MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Carnegie Mellon University

Causal Representation Learning

Guangyi Chen

https://chengy12.github.io/

Carnegie Mellon University, Pittsburgh PA, USA

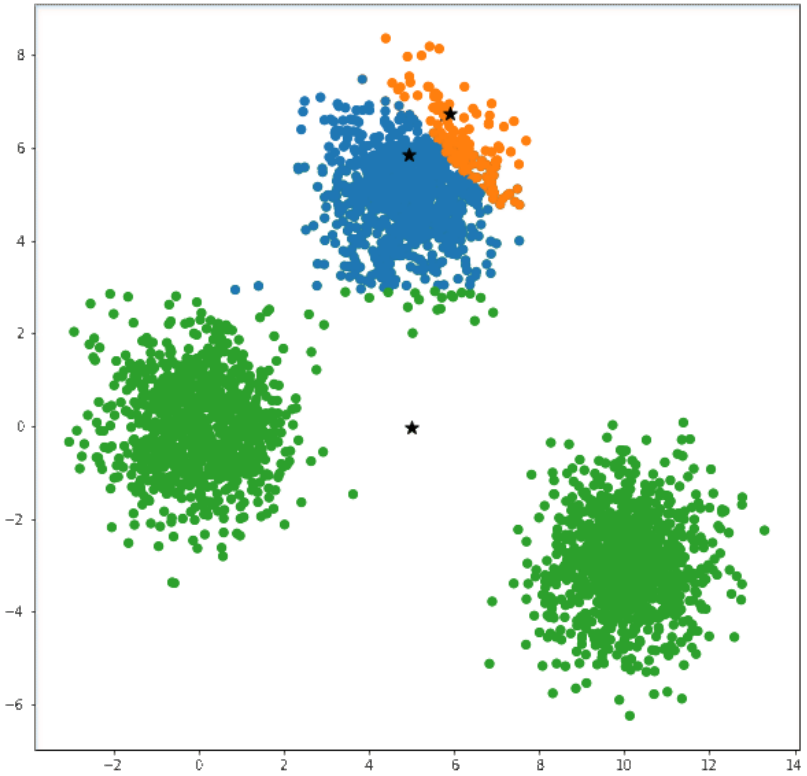
Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

Swarma Pattern, 1st September

Unsupervised Representation Learning

- In most cases, we need to learn the representation in an unsupervised may

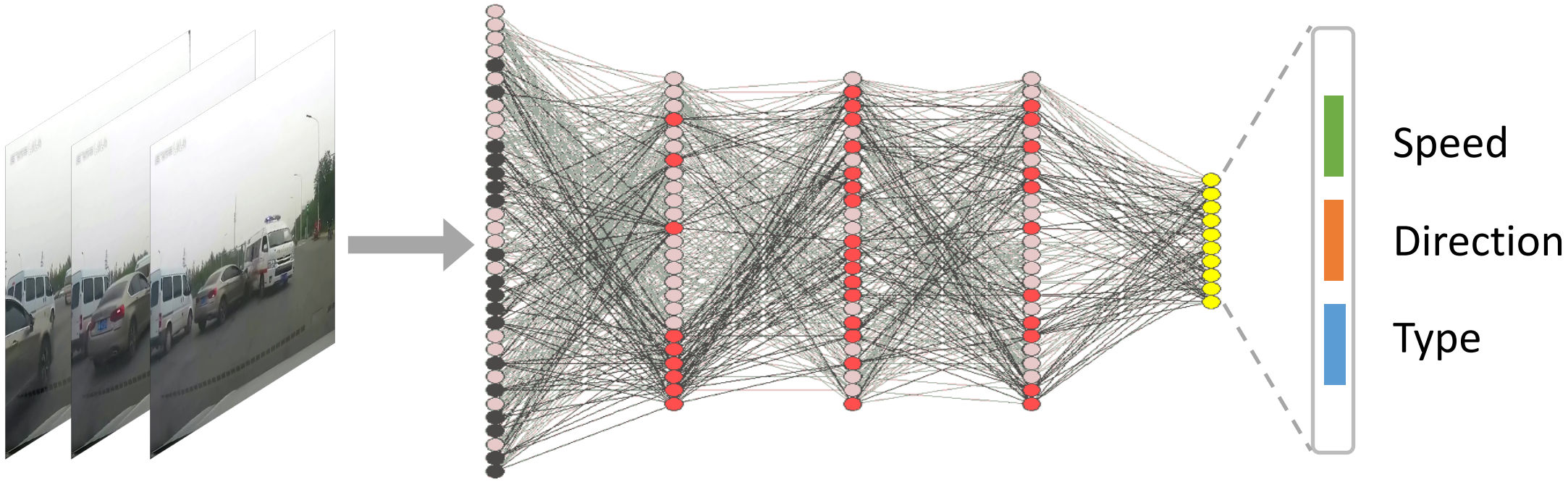
Massive unsupervised data



- We don’t know what we are looking for
- Labels are not always available
- Labels might not be informative

Disentangled Representation

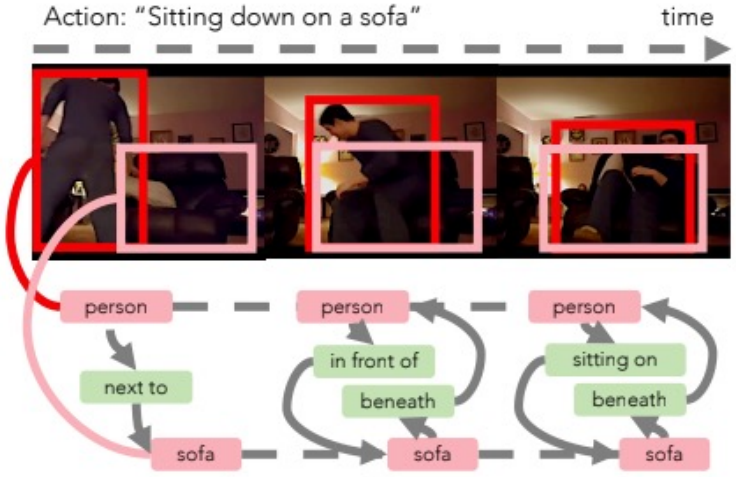
➢ What is good representation?



Compact, discriminative, high-level, informative, independent, disentangled, explainable, controllable

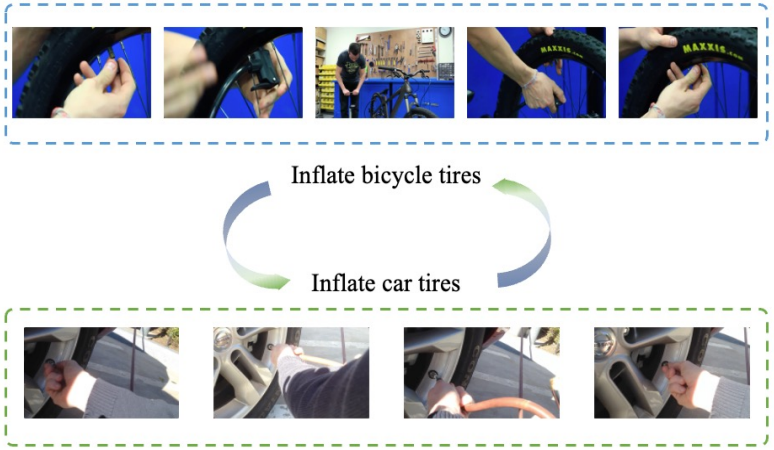
Disentangled Representation

- The advantages of disentangled representation



(Ji et al., 2020)

Explainable

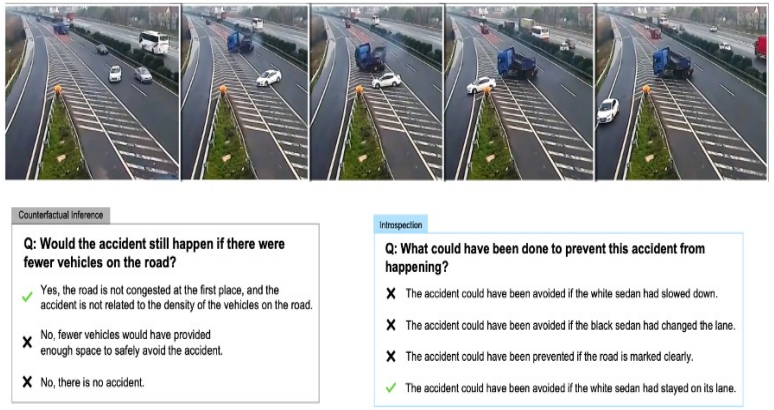


(Li et al., 2023)

Transferable

(Xiang et al., 2024)

Controllable

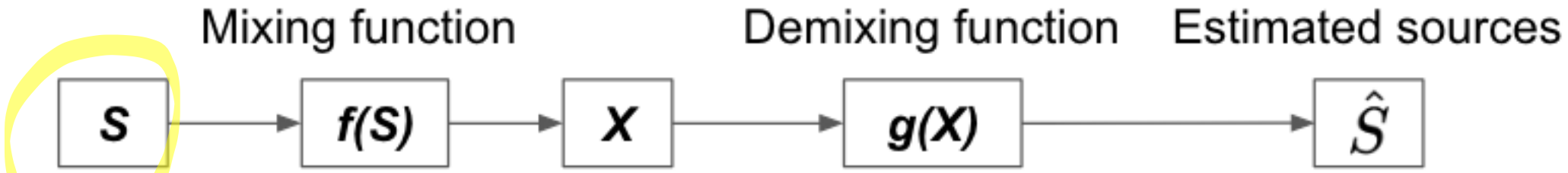


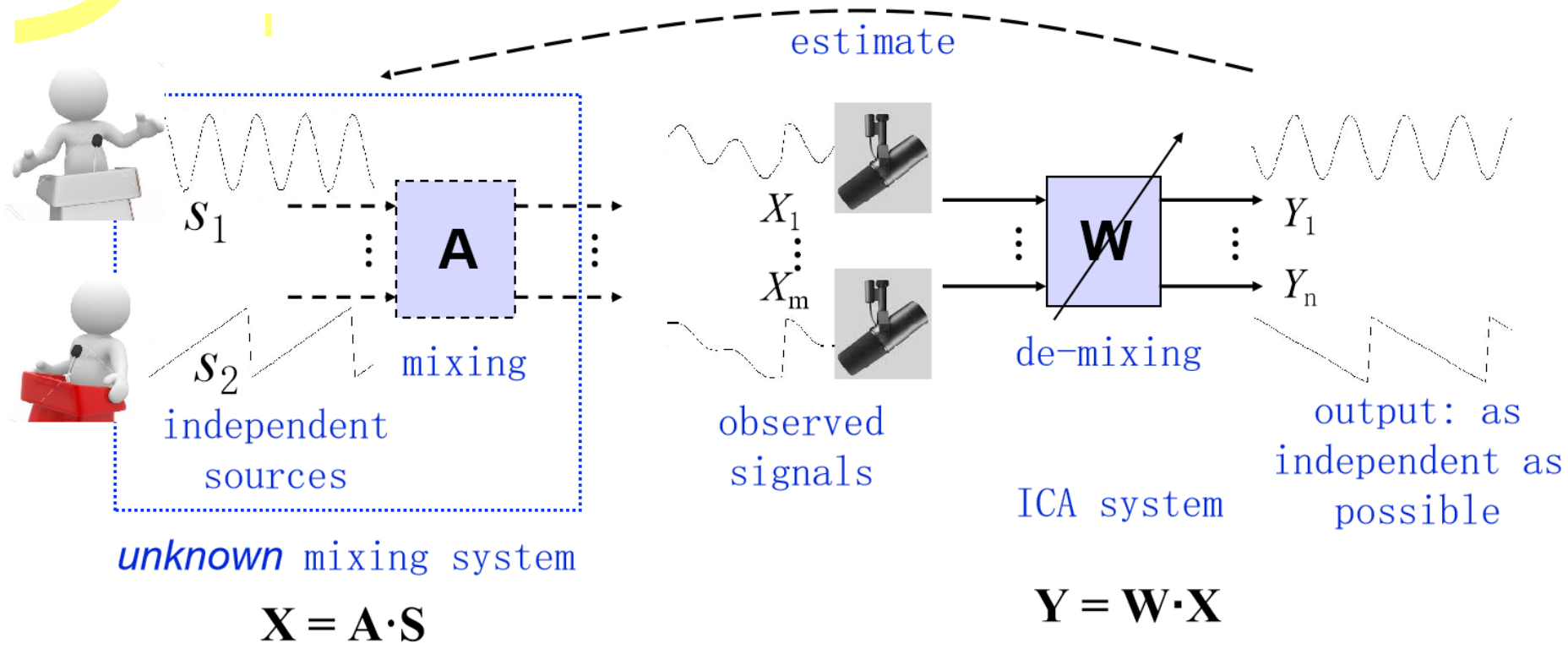
(Xu et al., 2021)

Reasonable

Independent Component Analysis Framework

- One potential formalization: Disentanglement via Independent Component Analysis

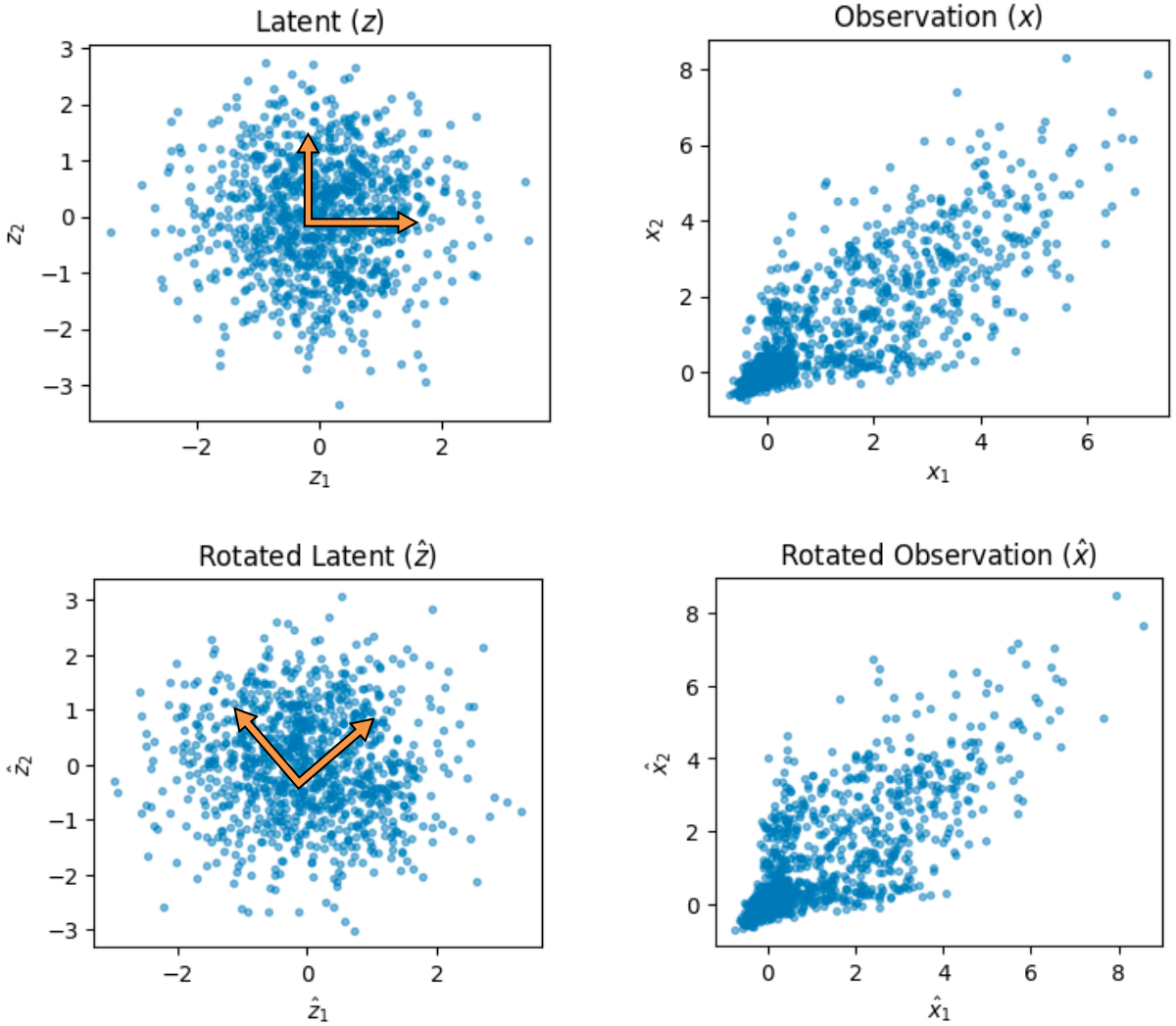




[Credit to Kun]

Causal Representation Learning is Hard

- For the linear Gaussian case, when we add a rotation on the latent variable, the generated observation distribution is unchanged.



- Original generation process

$$z \sim \mathcal{N}(\mathbf{0}, I) \qquad x = g(z)$$

- Adding a rotation on the latent variable doesn't change the observation

$$\hat{z} = \sigma(z) \qquad \hat{x} = g(\hat{z})$$

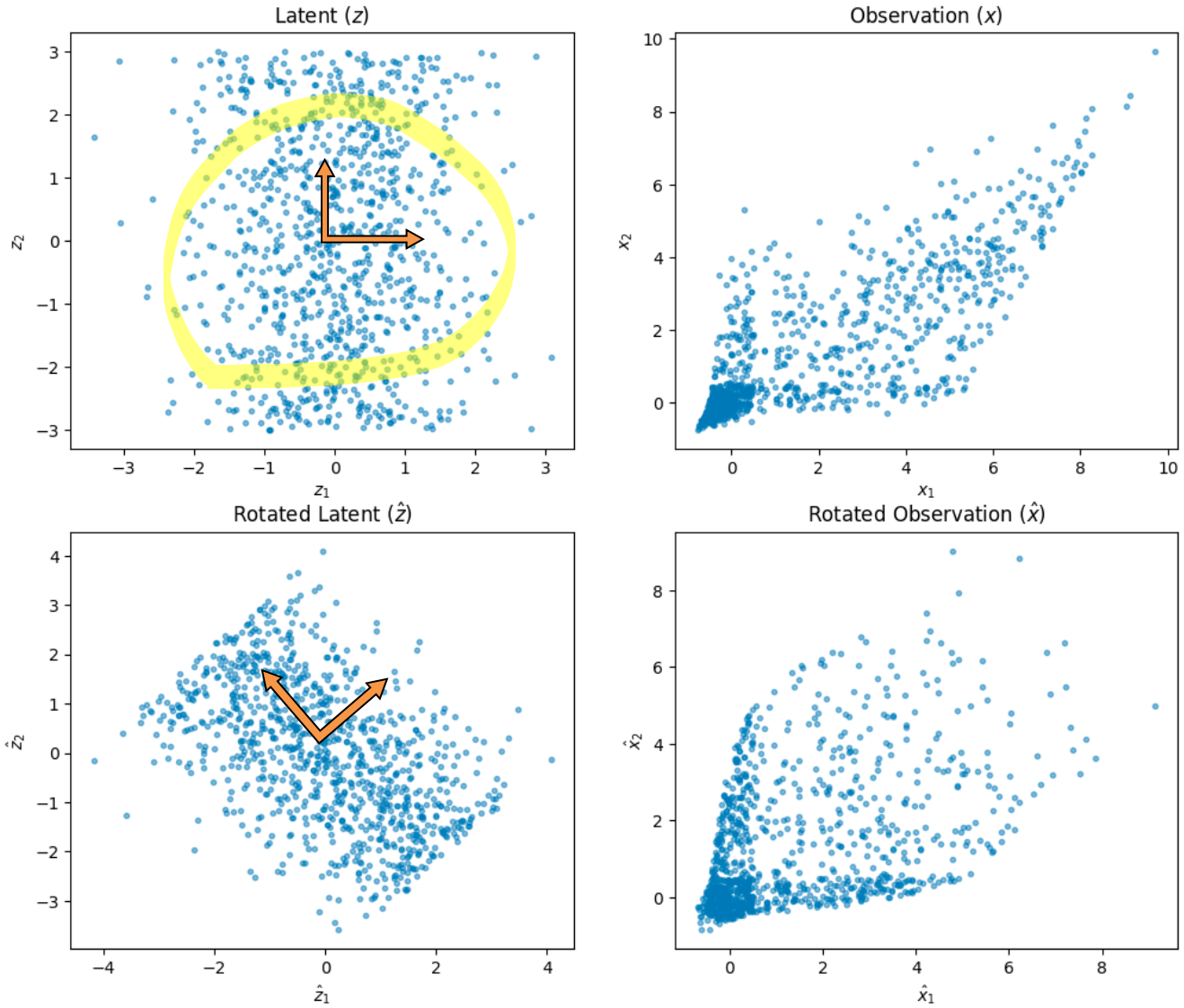
$$p(\hat{z}) = p(\sigma(z)) = p(z)$$

- We may learn the representations with the entanglement

$$\hat{g}^{-1} = g^{-1} \qquad \hat{g}^{-1} = \sigma \circ g^{-1}$$

Causal Representation Learning Is Hard

- For the linear non-Gaussian case, we can recover the latent variables.

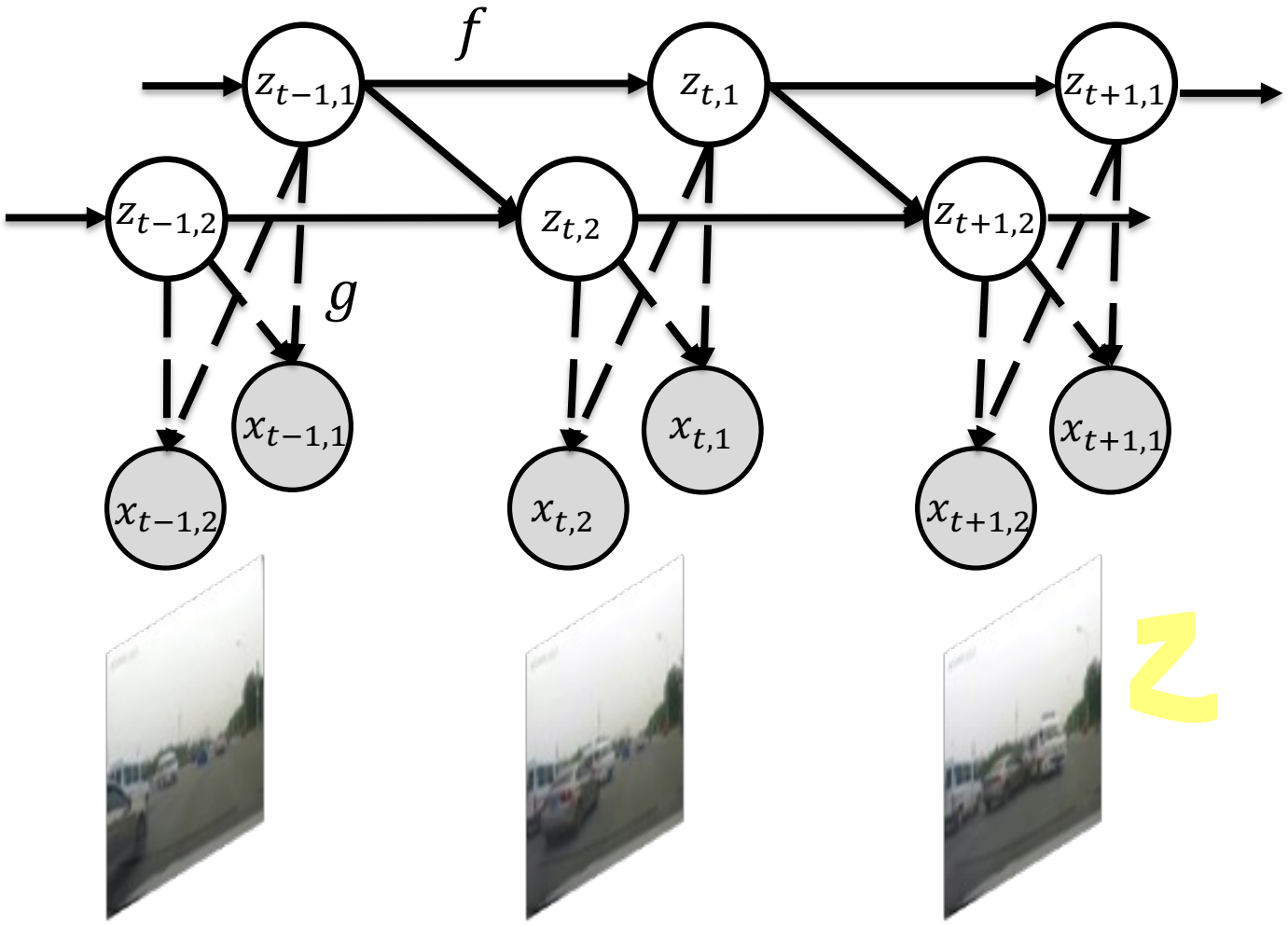


- The Non-Gaussian property provides the "changeability" (sufficient change)
- By maximizing non-Gaussianity, we could identify true sources.

[Credit to Xiangchen]

Formulation

- Causal representation learning (CRL) aims to recover the data generation process from the observation to obtain the disentangled latent representation.



- Data generation process:

$$x_t = g(z_t)$$

- Latent causal process:

$$z_{it} = f_i(\mathbf{Pa}(z_{it}), \varepsilon_{it})$$

- Representation learning:

$$\hat{z}_t = \hat{g}^{-1}(x_t)$$

- Component-wise identifiability

$$p_{\hat{g},\hat{f},\hat{p}_\epsilon}(x_t) = p_{g,f,p_\epsilon}(x_t)$$
$$\Longrightarrow \quad \hat{g} = g \circ T \circ \pi$$

Causal Representation Learning

- When the learned representation can be identifiable and how to learn?
- What can we do if causal representation is learned?

Identification Theorem

Real-world Application

Identification Theorem with Conditional Independence

Generation Process contains the stationary latent temporal dynamic transition and the invertible mixing function:

$$\mathbf{x}_t = \mathbf{g}(\mathbf{z}_t), \quad z_{it} = f_i\left(\mathbf{z}_H, \epsilon_{it}\right).$$

Theorem (Identifiability under Stationary Process). *For a series of observations* $\mathbf{x}_t$ *and estimated latent variables* $\hat{\mathbf{z}}_t$, *suppose there exists function* $\hat{\mathbf{g}}$ *which is subject to observational equivalence,*

$$\mathbf{x}_t = \hat{\mathbf{g}}(\hat{\mathbf{z}}_t).$$

If assumptions

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

$$\mathbf{v}_{lt} \triangleq \left(\frac{\partial^2 \eta_{1t}}{\partial z_{1t} \partial z_{l,H}}, \frac{\partial^2 \eta_{nt}}{\partial z_{nt} \partial z_{l,H}}, \frac{\partial^3 \eta_{1t}}{\partial z_{1t}^2 \partial z_{l,H}}, \frac{\partial^3 \eta_{nt}}{\partial z_{nt}^2 \partial z_{l,H}}\right)^\intercal,$$

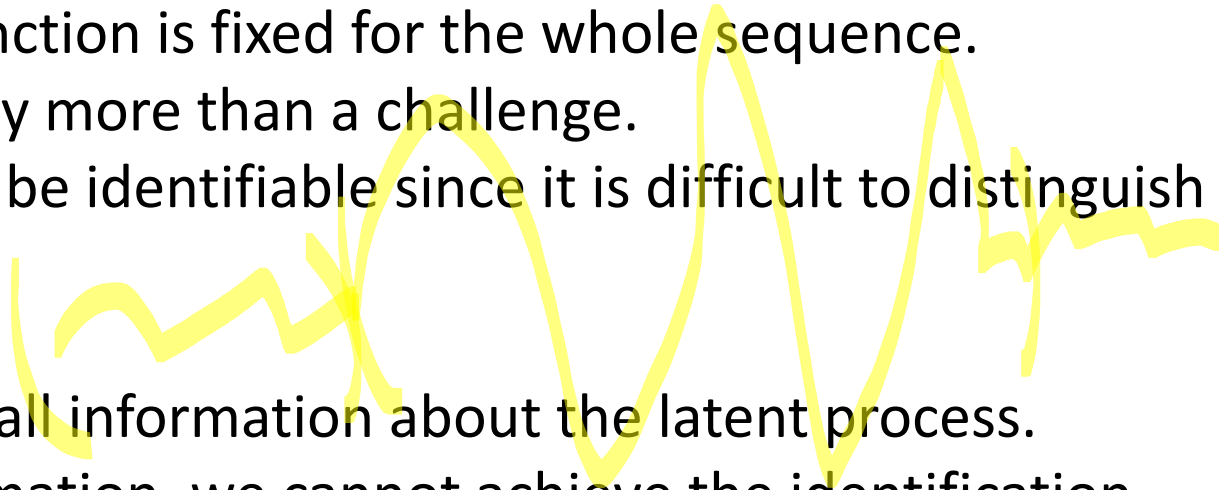
for $l = 1, 2, \cdots, n$. *For each value of* $\mathbf{z}_t$, *there exists* $2n$ *different values of* $z_{l,H}$ *such that the* $2n$ *vector functions* $\mathbf{v}_{lt} \in \mathbf{R}^{2n}$ *are linearly independent,*

are satisfied, then $\mathbf{z}_t$ *must be a component-wise transformation of a permuted version of* $\hat{\mathbf{z}}_t$.

(Yao, Chen, & Zhang, 2022)

Identification Theorem with Conditional Independence

- Stationary: The temporal transition function is fixed for the whole sequence.
  Known non-stationary is an opportunity more than a challenge.
  Unknown non-stationary cases cannot be identifiable since it is difficult to distinguish the domain change or variable change.



- Invertibility: The observation contains all information about the latent process.
  If we cannot recover the missing information, we cannot achieve the identification.

- Conditional independence: there are no instantaneous relations among the latent variables.
  If two variables always change jointly, we cannot say what is the unique effect of each.

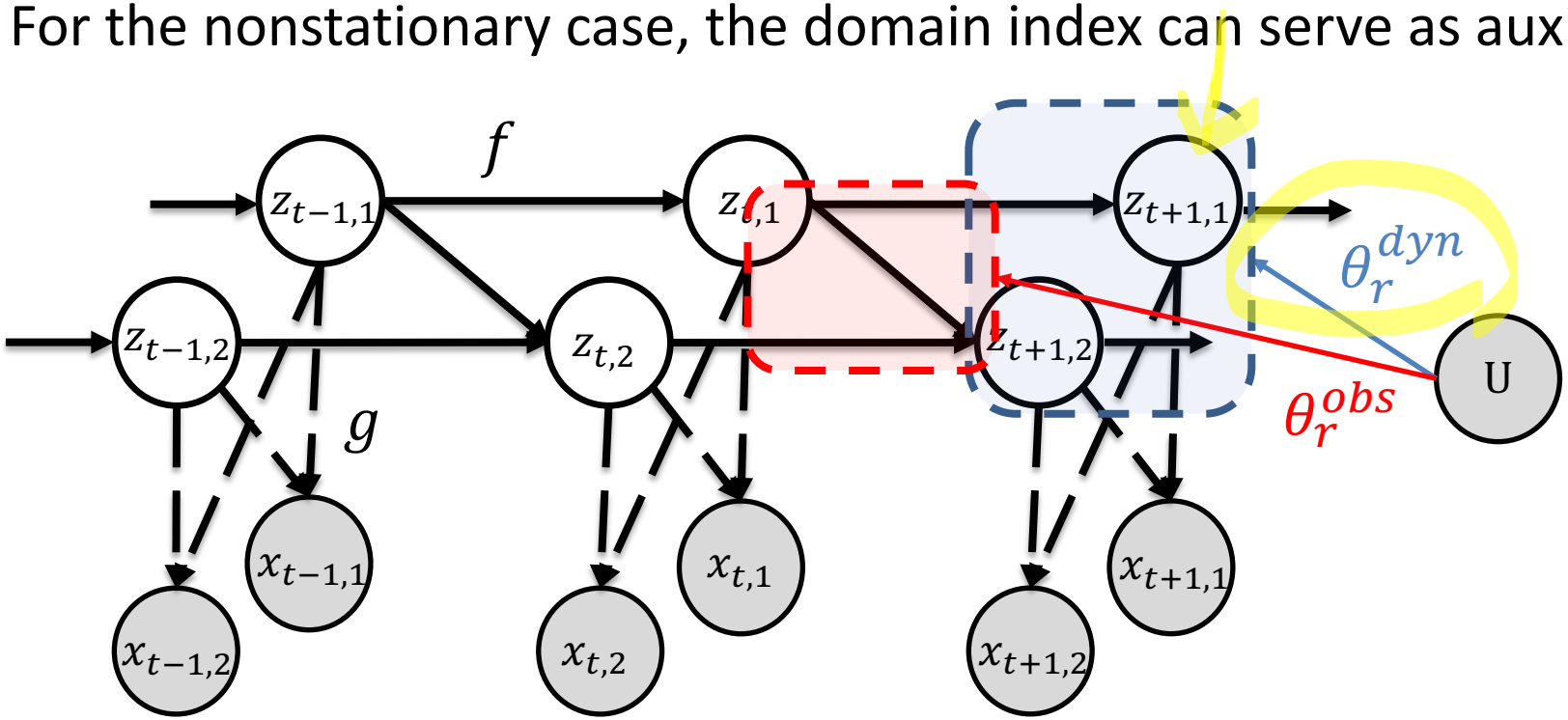
- Sufficiency: It means that the conditional independent change of the latent variables has sufficient influence and these changes can be captured from the observation.
  Image: requiring domain index or labels
  Video: using the historical state as the side information

Identification under More Challenges

Challenges	Scenarios	Extra Assumption	Reference
Unknown non-stationary	Bioinformatics, Speech, Volleyball Game	1) Domain index follows the Markov process, 2) Mechanism Sparsity	(Song et al., 2023) (Song et al., 2024) (Li et al., 2024)
Non-invertible mixing function	Occlusion, Video, Motion Blue	Context exists to complete the missing information.	(Chen et al., 2024)
Instantaneous relations	Skeleton, Stock	Sparse relations	(Li et al., 2024)

Nonstationary Provides More Information

- For the nonstationary case, the domain index can serve as auxiliary variables

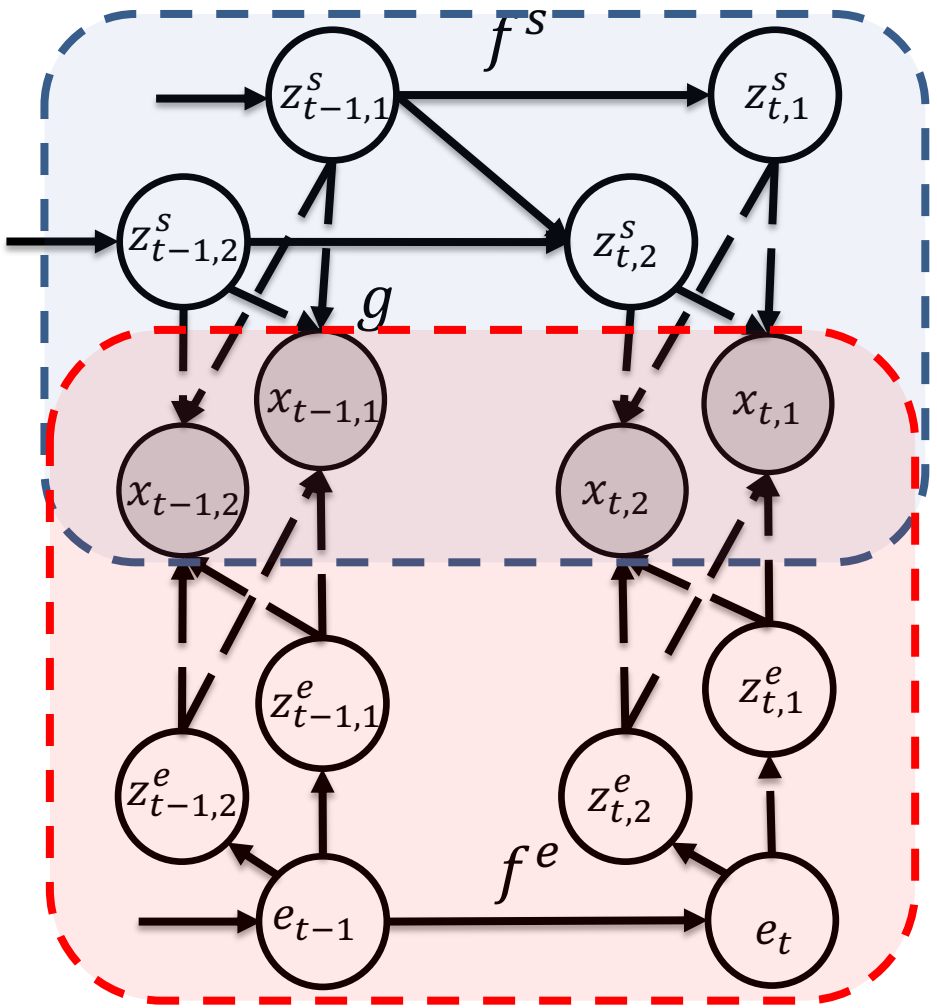


Domain information is observed

$$
\begin{cases}
z_{s,t}^{\text{fix}} &= f_s\left(\left\{z_{i,t-\tau} \middle| z_{i,t-\tau} \in \mathbf{Pa}(z_{s,t}^{\text{fix}})\right\}, \epsilon_{s,t}\right), \\
z_{c,t}^{\text{chg}} &= f_c\left(\left\{z_{i,t-\tau} \middle| z_{i,t-\tau} \in \mathbf{Pa}(z_{c,t}^{\text{chg}})\right\}, \boldsymbol{\theta}_r^{\text{dyn}}, \epsilon_{c,t}\right), \\
z_{o,t}^{\text{obs}} &= f_o\left(\boldsymbol{\theta}_r^{\text{obs}}, \epsilon_{o,t}\right), \\
\mathbf{x}_t &= \mathbf{g}(\mathbf{z}_t).
\end{cases}
$$

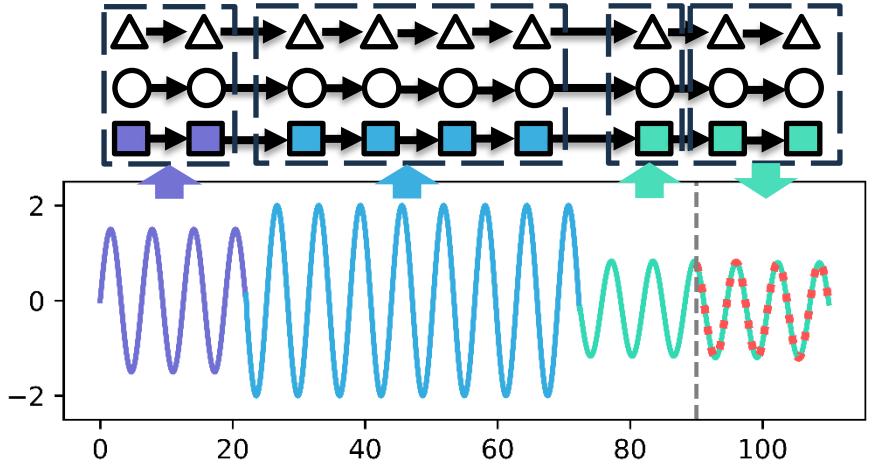
Nonstationary Information Is Unknown

➢ When the nonstationary information cannot be observed, we need extra assumptions.



Stationary process

Nonstationary process



➢ Data generation process:

$$x_t = g(z_t^e, z_t^s)$$

➢ Stationary latent causal process:

$$z_{it}^s = f_i^s(\mathbf{Pa}(z_{it}^s), \varepsilon_{it}^s)$$

➢ Nonstationary latent causal process

$$e_1, \ldots, e_T \sim Markov\ Chain(A)$$

$$z_{jt}^e = f_j^e(e_t, \varepsilon_{it}^e)$$

Identification Theorem under Unknown Nonstationary

Data generation process: Contain stationary and nonstationary latent temporal transition and the invertible mixing function.

$$x_t = g(z_t^e, z_t^s), \quad z_{it}^s = f_i^s(\mathbf{Pa}(z_{it}^s), \varepsilon_{it}^s), \quad z_{it}^e = f_i^e(e_t, \varepsilon_{it}^e)$$

Theorem (Block-wise Identifiability of the Stationary and Nonstationary Latent Variables.) We follow the data generation process with stationary and nonstationary latent variables, then if the following assumptions

(Latent Markov Process): The latent environments e_t are generated from a Markov process

(Smooth and Positive Density): The probability density function of latent variables is smooth and positive.

(Linear Independent): For any $z_t^e \subseteq R^{n_e}$, $\mathrm{v}_{1,t-1}, \ldots, \mathrm{v}_{l,t-1}, \ldots \mathrm{v}_{n_e,t-1}$ as n_e vector functions in $z_{1,t-2}, \ldots z_{n_e,t-2}$ are linear independent, where $\mathrm{v}_{l,t-1}$ are formalized as follows:

$$\mathrm{v}_{l,t-1} = \frac{\partial^2 \log p(z_t^e | z_{t-1}^e, z_{t-2}^e)}{\partial z_{k,t}^e \partial z_{l,t-2}^s}$$

(Variability of Historical Information): There exist two values of $\boldsymbol{u} = \{z_{t-1}^e, z_{t-2}^e\}, i.e., \boldsymbol{u}_1$ and $\boldsymbol{u}_2$, s.t., for any set $A_{z_t} \subseteq Z_t$ with non-zero probability measure and A_{z_t} cannot be expressed as $B_{z_t} \times Z_t^e$, for any $B_{z_t} \subset Z_t^s$, we have:

$$\int_{z_t \in A_{z_t}} p(z_t | \boldsymbol{u}_1) d_{z_t} \neq \int_{z_t \in A_{z_t}} p(z_t | \boldsymbol{u}_2) d_{z_t}$$

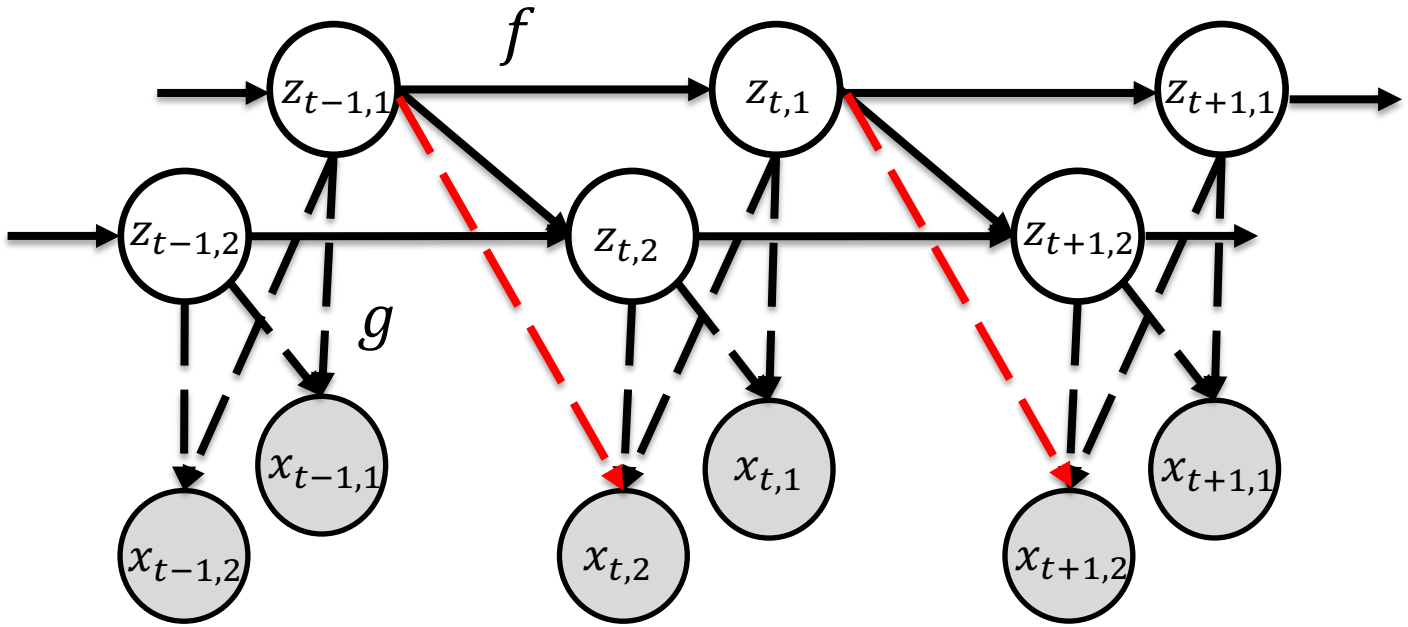
Then, by learning the data generation process, the stationary and nonstationary variables are block-wise identifiable.

(Li et al., 2024)

Identification under Non-Invertibility

- Usually, the maxing function is non-invertible, caused by occlusion or mappings like 3d-2d

(a) Non-invertibility by occlusion



- Non-invertible data generation process:

$$x_t = g(z_{t:t-\tau})$$

- Latent causal process:

$$z_{it} = f_i(\mathbf{Pa}(z_{it}), \varepsilon_{it})$$

- There exists a mapping from x to z, such as:

$$z_t = m(x_{t:t-u})$$

- Key intuition: leveraging the context to recover the lost information.

Identification under Non-Invertibility

Non-invariable Generation Process contains the stationary latent temporal dynamic transition and the non-invertible mixing function, and there exist a mapping from $\mathbf{x}_{t:t-\mu}$ to $\mathbf{z}_t$:

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

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

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

If assumptions

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

$$\mathbf{v}_{lt} \triangleq \left(\frac{\partial^2 \eta_{1t}}{\partial z_{1t}\partial z_{l,t-r-1}}, \cdots, \frac{\partial^2 \eta_{nt}}{\partial z_{nt}\partial z_{l,t-r-1}}, \frac{\partial^3 \eta_{1t}}{\partial z_{1t}^2\partial z_{l,t-r-1}}, \cdots, \frac{\partial^3 \eta_{nt}}{\partial z_{nt}^2\partial z_{l,t-r-1}}\right)^\intercal,$$

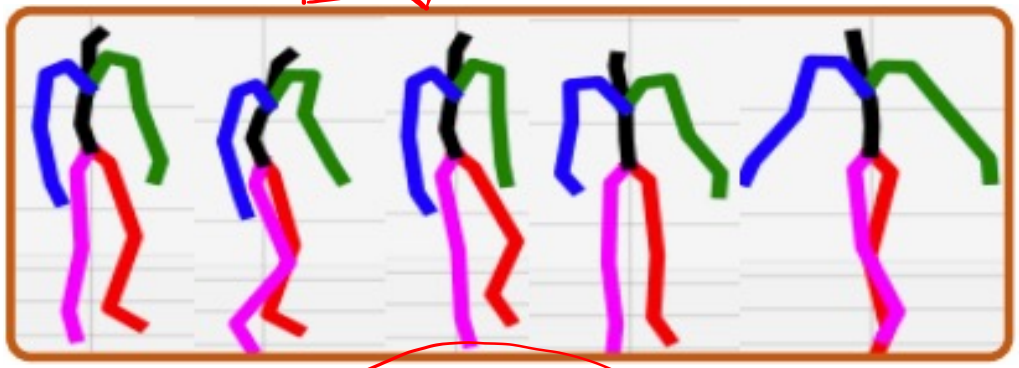
  for $l = 1, 2, \cdots, n$. For each value of $\mathbf{z}_t$, there exists $2n$ different values of $z_{l,t-r-1}$ such that the $2n$ vector functions $\mathbf{v}_{lt} \in \mathbf{R}^{2n}$ are linearly independent,

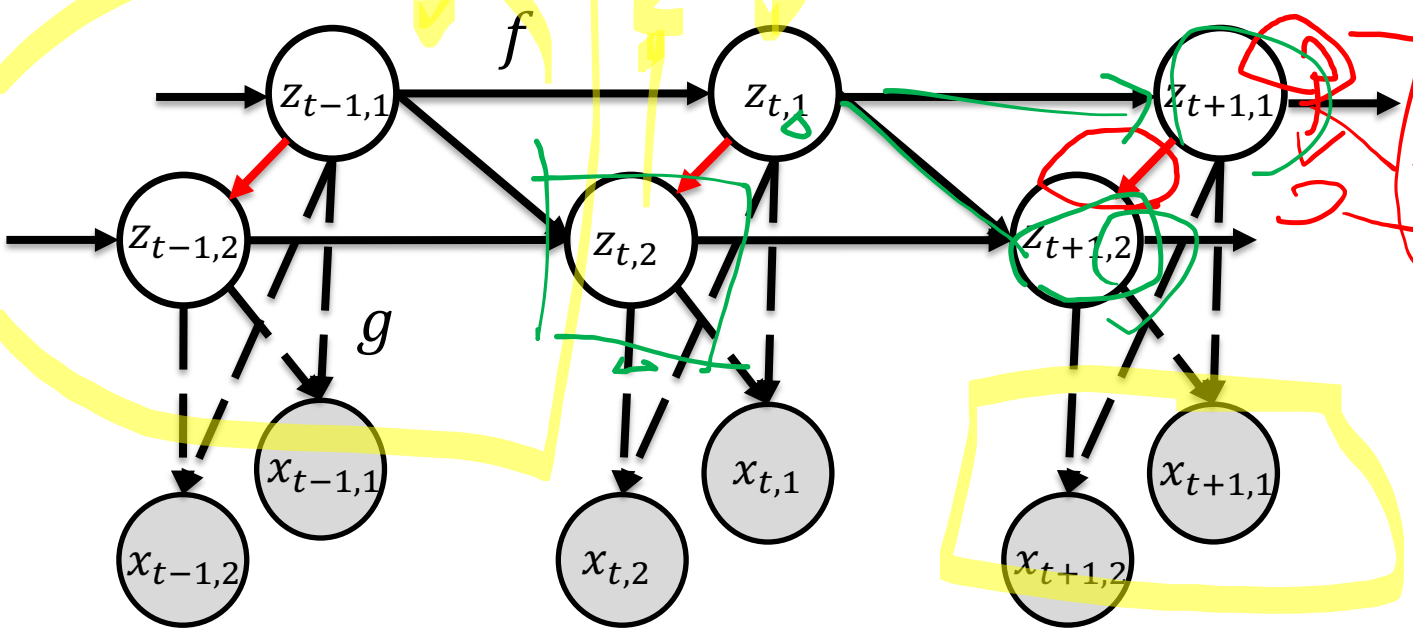
are satisfied, then $\mathbf{z}_t$ must be a component-wise transformation of a permuted version of $\hat{\mathbf{z}}_t$ with regard to context $\{\mathbf{x}_j \mid \forall j = t, t-1, \cdots, t-\mu-r\}$.

(Chen et al., 2024)

Instantaneous Dependency

- Current methods assume the absence of instantaneous dependencies, which may not hold true when the sampling frequency is low, or with solid relations.





- Data generation process:

$$x_t = g(z_t)$$

- Latent causal process with time-delay and instantaneous dependencies:

$$z_{it} = f_i(\mathbf{Pa_d}(z_{it}), \mathbf{Pa_e}(z_{it}), \varepsilon_{it})$$

Identification Theorem under Instantaneous Dependency

Data generation process: Contain stationary and nonstationary latent temporal transition and the invertible mixing function.

$$x_t = g(z_t), \quad z_{it} = f_i(\mathbf{Pa_d}(z_{it}), \mathbf{Pa_e}(z_{it}), \varepsilon_{it})$$

Theorem 1 (Relationships between Ground-truth and Estimated Latent Variables): For a series of observations x_t and estimated latent variables $\hat{\boldsymbol{z}}_t$, suppose the process subject to observational equivalence $x_t = \hat{g}(\hat{z}_t)$. Let $c_t = \{z_{t-1}, z_t\}$ and M_{c_t} be the variable set of two consecutive timestamps and the corresponding Markov network respectively. Suppose the following assumptions hold

(Smooth and Positive Density): The probability density function of latent variables is smooth and positive.

(Sufficient Variability): Denote $|M_{c_t}|$ as the number of edges in Markov network M_{c_t}, let $w(m)$ are linearly independent.

$$w(m) = \left(\frac{\partial^3 \log p(\boldsymbol{c}_t|z_{t-2})}{\partial c_{1,t}^2 \partial z_{m,t-2}}, \ldots \frac{\partial^3 \log p(\boldsymbol{c}_t|z_{t-2})}{\partial c_{2n,t}^2 \partial z_{m,t-2}}\right) \oplus \left(\frac{\partial^2 \log p(\boldsymbol{c}_t|z_{t-2})}{\partial c_{1,t} \partial z_{m,t-2}}, \ldots \frac{\partial^2 \log p(\boldsymbol{c}_t|z_{t-2})}{\partial c_{2n,t} \partial z_{m,t-2}}\right) \oplus \left(\frac{\partial^3 \log p(\boldsymbol{c}_t|z_{t-2})}{\partial c_{i,t} \partial c_{j,t} \partial z_{m,t-2}}\right)$$

$$\text{Then } \frac{\partial c_{i,t}}{\partial \hat{c}_{k,t}} \cdot \frac{\partial c_{i,t}}{\partial \hat{c}_{l,t}} = 0, \frac{\partial c_{i,t}}{\partial \hat{c}_{k,t}} \cdot \frac{\partial c_{j,t}}{\partial \hat{c}_{l,t}} = 0, \frac{\partial^2 c_{i,t}}{\partial \hat{c}_{k,t} \partial \hat{c}_{l,t}} = 0$$

Theorem 2 (Component-wise Identification of Latent Variables with instantaneous dependencies.) Except for the smooth, positive density and sufficient variability assumptions, we further make the following assumption:

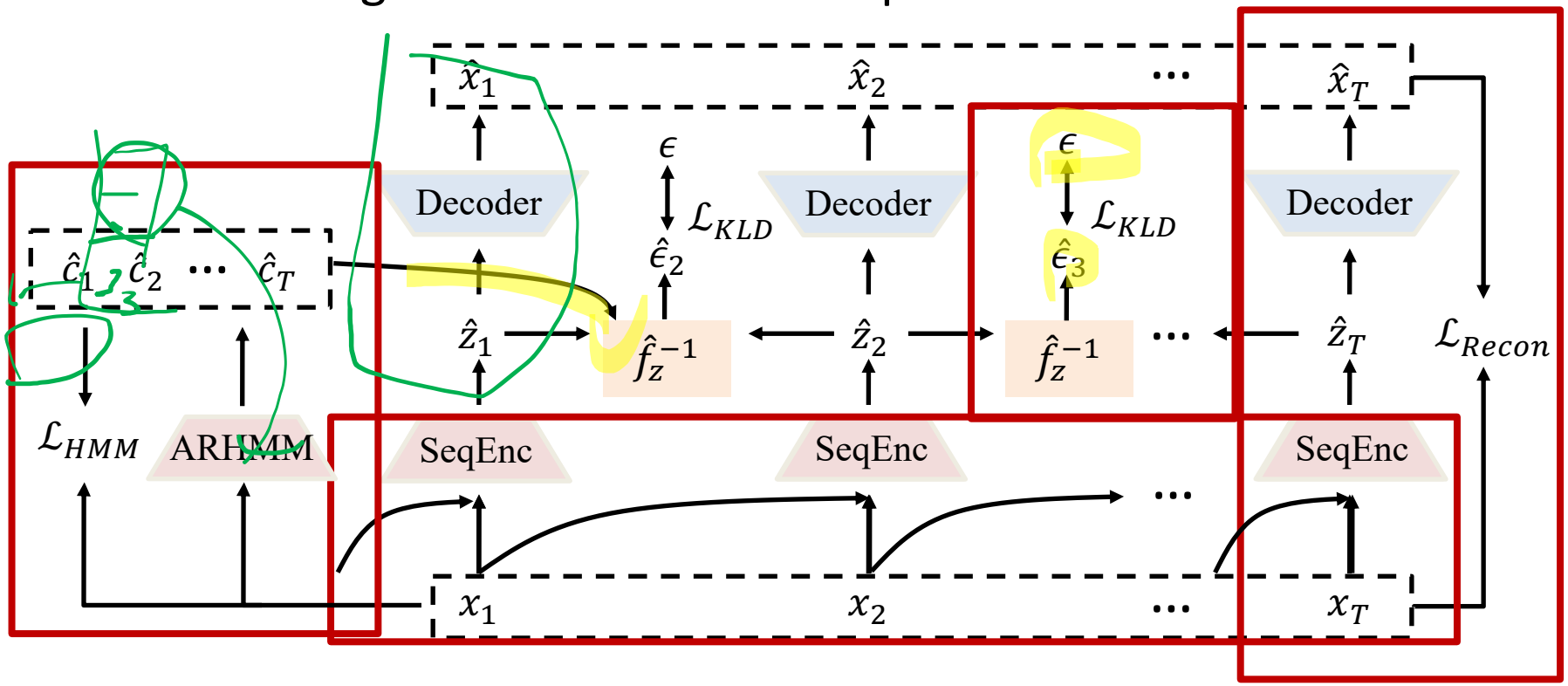
(Latent Process Sparsity): For any $z_{i,t}$, the intimate neighbor set of $z_{i,t}$ is an empty set.

When the observational equivalence is achieved with the minimal number of edges of the estimated Markov network of $\widehat{M}_{c_t}$, then the latent variables are component-wise identifiable.

(Li & Chen et al, 2024)

Generative Learning Framework

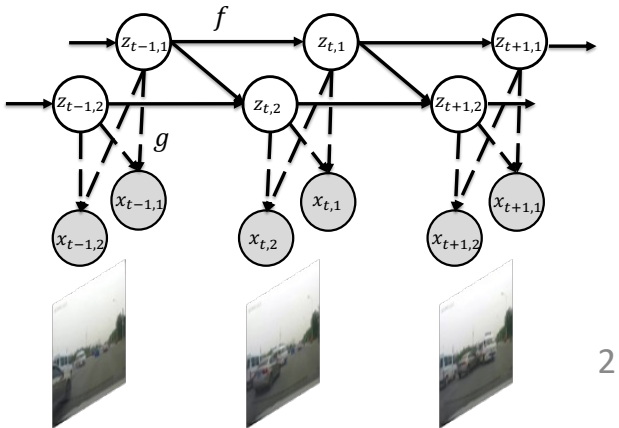
- **Auto-Encoder Model**: Estimates the mixing function $x_t = g(z_t)$ and de-mixing function $\hat{z}_t = \hat{g}^{-1}(x_t)$. Invertibility is enforced by reconstruction loss.
- **Prior Network**: Estimates the prior distribution $p(z_t|z_H)$ by learning the inverse dynamics f_z^{-1}, by change of variable $p(z_t|z_H) = p_\varepsilon\left(f_z^{-1}(z_t, z_H)\right)\left|\frac{\partial f_z^{-1}}{\partial z_t}\right|$. By constraining this prior, we encourage the conditional independence.



- **Sequential encoder** to leverage the context to recover the lost information
- **Autoregressive hidden Markov module** to estimate transition matrix of the unknown non-stationary.

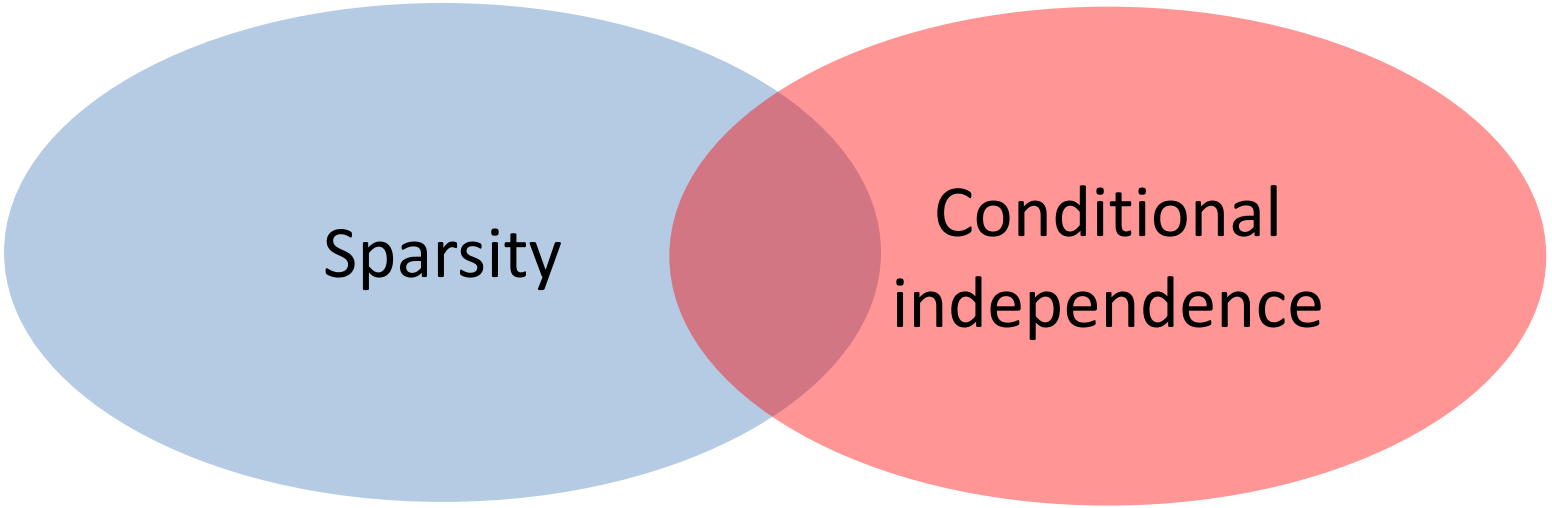
Summary of “Conditional Independence”

- In the linear case, identifiability can be achieved through non-Gaussianity.
- When the mixing function is non-linear, additional information is needed to ensure the “change” is sufficient, such as auxiliary variables like domain index.
- For stationary time series, auxiliary variables are not required; historical data can be used instead.
- In the case of non-stationary time series, the observed domain index provides sufficient change, no matter whether it involves dynamic or observational variations.
- When non-stationarity is unknown, additional assumptions, such as a Markov prior, may be necessary.
- If the mixing function is non-invertible, the lost information can be recovered from the context.
- Sparsity can help address instantaneous dependencies.



Causal Representation Learning with Sparsity

- Conditional independence requires more information on the observed distribution for sufficient change.
- Sparsity provides the constraints on the data structures.



Causal Representation Learning with Sparsity

- Sparsity provides the constraints on the data structures.

Theorem 1. *Let the observed data be sampled from a nonlinear ICA model as defined in Eqs. (1) and (2). Suppose the following assumptions hold:*

i. Mixing function $\mathbf{f}$ *is invertible and smooth. Its inverse is also smooth.*

ii. For all $i \in \{1, \dots, n\}$, *there exist* $\{\mathbf{s}^{(\ell)}\}_{\ell=1}^{|\mathcal{F}_{i,:}|}$ *and* T *s.t.* $\operatorname{span}\{\mathbf{J}_{\mathbf{f}}(\mathbf{s}^{(\ell)})_{i,:}\}_{\ell=1}^{|\mathcal{F}_{i,:}|} = \mathbb{R}^{n}_{\mathcal{F}_{i,:}}$ *and* $\left[\mathbf{J}_{\mathbf{f}}(\mathbf{s}^{(\ell)})\mathrm{T}\right]_{i,:} \in \mathbb{R}^{n}_{\hat{\mathcal{F}}_{i,:}}$.

iii. $|\hat{\mathcal{F}}| \leq |\mathcal{F}|$.

iv. *(Structural Sparsity)* *For all* $k \in \{1, \dots, n\}$, *there exists* $\mathcal{C}_k$ *such that*

$$\bigcap_{i \in \mathcal{C}_k} \mathcal{F}_{i,:} = \{k\}.$$

Then $\mathbf{h} := \hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ *is a composition of a component-wise invertible transformation and a permutation.*

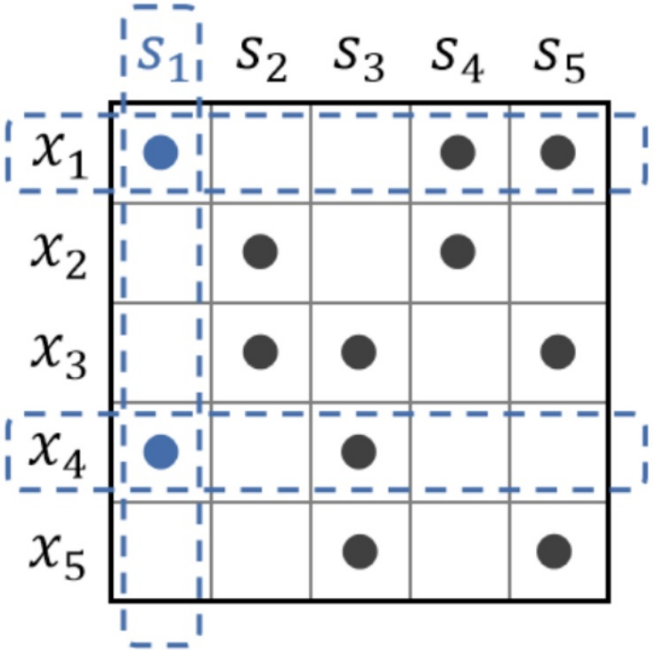
(Zheng et al., 2024)

Structural Sparsity

- Sparsity provides the constraints on the data structures.

(Structural Sparsity) *For all* $k \in \{1, \ldots, n\}$, *there exists* $\mathcal{C}_k$ *such that*

$$\bigcap_{i \in \mathcal{C}_k} \operatorname{supp}(\mathbf{J_f}(\mathbf{s})_{i,:}) = \{k\}.$$



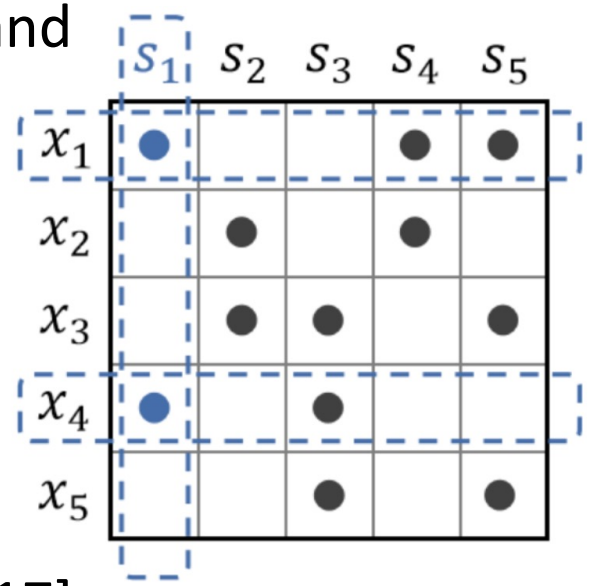
Implication: for every latent source $\boldsymbol{s_i}$, there exists a set of observed variable(s) such that $\boldsymbol{s\_i}$ is the only latent source that participates in the generation of all observed variables in the set.

Graphically, for every latent source $\boldsymbol{s_i}$, there exists a set of observed variable(s) such that the intersection of their/its parent(s) is $\boldsymbol{s_i}$

Example: for $\boldsymbol{s_1}$, there exists $\boldsymbol{x_1}$ and $\boldsymbol{x_4}$ such that the intersection of their parents is $\boldsymbol{s_1}$

Structural Sparsity

- Sparsity as a principle of simplicity
- Important in the disentanglement of latent factors both empirically and theoretically.
- In causality, various versions of **Occam's razor** are fundamental: e.g., faithfulness, the minimality principle.
- More likely to hold when the influence is "simple"



In biology, active interactions may often be sparse [Busiello et al., 2017].

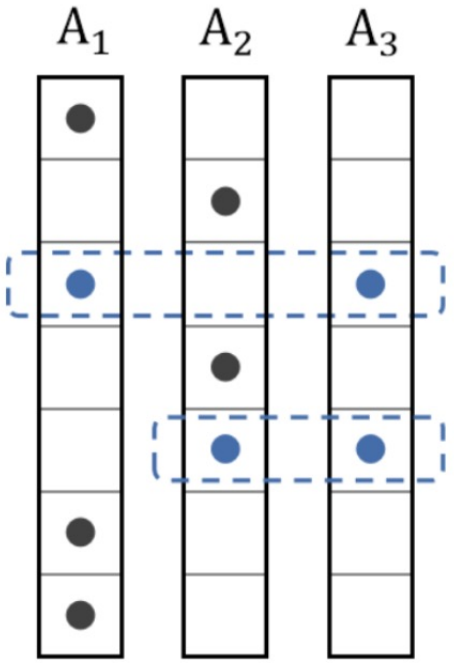
In physics, a relatively small set of laws govern complicated observed phenomena. [Einstein, 1905; Nash, 1963]

[Credit to Yujia]

Undercomplete Nonlinear ICA with Structural Sparsity

- Structural sparsity can provide identification for the undercomplete Nonlinear ICA.

Theorem 3. *Given a nonlinear ICA model defined in Eqs. (1) and (2), where* $\mathbf{f}$ *is the true mixing function. Consider* $\hat{\mathbf{f}} = \mathbf{f} \circ \mathbf{G}^{-1} \circ \mathbf{U} \circ \mathbf{G}$*, where* $\mathbf{G}$ *denotes an invertible Gaussianization*[3] *that maps the distribution to an standard isotropic (rotation-invariant) Gaussian,* $\mathbf{U}$ *denotes a rotation, and* $\mathbf{G}^{-1}$ *maps the distribution back to that before applying* $\hat{\mathbf{U}} \circ \mathbf{G}$*. If Assumptions ii, iii and iv of Thm. 2 are satisfied by replacing* $\mathbf{A}$ *with* $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ *and* $\hat{\mathbf{A}}$ *with* $\mathbf{J}_{\hat{\mathbf{f}}}(\mathbf{s})$*, then function* $\mathbf{h} := \hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ *is a composition of a component-wise invertible transformation and a permutation.*



Sparsity and Sufficiency

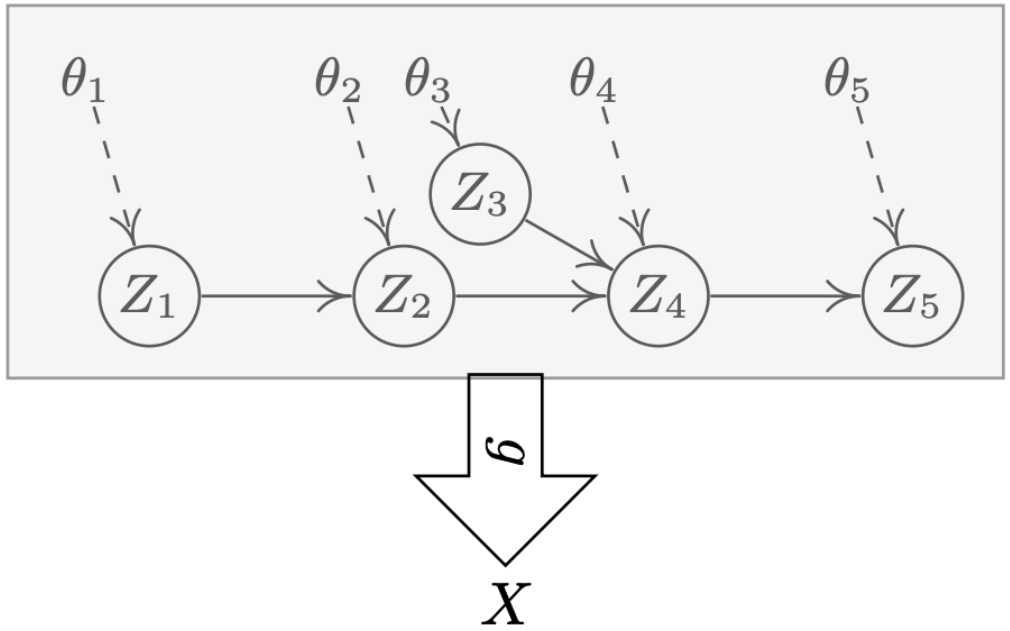
- Can Sparsity and Sufficiency work together? Yes

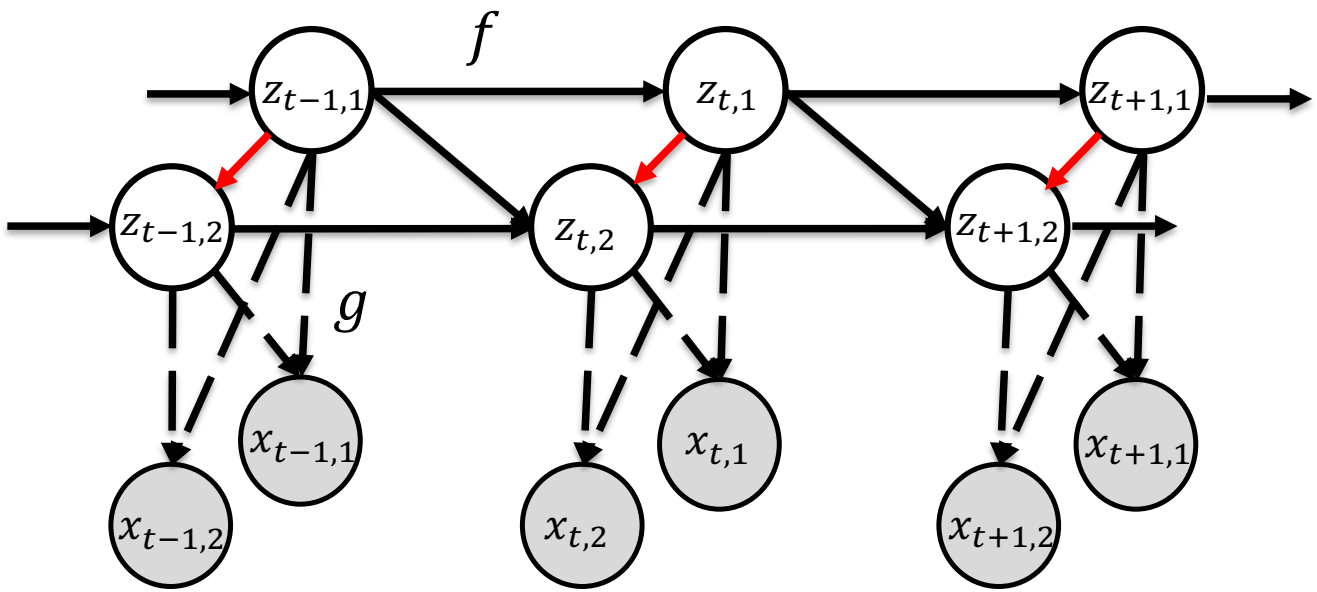
Sparsity V.S. Sufficiency

Sparsity + Sufficiency

Non-independent Sources

- Structure Sparsity and Conditional Independence both require independent sources.
- However, in complex real work, there are relations among the latent variables
- Static: images with relational latent concepts; Dynamic: low - frequency sampling





How Can We Identify from Sufficient Change

Proposition 1. *Let the observations be sampled from the data generating process in Eq. (1), and $\mathcal{M}_Z$ be the Markov network over Z. Suppose the following assumptions hold:*

- *A1 (Smooth and positive density): The probability density function of latent causal variables, i.e., p_Z, is twice continuously differentiable and positive in $\mathbb{R}^n$.*
- *A2 (Sufficient changes): For each value of Z, there exist $2n+|\mathcal{M}_Z|+1$ values of θ, i.e., $\theta^{(u)}$ with $u = 0, \ldots, 2n+|\mathcal{M}_Z|$, such that the vectors $w(Z,u) - w(z,0)$ with $u = 1, \ldots, 2n + |\mathcal{M}_Z|$ are linearly independent, where vector $w(Z,u)$ is defined as follows:*

$$w(Z,u) = \left(\frac{\partial \log p(Z;\theta^{(u)})}{\partial Z_i}\right)_{i\in[n]} \oplus \left(\frac{\partial^2 \log p(Z;\theta^{(u)})}{\partial Z_i^2}\right)_{i\in[n]} \oplus \left(\frac{\partial^2 \log p(Z;\theta^{(u)})}{\partial Z_i \partial Z_j}\right)_{\{Z_i,Z_j\}\in\mathcal{E}(\mathcal{M}_Z),\, i<j}.$$

Suppose that we learn $(\hat{g}, \hat{f}, p_{\hat{Z}}, \hat{\Theta})$ to achieve Eq. (2). Then, for every pair of estimated latent variables $\hat{Z}_k$ and $\hat{Z}_l$ that are **not adjacent in the Markov network** *$\mathcal{M}_{\hat{Z}}$ over $\hat{Z}$, we have the following statements:*

(a) For each true latent causal variable Z_i, we have

$$\frac{\partial Z_i}{\partial \hat{Z}_k}\frac{\partial Z_i}{\partial \hat{Z}_l} = 0.$$

(b) For each pair of true latent causal variables Z_i and Z_j that are adjacent in the Markov network $\mathcal{M}_Z$, we have

$$\frac{\partial Z_i}{\partial \hat{Z}_k}\frac{\partial Z_j}{\partial \hat{Z}_l} = 0.$$

(Zhang et al. 2024)

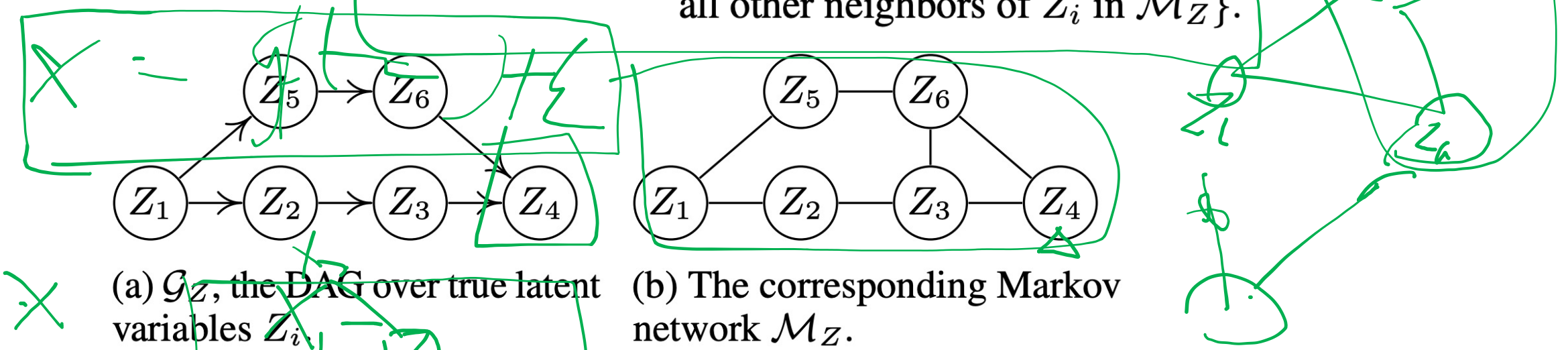
DAG V.S. Markov network

- Markov network describes the undirected relationship among variables

$$\{Z_i, Z_j\} \in E(M_Z) \;\; iff \;\; Z_i \not\perp\!\!\!\perp Z_j \mid Z_{[n]\setminus\{i,j\}}$$

Intimate neighbors

$$\Psi_{Z_i} := \{Z_j \mid Z_j, j \neq i, \text{ is adjacent to } Z_i \text{ and all other neighbors of } Z_i \text{ in } \mathcal{M}_Z\}.$$



(a) $\mathcal{G}_Z$, the DAG over true latent variables Z_i.

(b) The corresponding Markov network $\mathcal{M}_Z$.

Assumption 1 (Single adjacency-faithfulness (SAF)). *Given a DAG $\mathcal{G}_Z$ and distribution $P_{Z;\theta}$ over the variable set Z, if two variables Z_i and Z_j are adjacent in $\mathcal{G}_Z$, then $Z_i \not\perp\!\!\!\perp Z_j \mid Z_{[n]\setminus\{i,k\}}$.*

Assumption 2 (Single unshielded-collider-faithfulness (SUCF) (Ng et al., 2021)). *Given a latent causal graph $\mathcal{G}_Z$ and distribution $P_{Z;\theta}$ over the variable set Z, let $Z_i \to Z_j \leftarrow Z_k$ be any unshielded collider in $\mathcal{G}_Z$, then $Z_i \not\perp\!\!\!\perp Z_k \mid Z_{[n]\setminus\{i,k\}}$.*

How Can We Identify from Sufficient Change

- If two variables z_i and z_j are adjacent in the ground truth Markov network, then the estimated $\hat{z}_i$ and $\hat{z}_j$ are adjacent. (invertibility)
- When the estimated Markov network is with minimal edges, then $M_Z \cong M_{\hat{Z}}$
- Each latent variables is recovered as a function of itself and its intimate neighbors in the Markov network
- When the structure of latent variables is as sparse as that intimate neighbors are empty, we can establish the component-wise identification.
- A's neighbor B brings confusion only when there are no other neighbors C not adjacent to B.
- Non-linear ICA (with auxiliary variables) may be viewed as a special case

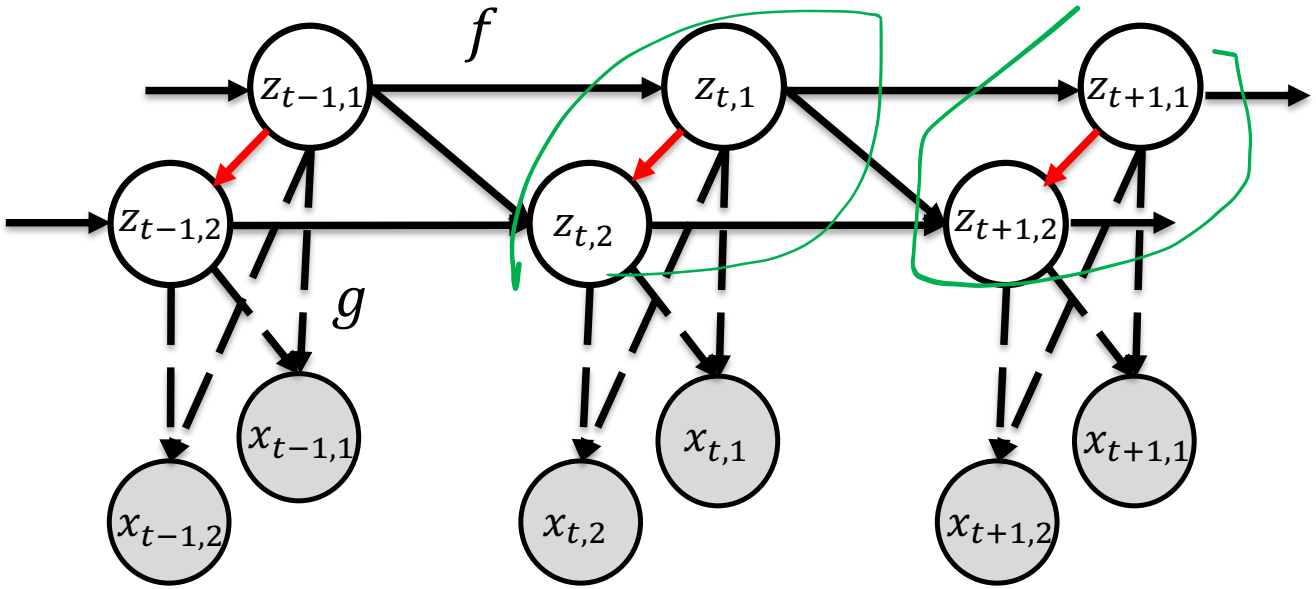
$$\frac{\partial Z_i}{\partial \hat{Z}_k}\frac{\partial Z_i}{\partial \hat{Z}_l} = 0. \qquad \frac{\partial Z_i}{\partial \hat{Z}_k}\frac{\partial Z_j}{\partial \hat{Z}_l} = 0.$$

Structure Sparsity and Sparsity Constraints

- **Sparsity Constraints:** Make sure the learned structure of latent variables is somehow equivalent to the group truth structure

- **Structure Sparsity:** When the structure of latent variables is as sparse as intimate neighbors are empty, we can establish the component-wise identification.

- **Temporal Infomation**: The historical information may provide side information to further disentangle the instantaneous intimate variables.

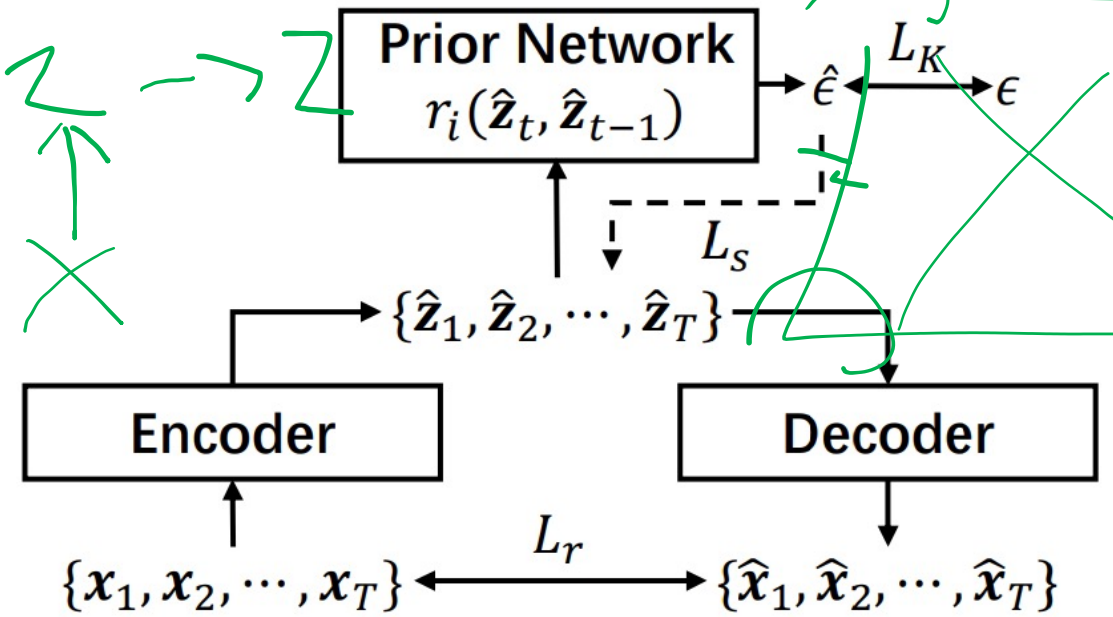


$$\frac{\partial Z_i}{\partial \hat{Z}_k}\frac{\partial Z_i}{\partial \hat{Z}_l} = 0.$$

$$\frac{\partial Z_i}{\partial \hat{Z}_k}\frac{\partial Z_j}{\partial \hat{Z}_l} = 0.$$

How to Add Constraints

- **Auto-Encoder Model**: Estimates the mixing function $x_t = g(z_t)$ and de-mixing function $\hat{z}_t = \hat{g}^{-1}(x_t)$. Invertibility is enforced by reconstruction loss.
- **Prior Network**: Estimates the prior distribution $p(z_t|z_H, z_C)$ by learning the inverse dynamics f_z^{-1}, by change of variable $p(z_t|z_H, z_C) = p_\varepsilon\left(f_z^{-1}(z_t, z_H, z_C)\right)\left|\frac{\partial f_z^{-1}}{\partial z_t}\right|$. By constraining this prior, we encourage the conditional independence.
- **Sparsity**: Adding the L1 norm on the Jacobin matrix of f_z^{-1}



$$ELBO = \underbrace{\mathbb{E}_{q(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})} \ln p(\mathbf{x}_{1:T|}|\mathbf{z}_{1:T})}_{L_r}$$
$$- \alpha \underbrace{D_{KL}(q(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})||p(\mathbf{z}_{1:T}))}_{L_K},$$

$$L_S = ||\mathbf{J}_d||_1 + ||\mathbf{J}_e||_1.,$$

Application: Video Reasoning

- Video reasoning aims to answer the neural language reasoning questions based on the video content, whose challenge lies in understanding the latent causal process.

Counterfactual Inference

Q: Would the accident still happen if there were fewer vehicles on the road?

- ✓ Yes, the road is not congested at the first place, and the accident is not related to the density of the vehicles on the road.
- ✗ No, fewer vehicles would have provided enough space to safely avoid the accident.
- ✗ No, there is no accident.

Introspection

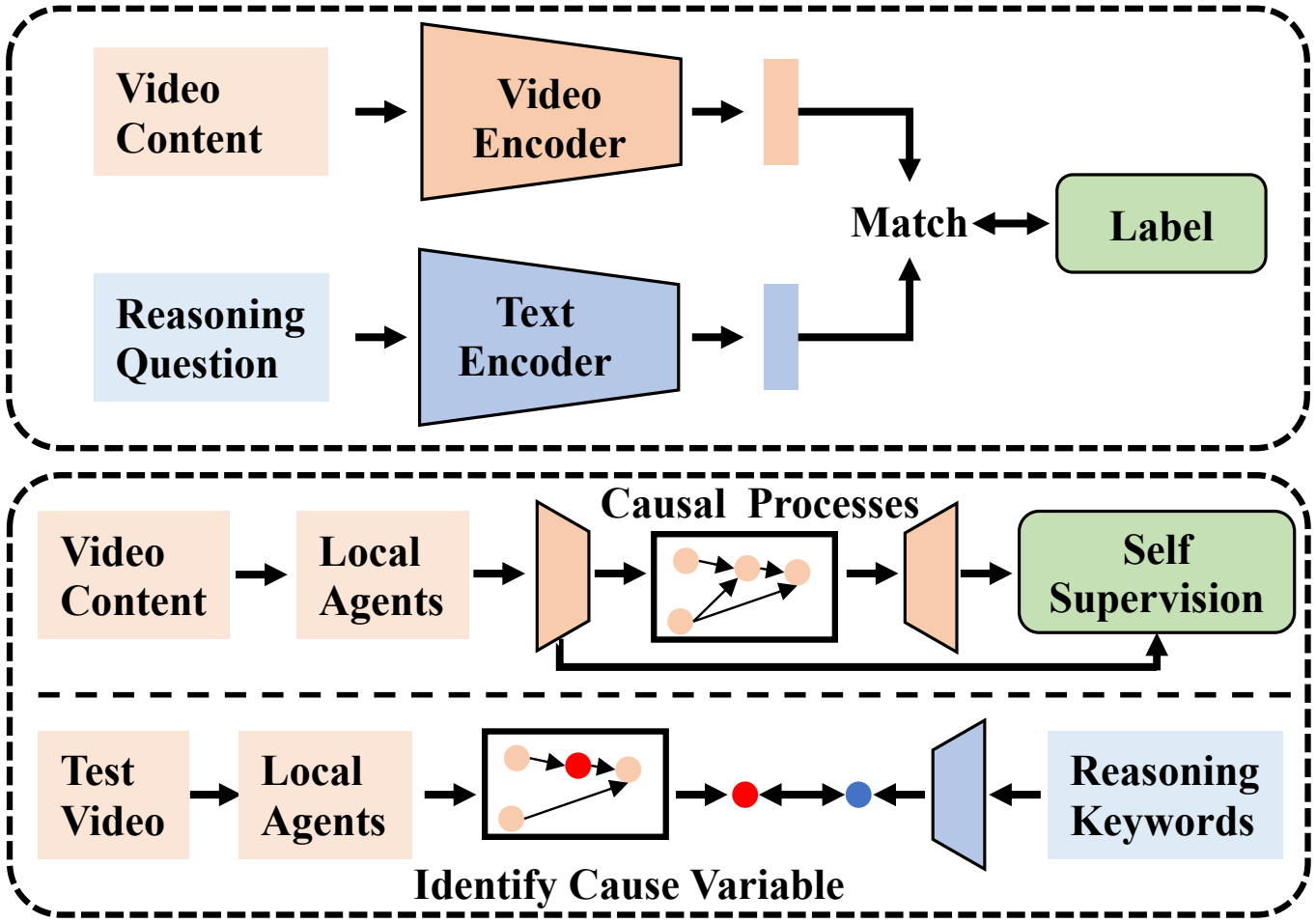
Q: What could have been done to prevent this accident from happening?

- ✗ The accident could have been avoided if the white sedan had slowed down.
- ✗ The accident could have been avoided if the black sedan had changed the lane.
- ✗ The accident could have been prevented if the road is marked clearly.
- ✓ The accident could have been avoided if the white sedan had stayed on its lane.

Credited to the SUTD-TrafficQA dataset (Xu. et.al. 2021)

Compared with Existing VideoQA methods

- Once we identify the causal dynamics, we can efficiently conduct video reasoning (such as attribution and counterfactual questions) as a causal inference process.



- Formulate VideoQA as cross-modality matching.
- Rely on question-answer (QA) pairs, Learn relations

☒

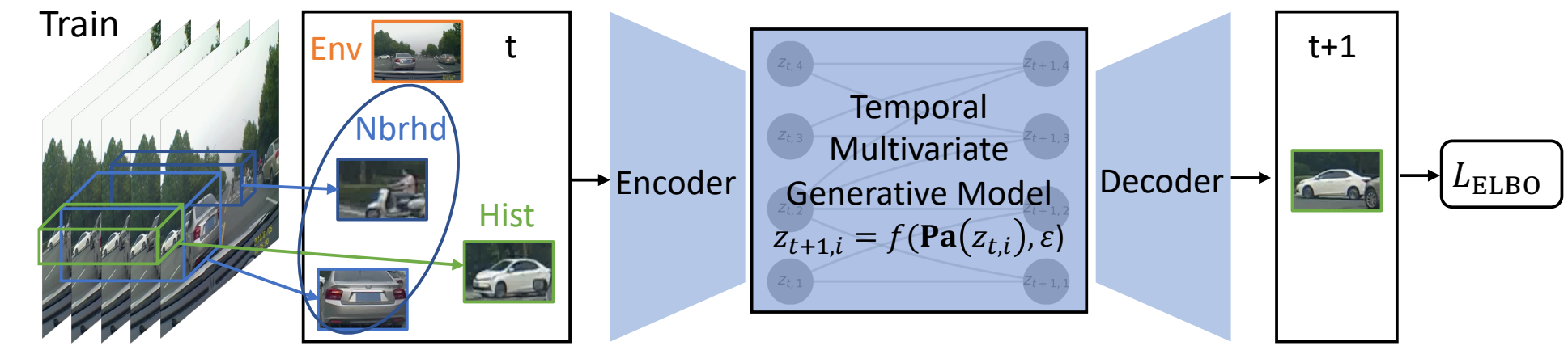
- Don't require the QA pairs.
- Can efficiently answer the attribution and counterfactual questions

☑

(Chen et al. 2024)

Overall Framework

- In the training, we learn the latent causal process from normal videos by self-supervision
- During inference, we first identify the root cause and use it to select answer



Algorithm Details

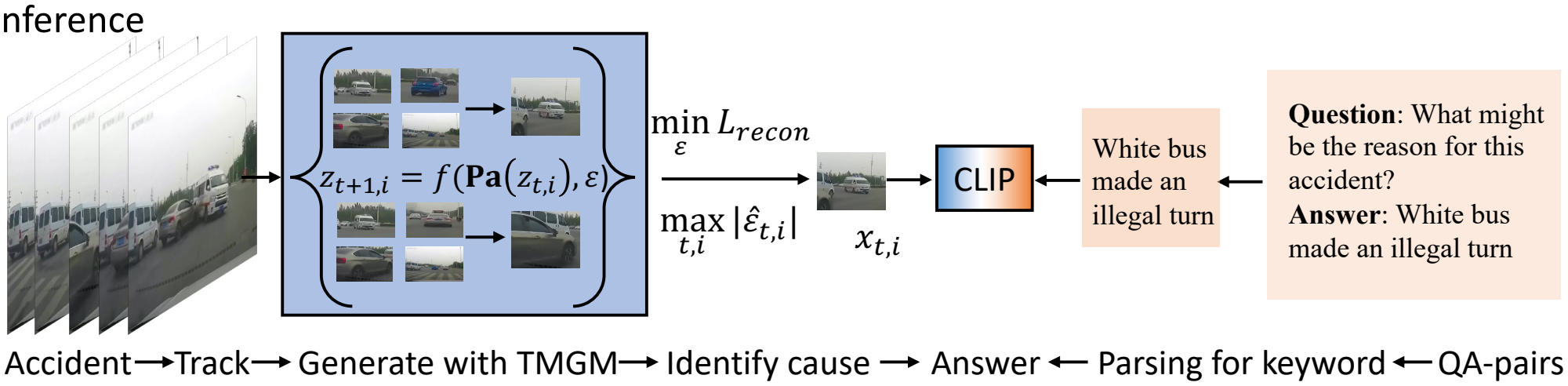
- **Identify root cause**: we find the root cause of abnormal videos by comparing them with learned causal structures. Specifically, we find the variable with a notable shift in the learned local causal processes.

$$\mathbf{x'_C} = \arg\max_{t,i} \left\| \hat{\mathbf{x}}'_{t,i} - \mathbf{x}'_{t,i} \right\|,$$

Or identify the variable that needs outlier noise for reconstruction:

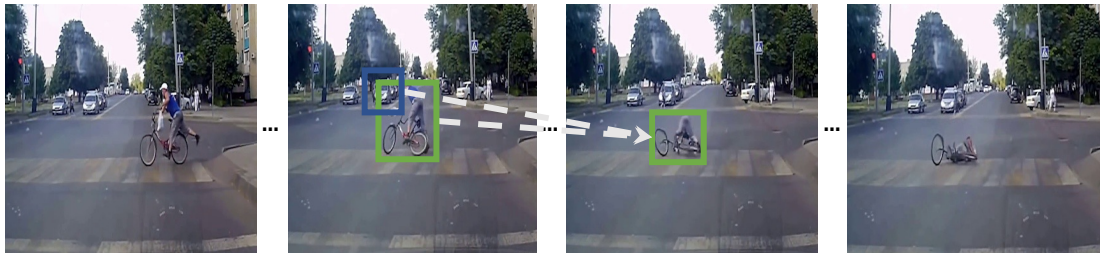
$$\mathbf{x'_C} = \arg\max_{t,i} \hat{\epsilon}_{t,i} \quad , \qquad \hat{\epsilon}_{t,i} = \min_{\epsilon_k} \left\| \hat{\mathbf{x}}'_{t,i}(\epsilon_k) - \mathbf{x}'_{t,i} \right\|$$

- **Counterfactual prediction**: 1) Estimate unique characters of the query video; 2) Change the state into the counterfactor; 3) Predict with estimated characters and counterfactual state.



Showcase of Experimental Results

➢ We can identify the causal relations and thus find the root cause.



Q: What types of vehicles that if get removed from the videos, there won't be an accident?

☑ A: Bicycle or tricycle or non-motor vehicles .

(a) Counterfactual (success)

Q: Could the accident be prevented if the involved vehicles change lane or turn properly?

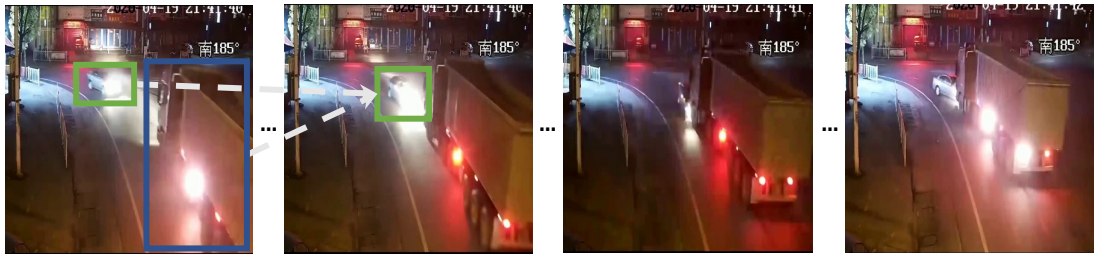
☑ A: Yes.

(b) Introspection (success)

Q: What might be the reason which led to this accident?

☑ A: The white sedan did an illegal lane changing.

(c) Attribution (success)



Q: Could any involved vehicles stop in time to prevent the accident?

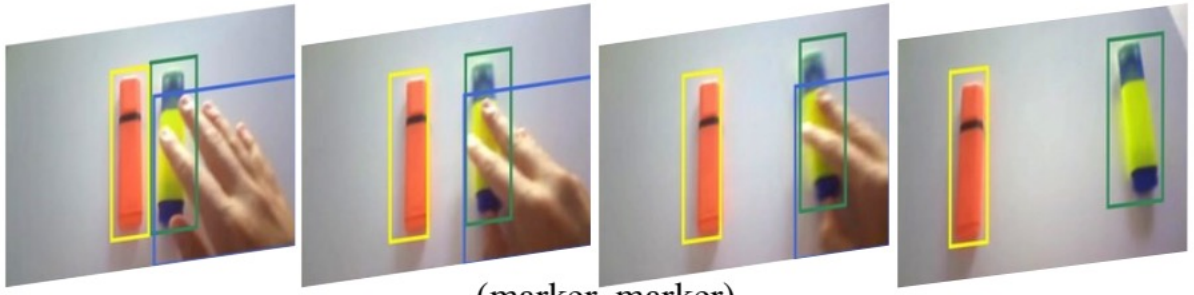
☒ A: Yes, there was enough time to react.

(d) Introspection (fail)

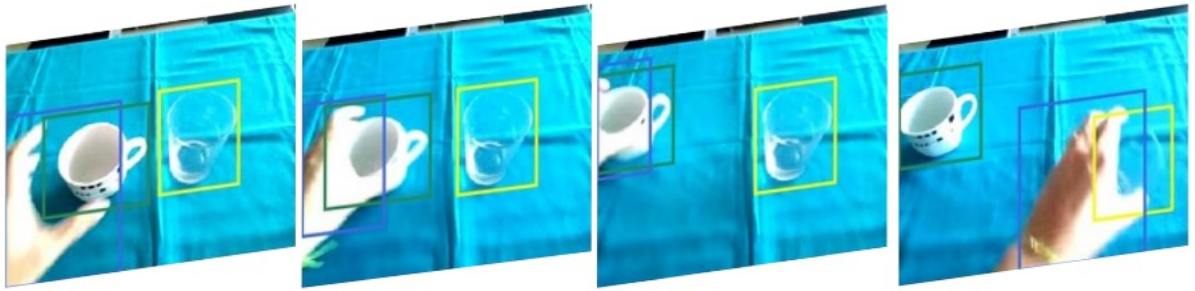
Application: Few-shot Action Recognition

- Few-shot action recognition targets to efficiently transfer the learned action recognition knowledge into the new domain with few-shot examples.

Moving [something] and [something] away from each other

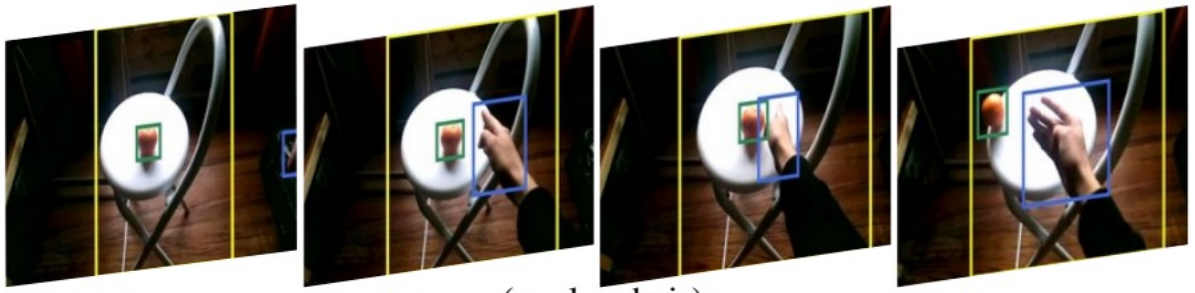


(marker, marker)

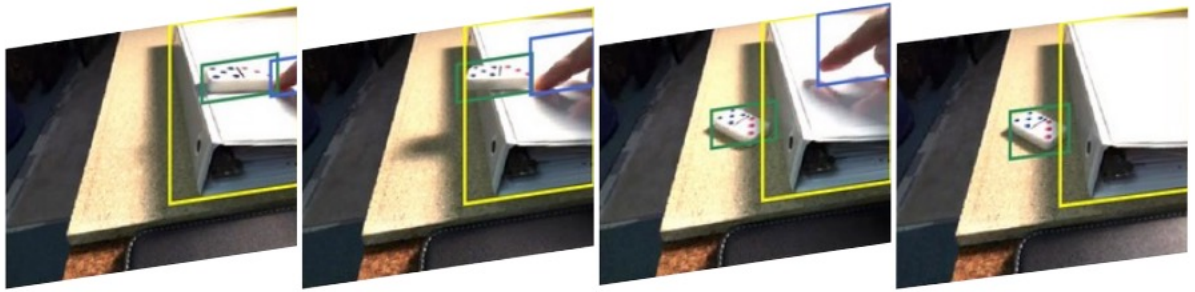


(cup, glass)

Pushing [something] off of [something]



(apple, chair)

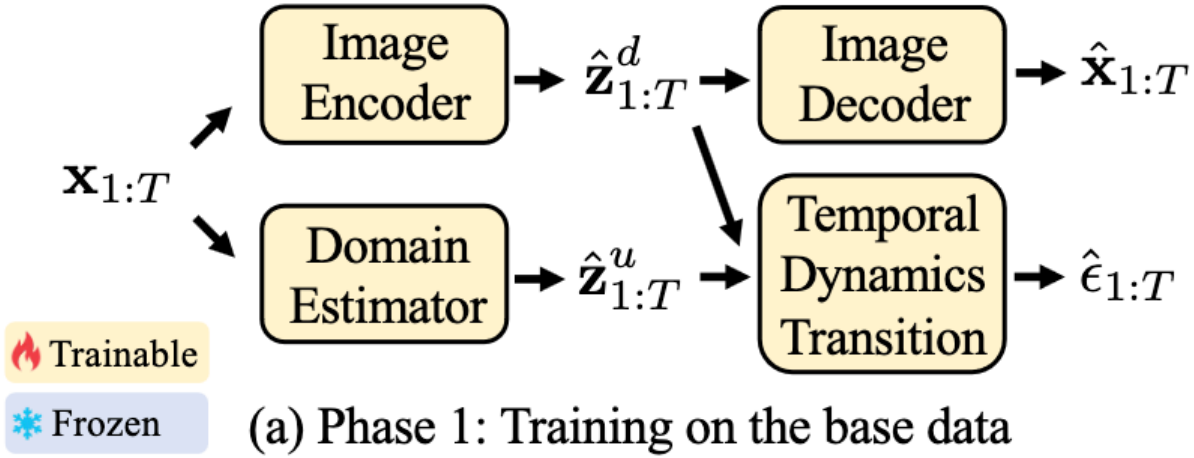


(domino, folder)

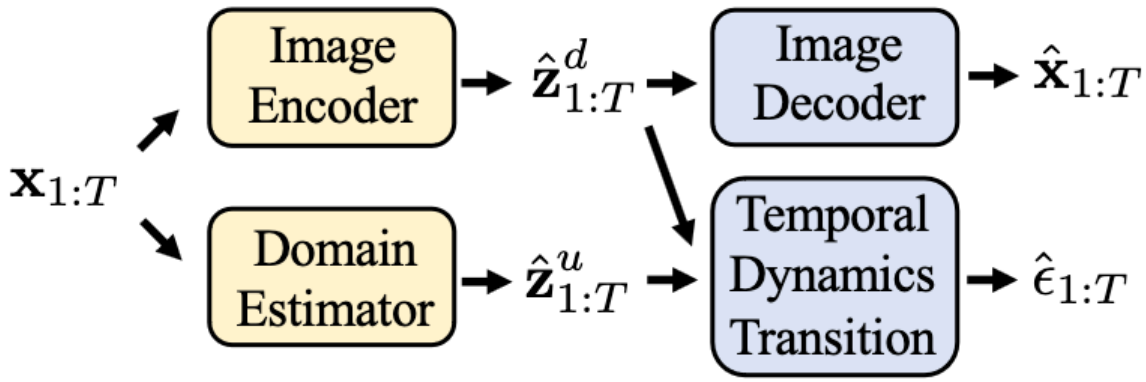
Credited to the Something-Else dataset (Materzynska. 2020)

Application: Few-shot Action Recognition

- Key insight: Despite different representations of actions, the underlying physical laws are invariant across different actions.
- Once the causal dynamic model is identified, we fix the temporal dynamic transition function (in each domain) as a constraint to adapt to novel data.



(a) Phase 1: Training on the base data



(b) Phase 2: Adaptation on the novel data

(Li & Chen et al., 2024)

Quantitative Experimental Results

➢ Significantly improve the few-shot action recognition performance over previous SOTAs

	SSv2				HMDB-51				UCF-101			
	2-shot	4-shot	8-shot	16-shot	2-shot	4-shot	8-shot	16-shot	2-shot	4-shot	8-shot	16-shot
XCLIP(Ni et al., 2022)	3.9	4.5	6.8	10.0	53.0	57.3	62.8	64.0	70.6	71.5	73.0	91.4
ActionCLIP(Wang et al., 2021b)	4.1	5.8	8.4	11.1	47.5	57.9	57.3	59.1	70.6	71.5	73.0	91.4
VicTR(Kahatapitiya et al., 2023)	4.2	6.1	7.9	10.4	60.0	63.2	66.6	70.7	87.7	92.3	93.6	95.8
VideoPrompt(Ju et al., 2022)	4.4	5.1	6.1	9.7	39.7	50.7	56.0	62.4	71.4	79.9	85.7	89.9
ViFi-CLIP(Rasheed et al., 2023)	6.2	7.4	8.5	12.4	57.2	62.7	64.5	66.8	80.7	85.1	90.0	92.7
VL Prompting(Rasheed et al., 2023)	6.7	7.9	10.2	13.5	63.0	65.1	69.6	72.0	**91.0**	93.7	95.0	96.4
VideoMAE (Tong et al., 2022)	8.2	10.0	15.1	18.2	63.7	69.4	70.9	75.3	**91.0**	94.1	94.8	97.7
CDTD$_{NCE}$ (ours)	**9.5**	**11.6**	**14.8**	**19.5**	**65.8**	**70.2**	**72.5**	**77.9**	90.6	**94.7**	**96.2**	**98.5**

➢ Also performs well in compositional action recognition task

	Loss	Sth-Else	
		5-shot	10-shot
ORViT (Herzig et al., 2022)	CE	33.3	40.2
SViT (Ben Avraham et al., 2022)		34.4	42.6
CDTD$_{CE}$ (ours)		**37.6**	**44.0**
ViFi-CLIP (Rasheed et al., 2023)	NCE	44.5	54.0
VL Prompting (Rasheed et al., 2023)		44.9	58.2
CDTD$_{NCE}$ (ours)		**48.5**	**63.9**

Takeaway Points

- Conditional independence constraint leverages the sufficient change of observed distribution to establish the identification.
- The identification results can be extended to more challenging scenarios, such as unknown non-stationary, non-invertibility, and instantaneous relations.
- The generative framework encourages conditional independence by adding the constraint between the learned posterior and conditional independent prior.
- Sparsity is another dual principle focusing on data structure
- Sparsity and Conditional independence can work together to provide more general identification results
- Learning causal representation can benefit lots of applications such as video reasoning, few-shot action recognition, and so on.

CRL Workshop @NeurIPS 2024

https://crl-community.github.io/neurips24.html

MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

Carnegie
Mellon
University

Thanks for your listening