed to model nonstationary noise, can be extended to model changing causal relations.

4.2 Real-world Applications: Perceptual Causality

Three public datasets including KiTTiMask (Klindt et al., 2020), Mass-Spring System (Li et al., 2020) and CMU MoCap database are used. The observations together with the true temporally causal latent processes are showcased in Figure 6. Note for CMU-Mocap, the true latent causal variables and time-delayed relations are not provided.

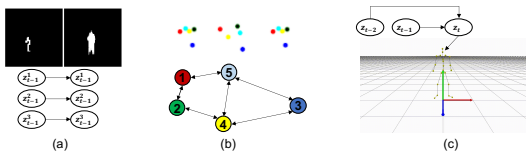

Figure 6: Real-world datasets: (a) KiTTiMask is a video dataset of binary pedestrian masks, (b) Mass-Spring System is a video dataset with ball movement rendered in color and invisible springs, and (c) CMU-Mocap is a 3D point cloud dataset of skeleton-based signals.

4.2.1 Independent Sources – KiTTiMask

LEAP with VAR transitions is used. We set latent size $n = 10$ and the lag number $L = 1$. As shown in Figure 7, the latent causal processes are recovered, as seen from (a) high MCC for independent sources; (b) latent factors estimated up to componentwise transformation; (c) the estimated state transition matrix, which is diagonal (independent sources); (d) latent traversals confirming that the three latent causal variables correspond to the vertical, horizontal and scale of pedestrian masks.

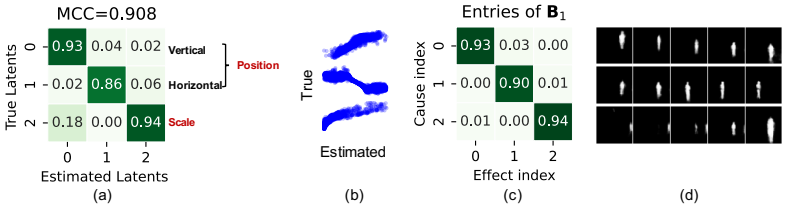

Figure 7: KiTTiMask dataset results: (a) MCC for independent sources; (b) Scatterplots between estimated and true factors; (c) Entries of $\mathbf{B}_1$; (d) Latent traversal on a fixed video frame.

4.2.2 Parametric Transition – Mass-Spring System

Mass-Spring system is a linear dynamical system with ball locations (x_t^i, y_t^i) as state variables and lag number $L = 2$. LEAP with VAR transitions is used. As shown in Figure 8, the time-delayed cross causal relations are recovered as: (a) causal variables keypoint (x_t^i, y_t^i) are successfully estimated; (b,d) the spring connections between balls are recovered, and (c) the latent variables correspond to each ball's location in the video frame.

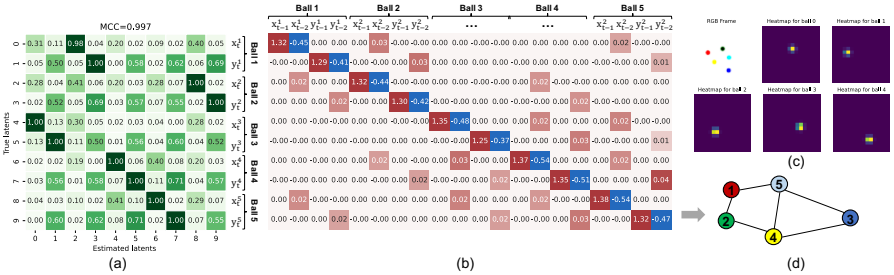

Figure 8: Mass-spring system results: (a) MCC for causally-related sources; (b) Entries of $\mathbf{B}_{1,2}$; (c) Latent variables for a fixed video frame; (d) Recovered causal skeletons.

4.2.3 Nonparametric Transition – CMU-MoCap

We fit LEAP with nonparametric transitions on 12 trials of motion capture data for subject 7 with 62 observed variables of skeleton-based measurements at each time step. The 12 trials contain walk cycles with slightly different dynamics (e.g., walk, slow walk, brisk walk). We set latent size $n = 8$ and lag number $L = 2$. The differences between trials are modeled by nonstationary noise with one regime for each trial. The results are in Figure 9. Three latent variables (i.e., pitch, yaw, roll) are found to explain most of the variances of human walk cycles (c). A video demonstration for latent traversal is available[1]. The learned latent coordinates show smooth cyclic patterns with slight differences between trials (a). Finally, we find that pitch (e.g., limbs) and roll (e.g., shoulders) of human walking are coupled while yaw has independent dynamics (b).

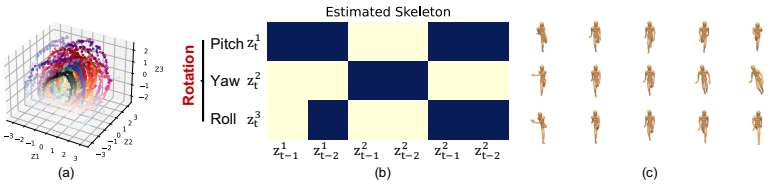

Figure 9: Mocap dataset results: (a) Latent coordinates dynamics for 12 trials; (b) Estimated skeleton; (c) Latent traversal by rendering the reconstructed point clouds into video frame.

5 Limitations and Future Work

In this work, we propose two provable conditions under which temporally causal latent processes can be identified from observed variables. The theories have been validated on a number of datasets with the properties required by the conditions. The limitations of this work lie in our two major assumptions: (1) instantaneous relations between latent causal variables are ignored due to conditional independence assumption, and (2) causal relations are fixed across time steps, both of which may not be true for some temporal data. The existence of instantaneous relations distorts identifiability results but the amount of these relations can be controlled by time resolution. While we do not establish theories under changing causal relations, we have shown the possibilities of generalizing our identifiability results to changing dynamics through experiments. Extending our identifiability theories and training framework for these properties are our future directions.

6 Reproducibility Statement

Our code for the proposed framework and experiments can be found at https://bit.ly/2Yg56Ao. For theoretical results, the assumptions and a complete proof of the claims are included in the Section A. For synthetic experiments, the data generation process is described in Section B.1.

[1] https://www.youtube.com/watch?v=vvW1IBo3J6s

REFERENCES

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Hadi Mohaghegh Dolatabadi, Sarah Erfani, and Christopher Leckie. Invertible generative modeling using linear rational splines. In *International Conference on Artificial Intelligence and Statistics*, pp. 4236–4246. PMLR, 2020.

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

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. 2016.

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

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

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

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

Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Bernhard Schölkopf, Michael C Mozer, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

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.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pp. 2649–2658. PMLR, 2018.

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.

Tejas Kulkarni, Ankush Gupta, Catalin Ionescu, Sebastian Borgeaud, Malcolm Reynolds, Andrew Zisserman, and Volodymyr Mnih. Unsupervised learning of object keypoints for perception and control. *arXiv preprint arXiv:1906.11883*, 2019.

Ismael Lemhadri, Feng Ruan, Louis Abraham, and Robert Tibshirani. Lassonet: A neural network with feature sparsity. *Journal of Machine Learning Research*, 22(127):1–29, 2021.

Yunzhu Li, Antonio Torralba, Animashree Anandkumar, Dieter Fox, and Animesh Garg. Causal discovery in physical systems from videos. *arXiv preprint arXiv:2007.00631*, 2020.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019.

Stanisław Mazur and Stanisław Ulam. Sur les transformations isométriques d'espaces vectoriels normés. *CR Acad. Sci. Paris*, 194(946-948):116, 1932.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

R. Silva, R. Scheines, C. Glymour, and P. Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7:191–246, 2006.

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

Charles Spearman. Pearson's contribution to the theory of two factors. *British Journal of Psychology*, 19:95–101, 1928.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Spring-Verlag Lectures in Statistics, 1993.

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

Peter L Spirtes, Christopher Meek, and Thomas S Richardson. Causal inference in the presence of latent variables and selection bias. *arXiv preprint arXiv:1302.4983*, 2013.

Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. *arXiv preprint arXiv:2010.04917*, 2020.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9593–9602, 2021.

Kun Zhang and Aapo Hyvärinen. A general linear non-gaussian state-space model: Identifiability, identification, and applications. In *Asian Conference on Machine Learning*, pp. 113–128. PMLR, 2011.

Supplementary Materials for Learning Temporally Causal Latent Processes from General Temporal Data

A Proof of Identifiability Theory

A.1 Preliminaries

Equivalent Relations on Latent Space Our proof of identifiability starts from deriving relations on estimated latent space from observational equivalence: the joint distribution $p_{\hat{g},\hat{f},\hat{p}_\epsilon}(\mathbf{x}_{\text{Hx}}, \mathbf{x}_t)$ matches $p_{g,f,p_\epsilon}(\mathbf{x}_{\text{Hx}}, \mathbf{x}_t)$ everywhere. Note that we consider only one future time step $\mathbf{x}_t$ for simplicity as the joint probability of the whole sequence can be decomposed into product of these terms. Since the learned unmixing function $\mathbf{x}_t = \hat{g}(\mathbf{z}_t)$ can be written as $\mathbf{x}_t = (g \circ (g)^{-1} \circ \hat{g})(\mathbf{z}_t)$, we can assume that $\hat{g} = g^* \circ h$ for some function h on the latent space where h. Our goal here is really to show that this function h is a permutation with component-wise nonlinearities. It has been proved in (Klindt et al., 2020) that:

1. Indeterminancy h on latent space can only be a bijection on the latent space if both g and $\hat{g}$ are injective functions;
2. Equation 13 can be directly derived from observational equivalence using the injective properties of $(g, \hat{g})$.

$$p(\mathbf{z}_t|\{\mathbf{z}_{t-\tau}\}) = p(h^{-1}(\mathbf{z}_t)|\{h^{-1}(\mathbf{z}_{t-\tau})\})\frac{p(\mathbf{z}_t)}{p(h^{-1}(\mathbf{z}_t))} \quad \forall(\mathbf{z}_t, \{\mathbf{z}_{t-\tau}\}) \tag{13}$$

Identifiability of Linear Non-Gaussian State-Space Model Linear state space model defined below in has been proved to be fully identifiable in (Zhang & Hyvärinen, 2011) when both the process noise ϵ_t and measurement error $\mathbf{e}_t$ are temporally white and independent of each other and at most one component of the process noise ϵ_t is Gaussian. The observation error $\mathbf{e}_t$ can be either Gaussian or non-Gaussian:

$$\mathbf{x}_t = \mathbf{A}\mathbf{z}_t + \mathbf{e}_t, \tag{14}$$

$$\mathbf{z}_t = \sum_{\tau=1}^{L} \mathbf{B}_\tau \mathbf{z}_{t-\tau} + \epsilon_t. \tag{15}$$

We'll make use of this property of linear Non-Gaussian SSM for deriving **Theorem 2**. The main idea is if we can prove h of parametric conditions (which also has vector autoregressive processes as in Equation 15 with non-Gaussian noise) is within affine transformations, the componentwise identifiability of true latent variables can be directly derived because we can treat the affine indeterminacy as a "high-level" affine mixing of sources same as $\mathbf{A}$ in Equation 14 without measurement error.

A.2 Proof of Theorem 1

Theorem A.1 (Nonparametric processes) *For nonparametric processes in Equation 16, where the transition function f_i can be any third-order differentiable functions and mixing function g is injective and differentiable almost everywhere. $\mathbf{Pa}(z_{it})$ denotes the set of parent nodes of z_{it}.*

$$\underbrace{\mathbf{x}_t = g(\mathbf{z}_t)}_{\textit{Nonlinear mixing}}, \quad \underbrace{z_{it} = f_i\left(\{z_{j,t-\tau}|z_{j,t-\tau} \in \mathbf{Pa}(z_{it})\}, \epsilon_{it}\right)}_{\textit{Nonparametric transition}} \; with \; \underbrace{\epsilon_{it} \sim p_{\epsilon_i|u}}_{\textit{Nonstationary noise}} \tag{16}$$

Assume:

1. *(Independent Noise): The noise terms ϵ_{it} are assumed to be jointly independent (i.e., spatially and temporally independent), and also independent from $\mathbf{Pa}(z_{it})$;*
2. *(Nonstationary Noise): Noise distribution $p_{\epsilon_i|u}$ is modulated (in any way) by the observed categorical auxiliary variables $\mathbf{u}$;*

3. *(Sufficient Variability): For any $\mathbf{z}_t \in \mathbb{R}^n$ there exist $2n+1$ values for $\mathbf{u}$, denoted by $\mathbf{u}_j$ such that $2n$ vectors $\mathbf{w} \in \mathbb{R}^{2n}$ given by: $[\mathbf{w}(\mathbf{z}_t, \mathbf{u}_1) - \mathbf{w}(\mathbf{z}_t, \mathbf{u}_0), \ldots, \mathbf{w}(\mathbf{z}_t, \mathbf{u}_{2n}) - \mathbf{w}(\mathbf{z}_t, \mathbf{u}_0)]$ are linearly independent with $\mathbf{w}(\mathbf{z}_t, \mathbf{u})$ defined below where q_i is log density and $\mathbf{z}_{Hx} = \{\mathbf{z}_{t-\tau}\}$ denotes history information up to maximum time lag L:*

$$\mathbf{w}(\mathbf{z}_t, \mathbf{u}) \triangleq \left(\frac{\partial q_1(z_{1t}|\mathbf{z}_{Hx}, \mathbf{u})}{\partial z_{1t}}, \ldots, \frac{\partial q_n(z_{nt}|\mathbf{z}_{Hx}, \mathbf{u})}{\partial z_{nt}}, \frac{\partial^2 q_1(z_{1t}|\mathbf{z}_{Hx}, \mathbf{u})}{\partial z_{1t}^2}, \ldots, \frac{\partial^2 q_n(z_{nt}|\mathbf{z}_{Hx}, \mathbf{u})}{\partial z_{nt}^2} \right). \quad (17)$$

Then the identifiability property of temporal latent causal processes is ensured.

Proof: We first extend Equation 13 to include conditioning on the nonstationary regime $\mathbf{u}$. This can be directly achieved because different regimes can be treated as separate datasets with every regime satisfying the latent space relations in Equation 13. We take logs on both sides,

$$q(\mathbf{z}_t|\{\mathbf{z}_{t-\tau}\}, \mathbf{u}) - q(h^{-1}(\mathbf{z}_t)|h^{-1}(\{\mathbf{z}_{t-\tau}\}, \mathbf{u})) = q(\mathbf{z}_t|\mathbf{u}) - q(h^{-1}(\mathbf{z}_t)|\mathbf{u}), \quad (18)$$

and using Independent Noise assumption in (1), the conditional log-pdf $q(\mathbf{z}_t|\{\mathbf{z}_{t-\tau}\}, \mathbf{u})$ and its estimated version $q(h^{-1}(\mathbf{z}_t)|\{h^{-1}(\mathbf{z}_{t-\tau})\}, \mathbf{u})$ are conditional independent (note this has been enforced in causal process network as constraints) and LHS can be factorized as:

$$\sum_i q_i(z_{it}|\{\mathbf{z}_{t-\tau}\}, \mathbf{u}) - q_i\left(\left[h^{-1}(\mathbf{z}_t)\right]_i|\{h^{-1}(\mathbf{z}_{t-\tau})\}, \mathbf{u}\right) = q(\mathbf{z}_t|\mathbf{u}) - q(h^{-1}(\mathbf{z}_t)|\mathbf{u}) \quad (19)$$

$$= \bar{q}(\mathbf{z}_t) - \bar{q}(h^{-1}(\mathbf{z}_t)). \quad (20)$$

where $\bar{q}$ is the marginal log-density of the components $\mathbf{z}_t$ when $\mathbf{u}$ is integrated out and it does not need to be factorial. Now we do the following simplification of notations. Let $h_i^{-1}(\mathbf{x}_t) = \left[h^{-1}(\mathbf{z}_t)\right]_i$. Denote the first-order and second-order derivatives by a superscript as:

$$q_i^1(z_{it}|\{\mathbf{z}_{t-\tau}\}, \mathbf{u}) = \frac{\partial q_i(z_{it}|\{\mathbf{z}_{t-\tau}\}, \mathbf{u})}{\partial z_{it}}, \quad (21)$$

$$q_i^2(z_{it}|\{\mathbf{z}_{t-\tau}\}, \mathbf{u}) = \frac{\partial^2 q_i(z_{it}|\{\mathbf{z}_{t-\tau}\}, \mathbf{u})}{\partial z_{it}^2}, \quad (22)$$

and take derivatives of both sides of Equation 20 with respect to z_{jt}, we have:

$$q_j^1(z_{jt}|\{\mathbf{z}_{t-\tau}\}, \mathbf{u}) - \sum_{i=1}^n q_i^1(h_i^{-1}(\mathbf{z}_t)|\{h^{-1}(\mathbf{z}_{t-\tau})\}, \mathbf{u}) \frac{\partial h_i^{-1}(\mathbf{z}_t)}{\partial z_{jt}} \quad (23)$$

$$= \bar{q}^j(\mathbf{z}_t) - \sum_i \bar{q}^j(h_i^{-1}(\mathbf{z}_t)) \frac{\partial h_i^{-1}(\mathbf{z}_t)}{\partial z_{jt}}. \quad (24)$$

Denote the first order derivative of h^{-1} as $v_i^j(\mathbf{z}_t) = \frac{\partial h_i^{-1}(\mathbf{z}_t)}{\partial z_{jt}}$ and $v_i^{jj'}(\mathbf{z}_t)$ is the second-order derivative with respect to a different component $z_{j't}$ for any $j \neq j'$. Taking another derivative with respect to $z_{j't}$ on both sides of Equation 24, the first term on LHS vanishes and we have:

$$\sum_i q_i^{11}(h_i^{-1}(\mathbf{z}_t)|\{h^{-1}(\mathbf{z}_{t-\tau})\}, \mathbf{u}) v_i^j(\mathbf{z}_t) v_i^{j'}(\mathbf{z}_t) + q_i^1(h_i^{-1}(\mathbf{z}_t)|\{h^{-1}(\mathbf{z}_{t-\tau})\}, \mathbf{u}) v_i^{jj'}(\mathbf{z}_t) = c^{jj'} \quad (25)$$

where $c^{jj'}$ the derivatives of RHS of Equation 24 which **does not depend on** $\mathbf{u}$. Similar to (Hyvarinen et al., 2019), collect all these equations in vector form by defining $\mathbf{a}_i(y)$ as a vector collecting all entries $\mathbf{u}) v_i^j(\mathbf{z}_t) v_i^{j'}(\mathbf{z}_t)$ for $j \in [1, n]$ and $j' \in [1, j-1]$ (we omit diagonal terms, and by symmetry, take only one half of the indices). Likewise, collect all the entries $v_i^{jj'}(\mathbf{z}_t)$ for $j \in [1, n]$ and $j' \in [1, j-1]$ in the vector $\mathbf{b}(\mathbf{z}_t)$. All the entries of $c^{jj'}$ are in $\mathbf{c}(\mathbf{z}_t)$. These $n(n-1)/2$ equations can be written a single system of equations:

$$\sum_i \mathbf{a}_i(y) q_i^{11}(h_i^{-1}(\mathbf{z}_t)|\{h^{-1}(\mathbf{z}_{t-\tau})\}, \mathbf{u}) + \mathbf{b}_i(\mathbf{z}_t) q_i^1(h_i^{-1}(\mathbf{z}_t)|\{h^{-1}(\mathbf{z}_{t-\tau})\}, \mathbf{u}) = \mathbf{c}(\mathbf{z}_t). \quad (26)$$

Now, collect the $\mathbf{a}$ and $\mathbf{b}$ into a matrix $\mathbf{M}$:

$$\mathbf{M}(\mathbf{z}_t) = (\mathbf{a}_1(\mathbf{z}_t), \dots, \mathbf{a}_n(\mathbf{z}_t), \mathbf{b}_i(\mathbf{z}_t), \dots, \mathbf{b}_n(\mathbf{z}_t)) \,. \tag{27}$$

Equation Equation 26 takes the form of the following linear system:

$$\mathbf{M}(\mathbf{z}_t)\mathbf{w}(\mathbf{z}_t, \mathbf{u}) = \mathbf{c}(\mathbf{z}_t), \tag{28}$$

where $\mathbf{w}$ are the vectors defined in Sufficient Variability assumption. Note the RHS of the linear system does not depend on $\mathbf{u}$ so we fix $\mathbf{z}_t$ and consider the $2n+1$ points $\mathbf{u}$ given for that $\mathbf{z}_t$ by the Sufficient Variability assumption. Collect Equation 28 above for $2n$ points starting from index 1:

$$\mathbf{M}(\mathbf{z}_t)\left(\mathbf{w}(\mathbf{z}_t, \mathbf{u}_1), \dots, \mathbf{w}(\mathbf{z}_t, \mathbf{u}_2 n)\right) = \left(\mathbf{c}(\mathbf{z}_t), \dots, \mathbf{w}(\mathbf{z}_t, \mathbf{u}_1)\right), \tag{29}$$

and collect the equation starting from index 0 for $2n$ points:

$$\mathbf{M}(\mathbf{z}_t)\left(\mathbf{w}(\mathbf{z}_t, \mathbf{u}_0), \dots, \mathbf{w}(\mathbf{z}_t, \mathbf{u}_{2n-1})\right) = \left(\mathbf{c}(\mathbf{z}_t), \dots, \mathbf{w}(\mathbf{z}_t, \mathbf{u}_1)\right). \tag{30}$$

Substract Equation 30 from Equation 29, we then have:

$$\mathbf{M}(\mathbf{z}_t)\underbrace{\left[\mathbf{w}(\mathbf{z}_t, \mathbf{u}_1) - \mathbf{w}(\mathbf{z}_t, \mathbf{u}_0), \dots, \mathbf{w}(\mathbf{z}_t, \mathbf{u}_{2n}) - \mathbf{w}(\mathbf{z}_t, \mathbf{u}_0)\right]}_{\mathbf{W}} = \mathbf{0}. \tag{31}$$

By Suffienct Variability assumption, the matrix $\mathbf{W}$ that has linearly independent columns and is a square matrix so is nonsingular. The only solution to the linear system above is thus:

$$\mathbf{M}(\mathbf{z}_t) = (\mathbf{a}_1(\mathbf{z}_t), \dots, \mathbf{a}_n(\mathbf{z}_t), \mathbf{b}_i(\mathbf{z}_t), \dots, \mathbf{b}_n(\mathbf{z}_t)) = \mathbf{0}. \tag{32}$$

In particular, $\mathbf{a}(\mathbf{z}_t)$ being zero implies no row of the Jacobian of $h^{-1}(\mathbf{z}_t)$ can have more than one non-zero entry. This holds for any $\mathbf{z}_t$. By continuity of the Jacobian and its invertibility, the non-zero entries in the Jacobian must be in the same places for all $\mathbf{z}_t$: If they switched places, there would have to be a point where the Jacobian is singular, which would contradict the bijection properties of h^{-1} derived in Section A.1. This means that each $h_i^{-1}(\mathbf{z}_t)$ is a function of only one z_{kt} for $k \in [1, n]$. The bijection h^{-1} also implies that each of the componentwise functions is invertible. Thus, we have proven that latent variables are identifiable up to permutation and componentwise invertible transformations and temporally causal latent processes with conditions required by **Theorem 1** are proved to be identifiable from observed variables. ■

A.3 Proof of Theorem 2

Theorem A.2 (Parametric processes) *For vector autoregressive process in Equation 33 where the state transition functions are linear and additive, mixing function g is injective and differentiable almost everywhere. Let $B_\tau \in \mathbb{R}^{n\times n}$ be the state transition matrix at lag τ. The process noise ϵ_t is assumed to be stationary and both spatially and temporally independent.*

$$\underbrace{\mathbf{x}_t = g(\mathbf{z}_t)}_{\textit{Nonlinear mixing}}, \quad \underbrace{\mathbf{z}_t = \sum_{\tau=1}^{L} \mathbf{B}_\tau \mathbf{z}_{t-\tau} + \epsilon_t}_{\textit{Linear additive transition}} \;\; with \;\; \underbrace{\epsilon_{it} \sim p_{\epsilon_i}}_{\textit{Independent noise}} \,. \tag{33}$$

Assume:

1. *(Generalized Laplacian Noise): Process noise $\epsilon_i \sim p_{\epsilon_i}$ is jointly independent and conforms to generalized Laplacian distribution $p_{\epsilon_i} = \frac{\alpha_i \lambda_i}{2\Gamma(1/\alpha_i)} \exp\left(-\lambda_i |\epsilon_i|^{\alpha_i}\right)$ with $\alpha < 2$;*
2. *(Nonsingular State Transitions): At least one state transition matrix $\mathbf{B}_\tau$ is of full rank.*

Then the identifiability property of temporal latent causal processes is ensured.

Proof: The following proof is inspired by Theorem 1 in (Klindt et al., 2020). The key differences are (i) allowing temporal causal relations B among the sources instead of independent sources assumption, (ii) extending the single time lag restriction to multiple time lags case.

Identifibility on causally-related sources Let us start from the simple case where the time lag $\tau = 1$. In this case, the transition dynamic in Equation (33) can be simplified as

$$\mathbf{x}_t = g(\mathbf{z}_t), \quad \mathbf{z}_t = \mathbf{B}\mathbf{z}_{t-1} + \epsilon_t. \tag{34}$$

Using Equation 13 and by applying the distributional forms of generalized Laplacian noise, we have:

$$\begin{aligned} & p(\mathbf{z}_t|\mathbf{z}_{t-1}) = p(h^{-1}(\mathbf{z}_t)|h^{-1}(\mathbf{z}_{t-1}))\frac{p(\mathbf{z}_t)}{p(h^{-1}(\mathbf{z}_t))} \\ \implies \quad & M||\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t-1}||_\alpha^\alpha - N||h^{-1}(\mathbf{z}_t) - \mathbf{B}h^{-1}(\mathbf{z}_{t-1})||_\alpha^\alpha = \log\frac{p(\mathbf{z}_t)}{p(h^{-1}(\mathbf{z}_t))}, \end{aligned} \tag{35}$$

where M, N is the constants appearing in the exponentials in $p(\mathbf{z}_t|\mathbf{z}_{t-1})$ and $p(h^{-1}(\mathbf{z}_t)|h^{-1}(\mathbf{z}_{t-1}))$.

Taking the derivative w.r.t $\mathbf{z}_{t-1}$ on both sides, we obtain

$$\frac{\partial||\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t-1}||_\alpha^\alpha}{\partial\mathbf{z}_{t-1}} = \frac{\partial||h^{-1}(\mathbf{z}_t) - \mathbf{B}h^{-1}(\mathbf{z}_{t-1})||_\alpha^\alpha}{\partial\mathbf{z}_{t-1}}. \tag{36}$$

For the left hand of the Equation (36), we can derive

$$\begin{aligned} & \frac{\partial||\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t-1}||_\alpha^\alpha}{\partial||\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t-1}||_\alpha}\frac{\partial||\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t-1}||_\alpha}{\partial\mathbf{z}_{t-1}} = \frac{\partial||h^{-1}(\mathbf{z}_t) - \mathbf{B}h^{-1}(\mathbf{z}_{t-1})||_\alpha^\alpha}{\partial\mathbf{z}_{t-1}} \\ \implies \quad & \alpha||\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t-1}||_\alpha^{\alpha-1}\frac{\partial||\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t-1}||_\alpha}{\partial\mathbf{z}_{t-1}} = \frac{\partial||h^{-1}(\mathbf{z}_t) - \mathbf{B}h^{-1}(\mathbf{z}_{t-1})||_\alpha^\alpha}{\partial\mathbf{z}_{t-1}} \\ \implies \quad & \alpha||\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t-1}||_\alpha^{\alpha-1}\frac{\partial||\epsilon_t||_\alpha}{\partial\epsilon_t}\frac{\partial\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t-1}}{\partial\mathbf{z}_{t-1}} = \frac{\partial||h^{-1}(\mathbf{z}_t) - \mathbf{B}h^{-1}(\mathbf{z}_{t-1})||_\alpha^\alpha}{\partial\mathbf{z}_{t-1}} \\ \implies \quad & -\alpha||\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t-1}||_\alpha^{\alpha-1}\frac{\partial||\epsilon_t||_\alpha}{\partial\epsilon_t}\mathbf{B} = \frac{\partial||h^{-1}(\mathbf{z}_t) - \mathbf{B}h^{-1}(\mathbf{z}_{t-1})||_\alpha^\alpha}{\partial\mathbf{z}_{t-1}} \\ \implies \quad & -\alpha(\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t-1}) \odot |\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t-1}|^{\alpha-2}\mathbf{B} = \frac{\partial||h^{-1}(\mathbf{z}_t) - \mathbf{B}h^{-1}(\mathbf{z}_{t-1})||_\alpha^\alpha}{\partial\mathbf{z}_{t-1}}. \end{aligned} \tag{37}$$

Make the same derivation process on the right hand and we obtain

$$\begin{aligned} & -\alpha(\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t-1}) \odot |\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t-1}|^{\alpha-2}\mathbf{B} \\ = & -\alpha(h^{-1}(\mathbf{z}_t) - \mathbf{B}h^{-1}(\mathbf{z}_{t-1})) \odot |h^{-1}(\mathbf{z}_t) - \mathbf{B}h^{-1}(\mathbf{z}_{t-1})|^{\alpha-2}\mathbf{B}\frac{\partial h^{-1}(\mathbf{z}_{t-1})}{\mathbf{z}_{t-1}}. \end{aligned} \tag{38}$$

For any $\mathbf{z}_t$ we can choose $\mathbf{z}_t = \mathbf{B}\mathbf{z}_{t-1}$ and thus the Equation (38) can be written as:

$$(h^{-1}(\mathbf{z}_t) - \mathbf{B}h^{-1}(\mathbf{z}_{t-1})) \odot |h^{-1}(\mathbf{z}_t) - \mathbf{B}h^{-1}(\mathbf{z}_{t-1})|^{\alpha-2}\mathbf{B}\frac{\partial h^{-1}(\mathbf{z}_{t-1})}{\mathbf{z}_{t-1}} = \mathbf{0}. \tag{39}$$

Considering the nonsingularity of matrix $\mathbf{B}$ (by assumption) and bijection of h, we can derive

$$\begin{aligned} & (h^{-1}(\mathbf{z}_t) - \mathbf{B}h^{-1}(\mathbf{z}_{t-1})) \odot |h^{-1}(\mathbf{z}_t) - \mathbf{B}h^{-1}(\mathbf{z}_{t-1})|^{\alpha-2} = \mathbf{0} \\ \implies \quad & (h_i^{-1}(\mathbf{z}_t) - \mathbf{B}h_i^{-1}(\mathbf{z}_{t-1}))|h_i^{-1}(\mathbf{z}_t) - \mathbf{B}h_i^{-1}(\mathbf{z}_{t-1})|^{\alpha-2} = 0 \end{aligned} \tag{40}$$

for all $i = 1, \ldots, d$. Apparently $h^{-1}(\mathbf{z}_t) = \mathbf{B}h^{-1}(\mathbf{z}_{t-1})$ is the only solution, thus

$$h^{-1}(\mathbf{B}\mathbf{z}_{t-1}) = \mathbf{B}h^{-1}(\mathbf{z}_{t-1}). \tag{41}$$

Substitute Equation (46) to the right hand of Equation (35), we have

$$\begin{aligned} & ||z_t - \mathbf{B}z_{t-1}||_\alpha^\alpha = ||h^{-1}(z_t) - \mathbf{B}h^{-1}(z_{t-1})||_\alpha^\alpha \\ \implies \quad & ||z_t - \mathbf{B}z_{t-1}||_\alpha^\alpha = ||h^{-1}(z_t) - h^{-1}(\mathbf{B}z_{t-1})||_\alpha^\alpha. \end{aligned} \tag{42}$$

This indicates that h^{-1} preserves the α-distances between points. Since h is bijective, then by Mazur-Ulam theorem Mazur & Ulam (1932), h must be an affine transform. According to Theorem 2 in (Zhang & Hyvärinen, 2011), the model is identifiable, which proves the theorem.

Extension to multiple time lags Similar to the simple case, we can extend the result in Equation (36) and have

$$\frac{\partial||\mathbf{z}_t - \sum_{\tau=1}^{L} \mathbf{B}_\tau \mathbf{z}_{t-\tau}||_\alpha^\alpha}{\partial \mathbf{z}_{t-i}} = \frac{\partial||h^{-1}(\mathbf{z}_t) - \sum_{\tau=1}^{L} \mathbf{B}_\tau h^{-1}(\mathbf{z}_{t-\tau})||_\alpha^\alpha}{\partial \mathbf{z}_{t-i}}, \tag{43}$$

where $\mathbf{z}_{t-i}$ is any lag latents with the limitation that the $\mathbf{B}_i$ corresponding to $\mathbf{z}_{t-i}$ is of full rank. Following the same above-mentioned derivation process, we can obtain

$$\begin{aligned}&(\mathbf{z}_t - \sum_{\tau=1}^{L} \mathbf{B}_\tau \mathbf{z}_{t-\tau}) \odot |\mathbf{z}_t - \sum_{\tau=1}^{L} \mathbf{B}_\tau \mathbf{z}_{t-\tau}|^{\alpha-2} \mathbf{B}_i \\ =&(h^{-1}(\mathbf{z}_t) - \sum_{\tau=1}^{L} \mathbf{B}_\tau h^{-1}(\mathbf{z}_{t-\tau})) \odot |h^{-1}(\mathbf{z}_t) - \sum_{\tau=1}^{L} \mathbf{B}_\tau h^{-1}(\mathbf{z}_{t-\tau})|^{\alpha-2} \mathbf{B}_i \frac{\partial h^{-1}(\mathbf{z}_{t-i})}{\mathbf{z}_{t-i}}\end{aligned}. \tag{44}$$

For any $\mathbf{z}_t$ we can choose $\mathbf{z}_t = \sum_{\tau=1}^{L} \mathbf{B}_\tau \mathbf{z}_{t-\tau}$, thus the Equation (44) can be written as:

$$(h^{-1}(\mathbf{z}_t) - \sum_{\tau=1}^{L} \mathbf{B}_\tau h^{-1}(\mathbf{z}_{t-\tau})) \odot |h^{-1}(\mathbf{z}_t) - \sum_{\tau=1}^{L} \mathbf{B}_\tau h^{-1}(\mathbf{z}_{t-\tau})|^{\alpha-2} \mathbf{B}_i \frac{\partial h^{-1}(\mathbf{z}_{t-i})}{\mathbf{z}_{t-i}} = \mathbf{0}. \tag{45}$$

As mentioned above, $h^{-1}(\mathbf{z}_t) = \sum_{\tau=1}^{L} \mathbf{B}_\tau h^{-1}(\mathbf{z}_{t-\tau})$ is the only solution, thus

$$h^{-1}(\sum_{\tau=1}^{L} \mathbf{B}_\tau \mathbf{z}_{t-\tau}) = \sum_{\tau=1}^{L} \mathbf{B}_\tau h^{-1}(\mathbf{z}_{t-\tau}). \tag{46}$$

Following the same procedure in the simple case, the theorem is proven. ■

B Experiment Settings

B.1 Synthetic Dataset

1. For nonparametric processes **(NP)**, we generate 150,000 data points according to Equation 2. The noises ϵ_{it} are sampled from i.i.d. normal distribution with variance modulated by nonstationary regime $u \in [1, 20]$. In each regime, the variance entries are uniformly sampled between $[0, 1]$. A 2-layer MLP with LeakyReLU units is used as the state transition function f_i. When we need sparse causal structure, a random input mask is added.
2. For parametric processes **(VAR)**, we generate 50,000 data points according to Equation 4. The noises ϵ_{it} are sampled from i.i.d. Laplace distribution $(\sigma = 0.1)$. The entries of state transition matrices $\mathbf{B}_\tau$ are uniformly distributed between $[-0.5, 0.5]$.
3. **(Generalized condition 1)** For time-variant causal relations, we generate 240,000 data points according to Equation 47.

$$\mathbf{x}_t = g(\mathbf{z}_t), \quad \mathbf{z}_t = \sum_{\tau=1}^{L} (\mathbf{B}_\tau|u)\mathbf{z}_{t-\tau} + \epsilon_t \quad with \quad \epsilon_{it} \sim p_{\epsilon_i} \tag{47}$$

   The noises ϵ_{it} are sampled from i.i.d. Laplace distribution with $\sigma = 0.1$. The entries of state transition matrices $\mathbf{B}_\tau$ are uniformly distributed between $[-0.5, 0.5]$ which modulated by time segment u across time.
4. **(Generalized condition 2)** For instantaneous causal relations, we generate 45,000 data points according to Equation 48.

$$\mathbf{x}_t = g(\mathbf{z}_t), \quad \mathbf{z}_t = \mathbf{A}\mathbf{z}_t + \sum_{\tau=1}^{L} \mathbf{B}_\tau \mathbf{z}_{t-\tau} + \epsilon_t \quad with \quad \epsilon_{it} \sim p_{\epsilon_i} \tag{48}$$

   where matrix $\mathbf{A}$ is a directed acyclic graph (DAG) which contains the coefficients of the linear instantaneous influence. The noises ϵ_{it} are sampled from i.i.d. Laplace distribution with $\sigma = 0.1$. The entries of state transition matrices $\mathbf{B}_\tau$ are uniformly distributed between $[-0.5, 0.5]$.

B.2 THE MCC METRIC

Following the common evaluation practice as in Independent Component Analysis (ICA), we employ MCC for performance evaluation. A high MCC represents the accurate recovery of the latents. First, MCC calculates the correlation coefficients between the ground-truth factors and the recovered latent variables. The possible permutation problem is addressed by solving a linear sum assignment problem in polynomial time. After finding the correct permutation, the correlation coefficients are computed again between ground-truth factors and the permuted latents. This is the reason why the MCC metric remains invariant by permutation and component-wise transformations. We use the Pearson coefficient as our correlation measure which evaluates the linear relationship between predictions and ground-truth values.

C IMPLMENTATION DETAILS

In this section, we present the disentangled encoder/decoder design for multi-object video dataset and the implementation of model ablations. The training settings are also summarized.

C.1 KEYPOINT DETECTOR

For the multi-object video dataset in this paper, the time-delayed causal variables use objects as the building blocks to factorize the scene. The object-centric representations contain object locations and some other attributes (e.g., color, size, etc.). We thus use the disentangled visual encoder/decoder design in Transporter (Kulkarni et al., 2019). Two separate CNNs, one for extracting visual features and the other for locating object locations with a spatial softmax unit. The decoder retrieves object features from feature maps using object locations and reconstructs the scene. During reconstruction, the visual object feature of the first video frame are directly "transported" to later frames by (1) removing original visual features in the first frame, and (3) rendering new visual features at the new object locations given by inference network changing across time.

C.2 MODEL ABLATIONS

We start with BetaVAE and add our proposed modules successively as model ablations. When causal process prior network is added to the baseline, the variant is equipped with recurrent inference network and learnable transition functions. However, during estimation of KL divergence, we use MSE loss directly, which corresponds to stationary Gaussian noise distribution, to replace the nonstationary flow estimator for evaluating the prior probability. When nonstationary flow estimator is further added to the variant, it's implemented by only disabling the noise discriminator from LEAP and do not change any training settings.

C.3 HYPERPARAMETERS AND TRAINING

Our framework is implemented in PyTorch. The VAE model is trained using AdamW optimizer for a maximum of 200 epochs and early-stops if validation error does not decrease for 5 epochs. Learning rate of 0.002 and mini-batch size of 32 are used. For the noise discriminator, we used SGD optimizer with the learning rate 0.001.

D RELATED WORK

Recent advances in identifiability provide a principled framework for unsupervised latent recovering. Several methods have been proposed in the literature and mainly focused on two aspects: 1) observed variables are assumed to be time series and have stationary temporal dependencies, 2) latent variables are conditional independent on some observed auxiliary variables which raise the nonstationarity. For stationary temporal dependency, Hyvarinen & Morioka (2017) developed the logistic regression network to separate the independent sources. Inspired by the idea of contrastive learning, their approach learns to discriminate between the ordered observed temporal window and permuted temporal window and the model is identifiable under the uniformly dependent assumption. Regarding nonstationarity in temporal data, Hyvarinen & Morioka (2016) proposes time-contrastive

learning (TCL) based on the independent sources assumption. It gives the very first identifiability proof for a nonlinear mixing model with limitations in terms of inaccurate and computationally expensive nonstationary data segmentation. Hälvä & Hyvarinen (2020) combines nonlinear ICA with a Hidden Markov Model (HMM) to remedy these problems, which can automatically model nonstationary data and remove the need for manual data segmentation. Along the same lines, Hyvarinen et al. (2019) subsequently augmented the observed data with an auxiliary variable and the nonlinear ICA was realized by distinguishing between ordered augmented data and data with randomized auxiliary variables. They assumed the distributions of the independent sources are changed conditioned on the auxiliary variables under the restriction to the exponential family. Building on these prior work, (Khemakhem et al., 2020) introduced VAE to nonlinear ICA and approximated the true joint distribution over observed and latent variables. Although the method is general enough to be compatible with any parameter estimation method, they needed to make additional assumptions on parametric conditions and auxiliary variables. A more recent development in causal-related nonlinear ICA is given by (Yang et al., 2021). They introduced a linear causal layer to transform independent exogenous factors into endogenous causal variables, while the assumptions on the auxiliary variable and parametric conditions are still needed.