TOWARDS IDENTIFIABLE REPRESENTATIONS UNDER MISSPECIFIED STRUCTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

The presence of noise that depends on the latent variables poses a fundamental challenge to identifiability. Existing results rely on conditional independence among the observations given the latent variables. We study a more general *misspecified structure*, where this conditional factorization does not hold, and establish both precise and approximate identifiability guarantees. We characterize structural misspecification as a perturbed factor analysis problem. For precise identifiability, we establish subspace identifiability under spectral separation and controlled perturbation, followed by component-wise identifiability under structural sparsity. When the precise condition is not guaranteed, we derive an approximate subspace-identifiability theorem. Based on these results, we develop an unsupervised variational estimator for recovering latent variables. Experiments demonstrate the effectiveness of the proposed framework.

1 INTRODUCTION

Learning meaningful representations from the data generating process with structured noise remains an open challenge. Several works treats the noise terms as known auxiliaries (Lachapelle et al., 2024b; Liang et al., 2023; Lachapelle et al., 2023; Liang et al., 2023; Lippe et al., 2023; Zheng et al., 2022; Yao et al., 2024; Lachapelle et al., 2024a; Li et al., 2025a; Song et al., 2024; Rajendran et al., 2024; Xu et al., 2024; Brady et al., 2025). One might argue that noise term can simply be absorbed into an expanded latent space. The approaches of (Kügelgen et al., 2021; Kong et al., 2022; Xie et al., 2023; Timilsina et al., 2026; Xu et al., 2026) pursue this route but require partitioning them into invariant / variant components across environments. Sun et al. (2025) assumes independence between noise and latent variables, while (Kong et al., 2023) requires that the relations between a pair of latent variables have to be sufficiently distinct.

When the noise itself depends on the latent variables, which we refer to as *dependent noise*, establishing identifiability becomes substantially more challenging. Prior work (Hu & Schennach, 2008; Zheng et al., 2025) addresses this setting under the assumption that the observations become conditionally independent given the latent variables. A few recent works extends this assumption to temporal data (Li et al., 2025b;c; Fu et al., 2026). However, this assumption can be violated in realistic systems. As illustrated in Fig. 1, in chest X-ray diagnosis, the latent variable corresponds to the patient's underlying lung-cancer state, the observation is the chest X-ray image, and the noise can be represented by the patient's inspiration level. The inspiration level may itself depend on the disease state, while both the disease state and the inspiration level affect multiple regions of the observed image. Consequently, different image regions can remain statistically dependent even after conditioning on the lung-cancer state. We refer to this setting as ***misspecified structure***, and study how to recover the latent variables under such misspecification.

We address this challenge by characterizing structural misspecification as a perturbed factor analysis problem. Equation 5 quantifies the resulting perturbation and provides the basis for two complementary identifiability regimes. First, when the perturbation is sufficiently small relative to the spectral separation, Theorem 1 establishes precise subspace identifiability, while Proposition 2 further reduces the remaining ambiguity to component-wise identifiability under structural sparsity. Second, when the sufficient condition for precise identifiability is no longer guaranteed, Theorem 2 establishes approximate subspace identifiability by bounding the identification error between the recovered representation and its admissible equivalence class. These theoretical contributions yields,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

to the best of our knowledge, one of the first general frameworks for uncovering latent variables in misspecified structure with proper identifiability guarantees.

Leveraging these theoretical insights, we propose an unsupervised method, which utilizes a variational inference-based learning objective specifically designed to uncover latent variables. The same learning objective is used in both the precise and approximate regimes. Experiments on synthetic and real-world data demonstrate improved latent recovery under a misspecified structure. In particular, we evaluate precise component-wise identification and approximate subspace identification separately, supporting the two theoretical regimes developed in this work.

2 Problem Setting

Let $\mathbf{x} \in \mathbb{R}^K$ denote the K dimensional observation, $\mathbf{z} \in \mathbb{R}^N$ denote the latent variable. Our data generating process is formulated by:

$$\mathbf{x} = g(\mathbf{z}, \epsilon), \quad \epsilon = e(\mathbf{z}, \eta) \tag{1}$$

We assume g to be nonlinear, nonparametric and smooth functions. Importantly, e denotes another nonlinear, nonparametric and smooth function, allowing the ϵ to be dependent on $\mathbf{z}$. η denotes an exogenous variable sampled from $\mathcal{N}(0, I)$.

Our primary goal is to identify latent variables $\mathbf{z}$ from observed data $\mathbf{x}$. Suppose there exists an invertible and differentiable transformation h, we further define subspace identifiability and component-wise identifiability as follows:

Definition 1. ***(Subspace Identifiability)*** *The mapping h is said to achieve subspace identifiability if it is invertible and the following transformation holds: $\hat{\mathbf{z}} = h(\mathbf{z})$, where $\hat{\mathbf{z}}$ denotes the estimation of $\mathbf{z}$.*

Definition 2. ***(Component-wise Identifiability)*** *For an individual component of the latent variable $\mathbf{z}^n$ ($\mathbf{n} \in [1, N]$), there exists a unique component $\hat{n}$ ($\hat{n} \in [1, N]$) of $\hat{\mathbf{z}}$ matches $\mathbf{z}^n$ up to a permutation π, such that $\hat{\mathbf{z}}^{\hat{n}} = h^{\hat{n}}(\mathbf{z}^{\pi(n)})$. Then $\mathbf{z}^n$ is component-wise identifiable.*

In this work, we breakdown the identifiability problem in two steps. First, we provide our findings on subspace identifiability in Section 3.1 and 3.2. Subsequently, we present the component-wise identifiability result in Section 3.3.

Remark on Misspecified Structure

Our main challenge arises from relaxing the conditional-independence structure commonly used in prior identification results (Hu & Schennach, 2008; Zheng et al., 2025; Fu et al., 2026; Li et al., 2025c;b). Following their multi-partition approach, we partition the observation as three disjoint blocks $\mathbf{x} = \{\mathbf{x_a}, \mathbf{x_b}, \mathbf{x_c}\}$ and the structural misspecification by:

Definition 3. ***(Misspecified Structure)*** *The conditional structure is misspecified when the three observed blocks do not factorize conditionally on $\mathbf{z}$, i.e., $p(\mathbf{x}|\mathbf{z}) \neq p(\mathbf{x_a}|\mathbf{z})p(\mathbf{x_b}|\mathbf{z})p(\mathbf{x_c}|\mathbf{z})$*

Figure 1 illustrates the misspecified structure using the lung-cancer example introduced in Section 1. The existing results based on conditional independence rely on a factorization of the observable distribution that no longer holds under Definition 3. Section 3 develops identifiability guarantees that remain valid under such misspecifications.

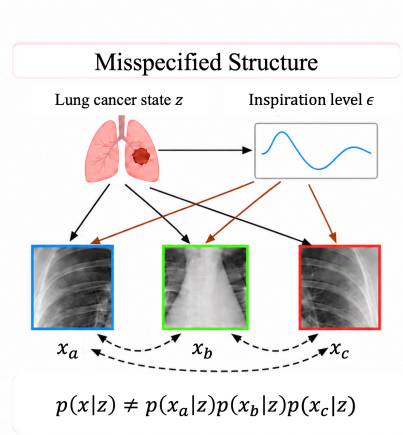


Figure 1: Visualization of the data generations of Eq. 1 under misspecified structure.

3 Identifiability Results

We first analyze the misspecified structure through factor analysis. To this end, we introduce two types of operators to characterize the structural misspecification:

Definition 4. ***(Integral Linear Operator)*** *Consider random variables $\mathbf{x}_*$ and $\mathbf{z}$ with supports $\mathcal{X}_*$ and $\mathcal{Z}$ respectively. The integral linear operator $L_{\mathbf{x}_*|\mathbf{z}}$ maps a function $f \in F(\mathcal{Z})$ to another function $L_{\mathbf{x}_*|\mathbf{z}} f \in F(\mathcal{X}_*)$, defined by:*

$$(L_{\mathbf{x}_*|\mathbf{z}} f)(\mathbf{x}_*) = \int_{\mathcal{Z}} p(\mathbf{x}_*|\mathbf{z}) f(\mathbf{z})\, d\mathcal{Z}, \quad \forall \mathbf{x}_* \in \mathcal{X}_* \tag{2}$$

where $p(\mathbf{x}_*|\mathbf{z})$ denotes the conditional density of $\mathbf{x}_*$ given $\mathbf{z}$.

Also, we present the definition of multiplication operator Λ:

Definition 5. ***(Multiplication Operator)*** *Let r and v be random variables with supports $\mathcal{R}$ and $\mathcal{V}$, respectively. For each fixed $r \in \mathcal{R}$, define the multiplication operator $\Lambda_{r|v} : (\Lambda_{r|v} f)(v) := p(r \mid v) f(v)$, where $f \in \mathcal{L}^2(\mathcal{V})$.*

With Definition 4 and 5 established, we impose the following Assumption 1 on $p(\mathbf{x}|\mathbf{z})$.

Assumption 1. *The joint distribution of $(\mathbf{x}, \mathbf{z})$ admits a bounded density with respect to a suitable product measure defined on their supports. Furthermore, all marginal and conditional densities derived from this joint distribution are also bounded.*

We begin by decomposing the bounded operator $L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}}$. The misspecified structure in Definition 3 renders the previous decomposition and identifiability results from Hu & Schennach (2008); Zheng et al. (2025) $L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}} = L_{\mathbf{x_a}|\mathbf{z}} \Lambda_{\mathbf{x_b}|\mathbf{z}} L_{\mathbf{z}|\mathbf{x_c}}$ (Section B.1) invalid. To account for this violation, we explicitly define the difference operator $\mathbf{d}(\mathbf{x_b})$ as follows:

$$\begin{aligned} \mathbf{d}(\mathbf{x_b}) &= L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}} - L_{\mathbf{x_a}|\mathbf{z}} \Lambda_{\mathbf{x_b}|\mathbf{z}} L_{\mathbf{z}|\mathbf{x_c}} \\ \mathbf{d} &= \int \mathbf{d}(\mathbf{x_b})\, d\mathbf{x_b} = L_{\mathbf{x_a}|\mathbf{x_c}} - L_{\mathbf{x_a}|\mathbf{z}} L_{\mathbf{z}|\mathbf{x_c}} \end{aligned} \tag{3}$$

where $\Lambda_{\mathbf{x_b}|\mathbf{z}}$ satisfies $\int_{\mathbf{x_b}} \Lambda_{\mathbf{x_b}|\mathbf{z}}\, d\mathbf{x_b} = I$, with I representing the identity operator. Therefore, $L_{\mathbf{x_a}|\mathbf{z}} L_{\mathbf{z}|\mathbf{x_c}} = L_{\mathbf{x_a}|\mathbf{x_c}} - \mathbf{d}$.

To the end of further decomposition, we also assume:

Assumption 2. *The operators $L_{\mathbf{x_a}|\mathbf{z}}$ and $L_{\mathbf{x_a}|\mathbf{x_c}}$ are injective and bounded-below.*

Assumption 2 implies that $L^{-1}_{\mathbf{x_a}|\mathbf{z}}$ and $L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}$ are well-defined and bounded. Leveraging all previous assumptions, we are able to derive the following:

$$L_{\mathbf{x_a}|\mathbf{z}} \Lambda_{\mathbf{x_b}|\mathbf{z}} L^{-1}_{\mathbf{x_a}|\mathbf{z}} = (L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}} - \mathbf{d}(\mathbf{x_b}))(L_{\mathbf{x_a}|\mathbf{x_c}} - \mathbf{d})^{-1} = L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}} L^{-1}_{\mathbf{x_a}|\mathbf{x_c}} + Per \tag{4}$$

where Per denotes the perturbation operator that represents the structural misspecification. Rather than requiring Per itself to be known, we characterize its upper bound through $||Per||_{op} \leq \overline{Per}$, which is derived in Lemma 1 in Section B.4. Consequently, our aim becomes to identify $\mathbf{z}$ from $L_{\mathbf{x_a}|\mathbf{z}} \Lambda_{\mathbf{x_b}|\mathbf{z}} L^{-1}_{\mathbf{x_a}|\mathbf{z}}$ of Eq. 4.

It is worth noting that Zheng et al. (2025) particularly proves the subspace identifiability for $Per = 0$ when the conditional independence holds. However, Definition 3 indicates that $Per \neq 0$, and (Zheng et al., 2025) can be seen as a special case of this work.

Before going further, we must address two critical challenges of $L_{\mathbf{x_a}|\mathbf{z}} \Lambda_{\mathbf{x_b}|\mathbf{z}} L^{-1}_{\mathbf{x_a}|\mathbf{z}}$: scaling ambiguity and spectral value degeneracy. The scaling ambiguity arises when there exists a non-zero constant c such that $\int cp(\mathbf{x_a}|\mathbf{z})\, d\mathbf{x_a} > 1$. However, this ambiguity are solved since the probability density constraint $\int p(\mathbf{x_a}|\mathbf{z})\, d\mathbf{x_a} = 1$ must always hold. The degeneracy occurs if spectral value are repeated, thereby associating multiple eigenfunctions to a single spectral value. To eliminate this degeneracy, we introduce the following additional assumption:

Assumption 3. *The operator $L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}} L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}$ has distinct spectral values with cardinality equal to that of $\Lambda_{\mathbf{x_b}|\mathbf{z}}$. Also, The map $\Theta : \mathcal{Z} \to \mathcal{D}, s.t. \Theta(\mathbf{z}) = p(\mathbf{x_b} \mid \mathbf{z})$ is injective.*

Under misspecification, Eq. 4 shows that the $L_{\mathbf{x_a}|\mathbf{z}} \Lambda_{\mathbf{x_b}|\mathbf{z}} L^{-1}_{\mathbf{x_a}|\mathbf{z}}$and $L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}} L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}$ differ by Per. We therefore quantify how far the spectral values of $L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}} L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}$ may move under this perturbation. To this end, we need to first justify the spectral analysis of Eq. 4 by establishing

that $L_{\mathbf{x_a}|\mathbf{z}}\Lambda_{\mathbf{x_b}|\mathbf{z}}L^{-1}_{\mathbf{x_a}|\mathbf{z}}$ admits a unique spectral resolution. Since $p(\mathbf{x_b}|\mathbf{z})$ is real-valued, $\Lambda_{\mathbf{x_b}|\mathbf{z}}$ is self-adjoint, which already admits a unique projection-valued measure. Its similarity transformation through $L_{\mathbf{x_a}|\mathbf{z}}$ then induces the corresponding unique spectral resolution of $L_{\mathbf{x_a}|\mathbf{z}}\Lambda_{\mathbf{x_b}|\mathbf{z}}L^{-1}_{\mathbf{x_a}|\mathbf{z}}$ (Dunford & Schwartz, 1971). Thus, $L_{\mathbf{x_a}|\mathbf{z}}\Lambda_{\mathbf{x_b}|\mathbf{z}}L^{-1}_{\mathbf{x_a}|\mathbf{z}}$ is equivalently the spectral resolution of $L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}}L^{-1}_{\mathbf{x_a}|\mathbf{x_c}} + Per$. The detailed proof is provided in Appendix B.2.

Let $\rho \in \sigma(L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}}L^{-1}_{\mathbf{x_a}|\mathbf{x_c}})$ and $\rho_\Lambda \in \sigma(L_{\mathbf{x_a}|\mathbf{z}}\Lambda_{\mathbf{x_b}|\mathbf{z}}L^{-1}_{\mathbf{x_a}|\mathbf{z}})$, where $\sigma(*)$ denotes the spectral value. The resolvent perturbation bound for bounded operator from (Trefethen, 2005; Amor, 2003) yields:

$$dist(\rho, \rho_\Lambda) \leq \kappa\overline{Per} \tag{5}$$

where $dist(\rho, \rho_\Lambda) = \inf_{\rho_\Lambda} |\rho - \rho_\Lambda|$, $\kappa = ||L_{\mathbf{x_a}|\mathbf{z}}||_{op}||L^{-1}_{\mathbf{x_a}|\mathbf{z}}||_{op}$, and $||*||_{op}$ denotes the operator norm. Notably, $\kappa \geq 1$, since$||L_{\mathbf{x_a}|\mathbf{z}}L^{-1}_{\mathbf{x_a}|\mathbf{z}}||_{op} = ||I||_{op} \leq ||L_{\mathbf{x_a}|\mathbf{z}}||_{op}||L^{-1}_{\mathbf{x_a}|\mathbf{z}}||_{op}$, where I denotes the identity operator. Let us define $\eta = \inf_{i\neq j} \frac{|\varrho^i - \varrho^j|}{2\kappa} - \alpha \geq 0$ for some constant $\alpha > 0$, where ϱ denotes the non-empty spectrum, i, j denote the indices of the spectrum. Similarly, let ϱ_Λ denotes the spectrum of ρ_Λ. We opt to measure $dist(\rho, \rho_\Lambda) = |\rho^i - \rho^i_\Lambda|$. Equation 5 forms the basis of the precise and approximate identifiability results developed next.

Assumption Intuitions
• Assumption 1 is a regularity condition to ensure that the operator expressions below are well defined.
• Assumption 2 is mild and prevents these operators from collapsing distinct functions to the same output.
• Assumption 3 prevents different latent variables from being indistinguishable through the same spectral value. Injectivity of Θ further ensures that distinct latent states induce distinct conditional distributions of $\mathbf{x_b}$.

3.1 Precise subspace Identifiability

After facilitating the structural misspecification in Eq. 5, we now establish the precise identifiability result that meets Definition 1. The key requirement is that $\overline{Per}$ needs to be sufficiently small so that distinct spectral values remain distinguishable.

Consider $||Per||_{op} \leq \overline{Per} < \eta$ [1] for some non-zero constant η, and substitute it into Eq. 4 resulting in:

$$L_{\mathbf{x_a}|\mathbf{z}}\Lambda_{\mathbf{x_b}|\mathbf{z}}L^{-1}_{\mathbf{x_a}|\mathbf{z}} < L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}}L^{-1}_{\mathbf{x_a}|\mathbf{x_c}} + \eta \tag{6}$$

Given Eq. 6, we propose the following for the distinctness of ρ^i_Λ:

Proposition 1. *Suppose Assumptions 1 $\sim$ 3 hold, and if both ϱ^i and ϱ^i_Λ include only ρ^i and ρ^i_Λ, respectively. Then, for all $i > 1$ and for all* $\mathbf{z}$*,*

$$\rho^i_\Lambda \in \left(\max\{\rho^i - \kappa\overline{Per}, \rho^{i-1} + \kappa\overline{Per}\}, \rho^i + \kappa\overline{Per}\right)$$

and if $i = 1$, for all $\mathbf{z}$*,*

$$\rho^i_\Lambda \in (\rho^i - \kappa\overline{Per}, \rho^i + \kappa\overline{Per})$$

guarantee the distinctness of ρ_Λ.

Insight: For fixed $\mathbf{x_b}$, $\sigma(\Lambda_{\mathbf{x_b}|\mathbf{z}})$ is determined by the essential range of $p(\mathbf{x_b}|\mathbf{z})$. Proposition 1 ensures that $\overline{Per}$ does not collapse two distinct ρ_Λ into the same regime. Together with the injectivity of Θ in Assumption 3, ρ_Λ remains invariant by re-labeling through permuting $\mathbf{z}$. This can be formally described by $\mathbf{z}' = h'(\mathbf{z})$, where h' denotes a bijection $h' : \mathcal{Z} \to \mathcal{Z}$, and $\mathbf{z}'$ is a permuted version of $\mathbf{z}$.

Proof Sketch: We proceed by contradiction. Assuming the existence of distinct spectral values ρ^1, ρ^2 satisfy $|\rho^1 - \rho^i_\Lambda| \leq \kappa\overline{Per}$ and $|\rho^2 - \rho^i_\Lambda| \leq \kappa\overline{Per}$. This leads directly to $\overline{Per} \geq \eta$, contradicting the condition $\overline{Per} < \eta$.

[1] We provide the sufficient condition of $\overline{Per} < \eta$ in Lemma 2 in Section B.4.

Notably, the bijection h' introduced above might not coincide with h defined in Definition 1. To meet the requirement of h being differentiable, we introduce the following assumption:

Assumption 4. *There exists a map $M : \mathcal{D} \to \mathcal{Z}$ such that $M(p(\mathbf{x_b}|\mathbf{z})) = M(p(\mathbf{x_b}|h'(\mathbf{z}))) = h(\mathbf{z})$, where $\mathcal{D}$ is the support of $p(\mathbf{x_b}|\mathbf{z})$, h is a differentiable transformation.*

We are now ready to state our main theoretical result on subspace identifiability:

Theorem 1. *Consider observed variables $\mathbf{x} \in \mathbb{R}^K$ and the estimated latent variables $\hat{\mathbf{z}} \in \mathbb{R}^N$, suppose that Proposition 1 and Assumption 4 hold, $\exists h, s.t., \hat{\mathbf{z}} = h(\mathbf{z}) = h'(\mathbf{z})$, In other words, $\mathbf{z}$ must be subspace identified.*

Insight: Theorem 1 characterizes a perturbation regime in which the subspace identifiability in Definition 1 survives from structural misspecification.

The detailed proofs of both Proposition 1 and Theorem 1 are in Sec.B.2.

Assumption intuition:
Assumption 4 ensures that the transformation h established through the Proposition 1 belongs to the transformation class required by Definition 1.

3.2 Approximate Subspace Identifiability

Theorem 1 establishes precise identifiability under the perturbation condition $\overline{Per} < \eta$. However, this inequality is not guaranteed when the sufficient condition in Lemma 2 is not satisfied. Consequently, Theorem 1 may no longer hold even under Assumptions 1-3. Motivated by this limitation, we consider *approximate identifiability*, which quantifies the deviation of the recovered representation from its admissible equivalence class:

Definition 6. ***(Approximate identifiability)*** *Let $\mathcal{H}$ denote the class of admissible invertible and smooth transformations in Definition 1. The latent variable $\mathbf{z}$ is ε-approximately identifiable with respect to $\mathcal{H}$ if*

$$d_{\mathcal{H}}(\hat{\mathbf{z}}, \mathbf{z}) := \inf_{h \in \mathcal{H}} |\hat{\mathbf{z}} - h(\mathbf{z})| = \varepsilon \tag{7}$$

where the identifiability error $\varepsilon \geq 0$. In particular, $\varepsilon = 0$ reduces to precisely subspace identifiability.

We are now ready to connect ε with $\overline{Per}$. To this end, we impose the following assumption and Theorem:

Assumption 5. *The latent space $\mathcal{Z}$ is compact, and the conditional-distribution map Θ in Assumption iii is continuously differentiable. Moreover, there exists a constant $c_\Theta > 0$ such that, for an admissible invertible transformation $h \in \mathcal{H}$, Θ satisfies the lower bound of Bi-Lipschitz continuity by:*

$$c_\Theta |\hat{\mathbf{z}} - h(\mathbf{z})| \leq \|\Theta(\mathbf{z}) - \widehat{\Theta}(\hat{\mathbf{z}})\|_{\mathcal{F}}. \tag{8}$$

Theorem 2. *Suppose Assumptions 1–3 and Assumption 5 hold, and the spectral perturbation bounds in Eq. 5 apply to the corresponding spectral values of the true and estimated representations. Then the approximate identification error in Definition 6 satisfies*

$$\varepsilon \leq \frac{\kappa \overline{Per} + \hat{\kappa} \widehat{\overline{Per}}}{c_\Theta} \tag{9}$$

Insight: Theorems 1 and 2 characterize two identifiability regimes. When $\overline{Per} < \eta$, the spectral value remain separated and precise subspace identifiability is guaranteed. When this condition cannot be guaranteed, Theorem 2 provides a relaxation: there exists an error ε bounded by Eq. 9.

Under Assumptions 1–3, the same spectral perturbation argument as in Eq. 5 yields $|\rho^i - \rho^i_\Lambda| \leq \kappa \overline{Per}$ and $|\rho^i - \hat{\rho}^i_\Lambda| \leq \hat{\kappa} \widehat{\overline{Per}}$. Therefore, the triangle inequality results in $|\rho^i_\Lambda - \hat{\rho}^i_\Lambda| \leq \kappa \overline{Per} + \hat{\kappa} \widehat{\overline{Per}}$. Combining this perturbation bound with the inverse-Lipschitz condition in Assumption 5 yields the upper bound in Eq. 9. The detailed proof is provided in Sec. B.3.

Assumption intuition:
Assumption 5 provides the quantitative measure to connect the discrepancy in $\mathcal{D}$ with discrepancy in $\mathcal{Z}$.

3.3 Precise Component-wise Identifiability

In this section, we introduce our precise component-wise identifiability upon Theorem 1. *Extending the result to approximate component-wise identifiability requires additional component-level information and is left for future work.*

Theorem 1 guarantees that $\hat{\mathbf{z}} = h(\mathbf{z})$ is not a function of ϵ, hence $\mathbf{z}$ and ϵ are disentangled. We now focus on identifying $\mathbf{z}$ in a component-wise manner. Consider that each dimensions of $\mathbf{z}$ are independent of each other ($p(\mathbf{z}) = \prod_{n=1}^{N} p(\mathbf{z}^n)$), Taking inspirations from (Lachapelle et al., 2023; Zheng & Zhang, 2024; Xu et al., 2024), we extend the definition of structural sparsity to the function from $\mathbf{z}$ to $p(\mathbf{x}|\mathbf{z})$ and establish the following proposition:

Proposition 2. *Suppose (Theorem 1) holds, and the following assumptions and regularization conditions are satisfied:*

1. **(Independence)***: For $n \in [N]$, $\mathbf{z}^n$ are independent of each other: $p(\mathbf{z}) = \prod_{n=1}^{N} p(\mathbf{z}^n)$.*
2. **(Structural sparsity)***: The map $\Phi : \mathcal{Z} \to \mathcal{P}$ is smooth, where $\mathcal{Z}$ denotes the support of $\mathbf{z}$, and $\mathcal{P}$ denotes the support of $p(\mathbf{x}|\mathbf{z})$ ($\mathcal{P} \subset \mathcal{L}^2(\mathcal{X}, \mu)$ ($\mathcal{X}$ denotes the space of $\mathbf{x}$, μ denotes the measure)). Let G and $\hat{G}$ be the binary support matrix of Jacobian $J_\Phi(\mathbf{z})$ and $J_{\hat{\Phi}}(\hat{\mathbf{z}})$, respectively, i.e., $G^{i,j} = 1$ if $||\frac{\partial p(\mathbf{x}^i|\mathbf{z})}{\partial \mathbf{z}^j}||_{\mathcal{L}^2} \neq 0$ and 0 for otherwise. For each $n \in [1, N]$, there exists a subset of indices $\mathcal{C}_k$ satisfying $\bigcap_{m \in \mathcal{C}_k} G^{m,:} = \{n\}$;*
3. *$|\hat{G}|_0 \leq |G|_0$, where $|*|_0$ is the ℓ_0 norm.*

Then, $\hat{\mathbf{z}}$ must correspond component-wise to a permutation of the true latent variables $\mathbf{z}$.

Insight: Theorem 1 first reduces the ambiguity from an arbitrary latent representation to an invertible transformation $\hat{\mathbf{z}} = h(\mathbf{z})$. Proposition 2 further uses structural sparsity to rule out transformations that mix multiple latent coordinates. Consequently, even under the misspecified structure, we are able to recover $\mathbf{z}$ in a component-wise manner.

Proof Sketch: The complete proof is deferred to the Appendix B.5. Here we highlight key steps. First, the subspace identifiability (Theorem 1) imply $\hat{\mathbf{z}} = h(\mathbf{z})$, which leads to: $J_\Phi(\mathbf{z}) = J_{\hat{\Phi}}(\hat{\mathbf{z}}) J_h(\mathbf{z})$. Subsequently, by leveraging the injectivity of Φ in Assumption 3, component-wise identifiability is proven by contradiction: any violation would contradict the structural sparsity in assumption 2.

Assumption Intuitions:
• Assumption 1 treats the coordinates of $\mathbf{z}$ as distinct latent sources. It provides the additional structure on which the sparsity condition operates.
• Assumption 2 requires each latent component to be sparsely uniquely localized through the conditional distributions of a subset of observed variables.
• $|\hat{G}_0| \leq |G|$ is the regularization that supports the structural sparsity in Assumption 2 for learning.

Assumption 1 in Proposition 2 claims the independence between each dimension of $\mathbf{z}$. To relax such a constraint, we can alternatively allow latent variables $\mathbf{z}$ to exhibit dependence via a known auxiliary domain variable $\mathbf{u}$. The details are discussed in Proposition 3 of Sec. B.6.

4 Approach

We now develop an unsupervised estimator for recovering the latent variable $\mathbf{z}$ under the data-generating process in Eq. 1. Our central objective aims to learn $p(\mathbf{x}|\mathbf{z})$. However, since $\mathbf{z}$ is unobservable from the data, we cannot directly optimize $p(\mathbf{x}|\mathbf{z})$ from the data. Accordingly, we instead learn the joint density in Eq. 1 through formalizing the following:

$$p(\mathbf{z}, \epsilon, \mathbf{x}) = p_\theta(\mathbf{x}|\mathbf{z}, \epsilon) p_\gamma(\epsilon|\mathbf{z}) p_\delta(\mathbf{z}) \tag{10}$$

where parameters θ denotes the parameters of g. γ denotes the parameters of e, and δ parameterizes $p(\mathbf{z})$. To uncover the latent variables $\mathbf{z}$ and ϵ from observed data $\mathbf{x}$, we introduce two encoders, $q_\psi(\mathbf{z}|\mathbf{x})$ and $q_\phi(\epsilon|\mathbf{x})$ parameterized by ψ and ϕ, respectively.

To the end of learning Eq. 10, we build our approach upon the framework of Beta-VAE (Higgins et al., 2016). The overall architecture of our framework is illustrated in Figure 2. In what follows, we detail each part of our proposed model.

4.1 Network Design

Learning Eq. 10 via variational inference suggests the architecture for our approach composing of the following key elements. Specifically, the architecture includes two encoders: $q_\psi(\hat{\mathbf{z}}|\mathbf{x})$ for inferring latent variables $\mathbf{z}$, and $q_\phi(\hat{\epsilon}|\mathbf{x})$ for estimating the posterior of noise term ϵ. These latent representations are then utilized by a decoder $p_\theta(\hat{\mathbf{x}}|\hat{\mathbf{z}}, \hat{\epsilon})$ to reconstruct the observations $\mathbf{x}$. Additionally, we regularize the latent variables by constraining their posterior distributions via the KL divergence to match the learned priors. We detail each of these modules below.

Encoder $q_\psi(\hat{\mathbf{z}}|\mathbf{x})$: We parameterize $q_\psi(\hat{\mathbf{z}}|\mathbf{x})$ as an isotropic Gaussian characterized by mean $\mu_\mathbf{z}$ and covariance $\sigma_\mathbf{z}$. To approximate this posterior, we employ a neural network encoder constructed with an MLP followed by a leaky ReLU activation:

$$\hat{\mathbf{z}} \sim \mathcal{N}(\mu_\mathbf{z}, \sigma_\mathbf{z}), \quad \mu_\mathbf{z}, \sigma_\mathbf{z} = \text{LeakyReLU}(\text{MLP}(\mathbf{x})) \tag{11}$$

Encoder $q_\phi(\hat{\epsilon}|\mathbf{x})$: Similarly, we parameterize $q_\phi(\hat{\epsilon}|\mathbf{x})$ as another isotropic Gaussian distribution:

$$\hat{\epsilon} \sim \mathcal{N}(\mu_\epsilon, \sigma_\epsilon), \quad \mu_\epsilon, \sigma_\epsilon = \text{LeakyReLU}(\text{MLP}(\mathbf{x})) \tag{12}$$

Prior Estimation $p_\delta(\mathbf{z}|)$: We estimate the prior $p_\delta(\mathbf{z}|)$ as a factorized Gaussian across latent dimensions, since we assume the independence of each dimension of $\mathbf{z}$ in Assumption 1 of Proposition 2:

$$p_\delta(\mathbf{z}) = \prod_{n=1}^{N} p_\delta(\mathbf{z}^n), \quad \mathbf{z}^n \sim \mathcal{N}(0, 1) \tag{13}$$

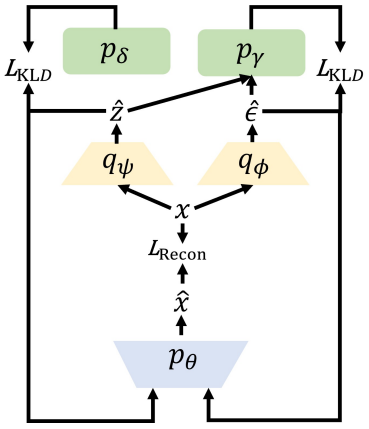


Figure 2: The overall framework of our proposed approach consists of: (1) two encoders q_ψ and q_ϕ that map observations $\mathbf{x}_t$ to $\hat{\mathbf{z}}$ and $\hat{\epsilon}$, respectively; (2) a decoder that reconstructs observations $\hat{\mathbf{x}}$ from $\hat{\mathbf{z}}$ and $\hat{\epsilon}$; and (3) two prior estimation modules p_δ and p_γ that models the prior of $\mathbf{z}$ and ϵ, respectively. We train the framework by L_{Recon} along with L_{KLD}.

Prior Estimation $p_\gamma(\epsilon|\mathbf{z})$: Direct estimation of the arbitrary density $p_\gamma(\epsilon|\mathbf{z})$ poses a substantial challenge. To overcome this, we introduce a transformation-based module leveraging normalizing flows, representing the prior distribution as a Gaussian transformed via an invertible mapping. Suppose each component of ϵ is independent conditioning on $\mathbf{z}$, $\forall m \in [1, M]$, the prior model is formulated through: $\hat{\eta}^m = \hat{e}^{-1,m}(\hat{\epsilon}^m|\hat{\mathbf{z}})$. Using the change-of-variable, the prior distribution of $\hat{\epsilon}^m$ is computed as: $p_\gamma(\hat{\epsilon}^m|\hat{\mathbf{z}}) = p(\hat{\eta}^m)\left|\frac{\partial \hat{e}^{-1,m}}{\partial \hat{\epsilon}^m}\right| = p_\gamma(\hat{e}^{-1,m}(\hat{\epsilon}^m|\hat{\mathbf{z}}))\left|\frac{\partial \hat{e}^{-1,m}}{\partial \hat{\epsilon}^m}\right|$. Aggregating across all dimensions, the complete prior distribution is given by:

$$p_\gamma(\hat{\epsilon}|\hat{\mathbf{z}}) = \prod_{m=1}^{M} p(\hat{\eta}^m)\left|\frac{\partial \hat{e}^{-1,m}}{\partial \hat{\epsilon}^m}\right| \tag{14}$$

The normalizing flow transformation $\hat{e}$ is implemented using a stacked MLP.

Decoder $p_\theta(\hat{\mathbf{x}}|\hat{\mathbf{z}}, \hat{\epsilon})$: The decoder generates the reconstructed observations $\hat{\mathbf{x}}$ from inferred latent variables $\hat{\mathbf{z}}$ and $\hat{\epsilon}$. It is implemented using an MLP followed by leaky ReLU activations:

$$\hat{\mathbf{x}} = \text{LeakyReLU}(\text{MLP}(\hat{\mathbf{z}}, \hat{\epsilon})) \tag{15}$$

4.2 TRAINING OBJECTIVE

Our complete ELBO is:

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{\hat{\mathbf{z}}\sim q_\psi}[\mathcal{L}_{\text{cond}}] - \beta_1 \underbrace{\mathbb{E}_{\hat{\mathbf{z}}\sim q_\psi}\big(\log q\left(\hat{\mathbf{z}}|\mathbf{x}\right) - \log p_\delta(\mathbf{z})\big)}_{\mathcal{L}_{\text{KLD of }\mathbf{z}}} + \mathcal{L}_{\text{sparse}} \tag{16}$$

where β_1 denotes the hyperparameter to penalize the KL divergence of $\mathbf{z}$. In the following, we explain each part in details.

We first construct the variational lower bound of $p(\mathbf{x}|\mathbf{z})$. For a fixed $\hat{\mathbf{z}}$, we can define:

$$\mathcal{L}_{\text{cond}} = \mathbb{E}_{\hat{\epsilon}\sim q_\phi}\left[\log p_\theta(\hat{\mathbf{x}}|\hat{\mathbf{z}},\hat{\epsilon})\right] - \beta_2 \mathbb{E}_{\hat{\epsilon}\sim q_\phi}\big(\log q\left(\hat{\epsilon}|\mathbf{x}\right) - \log p_\gamma(\hat{\epsilon}|\hat{\mathbf{z}})\big) \tag{17}$$

where β_2 is the hyperparameter that balance the KL divergence of ϵ.

Additionally, to learn the latent representation $\hat{\mathbf{z}}$, we further regularize its approximate posterior $q(\hat{\mathbf{z}}|\mathbf{x})$ toward the prior $p_\delta(\mathbf{z})$ using the KL divergence. We then take the expectation of both $\mathcal{L}_{\text{cond}}$ and KLD of $\mathbf{z}$ with respect to $q\psi(\mathbf{z} \mid \mathbf{x})$.

We also design the regularizer $\mathcal{L}_{\text{sparse}}$ by:

$$\mathcal{L}_{\text{sparse}} = \Big|\frac{\mathbb{E}_{\hat{\epsilon}\sim q_\phi}\left[\partial \log p_\theta(\hat{\mathbf{x}}|\hat{\mathbf{z}},\hat{\epsilon})\right]}{\partial \hat{\mathbf{z}}} + \frac{\mathbb{E}_{\hat{\mathbf{z}}\sim q_\psi}\left[\partial \log p_\gamma(\hat{\epsilon}|\hat{\mathbf{z}})\right]}{\partial \hat{\mathbf{z}}}\Big|_1 \tag{18}$$

$|*|_1$ denotes the ℓ_1 norm that enforces the sparsity regularization in the Regularization 3. We use the $\log p_\theta(\hat{\mathbf{x}}|\hat{\mathbf{z}},\hat{\epsilon})$ and $\log p_\gamma(\hat{\epsilon}|\hat{\mathbf{z}})$ instead of the actual $p_\theta(\hat{\mathbf{x}}|\hat{\mathbf{z}})$ is given the fact that $\frac{\partial p(\mathbf{x}^i|\mathbf{z})}{\partial \mathbf{z}^j} = 0 \iff \frac{\partial \log p(\mathbf{x}^i|\mathbf{z})}{\partial \mathbf{z}^j} = 0$. where the ℓ_1 penalty serves as a differentiable surrogate for the ℓ_0 norm. *Notably, we do not require any knowledge of the theoretical partitions of* $\mathbf{x}$ *in Definition 3*.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Data: We evaluate the precise and approximate identifiability results using synthetic data described in Section C.1, and the assumption justifications are provided in Section C.2. For precise identifiability, since Proposition 2 is established upon the subspace result of Theorem 1, we create a dataset evaluates component-wise recovery of $\mathbf{z}$. For approximate identifiability, we synthesize ***another*** dataset satisfying the assumptions required by Theorem 2. This setting evaluates approximate subspace recovery only, since approximate component-wise identifiability is left for future work.

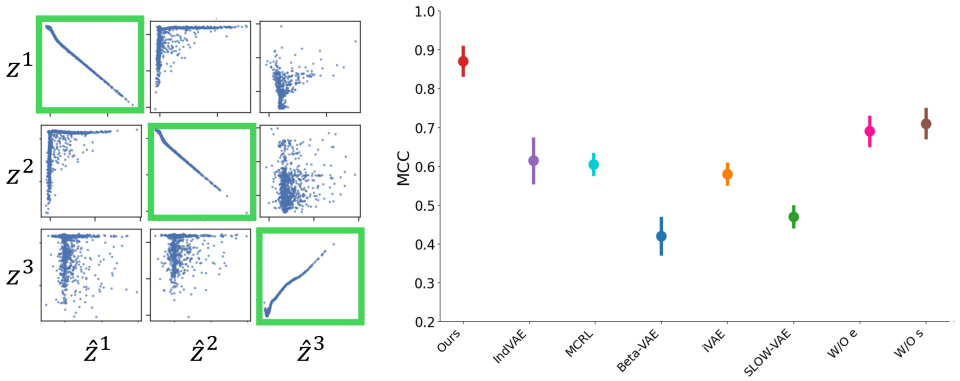
Metrics: For precise identifiability, we use the Mean Correlation Coefficient (MCC) to evaluate component-wise recovery. MCC measures the average absolute correlation between matched coordinates of the true and estimated latent variables. MCC scores range from 0 to 1, with larger values indicating better component-wise identification.

For approximate identifiability, we report the coefficient of determination R^2 after aligning $\hat{\mathbf{z}}$ with $\mathbf{z}$. A higher value of R^2 corresponds to a smaller discrepancy between $\hat{\mathbf{z}}$ and its admissible equivalent $\mathbf{z}$ and therefore indicates better approximate subspace identifiability.

Comparison Methods. We compare our method with several representative approaches for latent variable identification. IndVAE (Hu & Schennach, 2008; Zheng et al., 2025) assumes conditional independence among observations given the latent variables; its objective is detailed in Section C.4. MCRL (Sun et al., 2025) assumes that the noise term ϵ is independent of $\mathbf{z}$. We additionally compare against Beta-VAE (Higgins et al., 2016), iVAE (Khemakhem et al., 2020), and SLOW-VAE (Klindt et al., 2020). We further conduct ablations to asses the impact of latent-dependent noise and structural sparsity. "W/O e" removes the dependence $\epsilon = e(\mathbf{z})$ from Eq. equation 1, whereas "W/O s" removes the structural-sparsity regularization imposed on Φ. Section C.3 provides the implementation details of our approach.

5.2 RESULTS AND DISCUSSIONS:

Precise Identifiability: As shown in Figure 3(b), our method achieves the highest MCC among all compared approaches, demonstrating more accurate component-wise recovery of the latent variables



(a) (b)

Figure 3: (a) Visualization of the correlations between each component of true latent variables ($\mathbf{z}^i$) and their corresponding component of estimated latent variables ($\hat{\mathbf{z}}^i$) using our approach. The green bounding boxes highlight the components that are identified. (b) Mean Correlation Coefficient (MCC) scores comparing our framework with state-of-the-art approaches, including IndVAE, MCRL, BetaVAE, iVAE, and SlowVAE, as well as the ablation baselines W/O e and W/O s.

under misspecified structure. Figure 3(a) provides a complementary visualization: each true latent coordinate $\mathbf{z}^i$ exhibits a clear one-to-one correspondence with a single estimated coordinate $\hat{\mathbf{z}}^i$, while the off-diagonal pairs show substantially weaker dependence. This behavior is consistent with component-wise recovery.

The ablation results further illustrate the importance of modeling the data-generating structure. IndVAE performs substantially worse because its conditional-independence assumption is violated by our synthetic data. Removing the latent-dependent noise model ("W/O e") also reduces MCC comparing with our full model, indicating that explicitly modeling $\epsilon = e(\mathbf{z}, \eta)$ is important for separating the latent signal from the dependent noise. Also, "W/O s" yields lower MCC than the full model, supporting the impact of the structural-sparsity regularization.

Approximate Identifiability: We use the same model architecture and the same training objective as in the precise-identifiability experiment; only $\overline{Per} < \eta$ does not hold in the the synthetic data. This setting therefore evaluates whether the same estimator can still recover a representation close to the admissible equivalence class under the approximate identifiability regime.

Figure 4 reports the R^2 scores. Our method achieves the highest R^2 among all compared approaches, indicating that $\hat{\mathbf{z}}$ retains the strongest correspondence with $\mathbf{z}$ in the approximate-identifiability regime. This is consistent with Theorem 2.

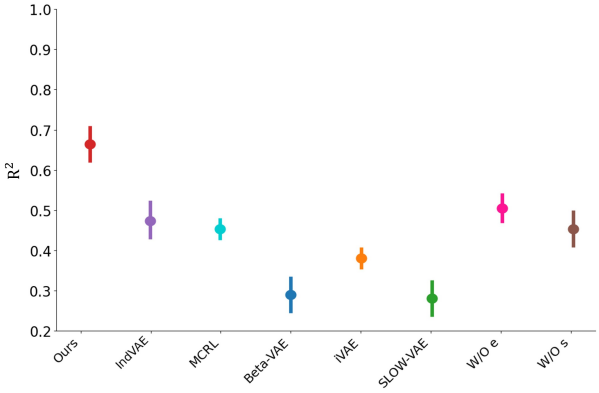


Figure 4: R^2 scores for comparison methods.

The comparison with IndVAE, MCRL, Beta-VAE, iVAE and SLOW-VAEfurther shows that explicitly modeling the misspecified latent-dependent-noise process improves latent recovery even outside the precise-identifiability regime. In particular, "W/O e" obtains a lower R^2 than the complete model, showing that modeling the dependence between ϵ and $\mathbf{z}$ remains beneficial for approximate subspace recovery. "W/O s" also decreases R^2, suggesting the benefit from structural sparsity.

Also, we postpone discussing the real-world experiments, and more ablation studies in Section D.

AI USE STATEMENT

In this work, we used generative AI tools solely to assist with language editing, including improving grammar, clarity, conciseness, and overall wording of the manuscript. Generative AI tools were not used to generate the research ideas, theoretical results, experimental design, or scientific conclusions. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final content of this work.

REFERENCES

Ali Ben Amor. An extension of henrici theorem for the joint approximate spectrum of commuting spectral operators. *Journal of the Australian Mathematical Society*, 75(2):233–246, 2003.

Jack Brady, Julius von Kügelgen, Sebastien Lachapelle, Simon Buchholz, Thomas Kipf, and Wieland Brendel. Interaction asymmetry: A general principle for learning composable abstractions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=cCl10IU836.

Nelson Dunford and Jacob T Schwartz. *Linear operators*. Wiley-interscience, 1971.

Minghao Fu, Biwei Huang, Zijian Li, Yujia Zheng, Ignavier Ng, Guangyi Chen, Yingyao Hu, and Kun Zhang. Learning general causal structures with hidden dynamic process for climate analysis. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=MZxhwFT7yF.

Yuqian Fu, Yu Xie, Yanwei Fu, and Yu-Gang Jiang. Styleadv: Meta style adversarial training for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24575–24584, 2023.

Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. doi: 10.1109/TNNLS.2023.3327962.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15013–15022, 2021.

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

Yingyao Hu and Susanne M Schennach. Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216, 2008.

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.

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

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

Lingjing Kong, Biwei Huang, Feng Xie, Eric Xing, Yuejie Chi, and Kun Zhang. Identification of nonlinear latent hierarchical models. *Advances in Neural Information Processing Systems*, 36: 2010–2032, 2023.

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=4pf_pOo0Dt.

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

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

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

Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 1405–1413, 2023.

Xinyao Li, Yuke Li, Zhekai Du, Fengling Li, Ke Lu, and Jingjing Li. Split to merge: Unifying separated modalities for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23364–23374, June 2024a.

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

Yuke Li, Yujia Zheng, Guangyi Chen, Kun Zhang, and Heng Huang. Identification of intermittent temporal latent process. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=6Pz7afmsOp.

Zijian Li, Minghao Fu, Junxian Huang, Yifan Shen, Ruichu Cai, Yuewen Sun, Guangyi Chen, and Kun Zhang. Towards identifiability of hierarchical temporal causal representation learning. *Advances in Neural Information Processing Systems*, 38:19680–19719, 2025b.

Zijian Li, Changze Zhou, Minghao Fu, Sanjay Manjunath, Fan Feng, Guangyi Chen, Yingyao Hu, Ruichu Cai, and Kun Zhang. Online time series forecasting with theoretical guarantees. *Advances in Neural Information Processing Systems*, 38:134299–134343, 2025c.

Wendong Liang, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal component analysis. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HszLRiHyfO.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. BISCUIT: Causal Representation Learning from Binary Interactions. In *The 39th Conference on Uncertainty in Artificial Intelligence*, 2023. URL https://openreview.net/forum?id=VS7Dn31xuB.

Hong Liu, Liang Hu, and Liqian Ma. Online rgb-d person re-identification based on metric model update. *CAAI Transactions on Intelligence Technology*, 2(1):48–55, 2017.

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

Xingchao Peng, Ben Usman, Neela Kaushik, Dequan Wang, Judy Hoffman, and Kate Saenko. Visda: A synthetic-to-real benchmark for visual domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2021–2026, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021a.

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

Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Kumar Ravikumar. From causal to concept-based representation learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=r5nev2SHtJ.

Mainak Singha, Harsh Pal, Ankit Jha, and Biplab Banerjee. AD-CLIP: Adapting domains in prompt space using CLIP. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 4355–4364, October 2023.

Xiangchen Song, Weiran Yao, Yewen Fan, Xinshuai Dong, Guangyi Chen, Juan Carlos Niebles, Eric Xing, and Kun Zhang. Temporally disentangled representation learning under unknown nonstationarity. *Advances in Neural Information Processing Systems*, 36, 2024.

Yuewen Sun, Lingjing Kong, Guangyi Chen, Loka Li, Gongxu Luo, Zijian Li, Yixuan Zhang, Yujia Zheng, Mengyue Yang, Petar Stojanov, Eran Segal, Eric P. Xing, and Kun Zhang. Causal representation learning from multimodal biomedical observations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=hjROBHstZ3.

Subash Timilsina, Hoang-Son Nguyen, Sagar Shrestha, and Xiao Fu. Content-style identification via differential independence. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=WcsBsGO6SW.

Lloyd N Trefethen. Spectra and pseudospectra: The behaviour of non-normal matrices and operators. In *The graduate student's guide to numerical analysis' 98: Lecture notes from the VIII EPSRC Summer School in Numerical Analysis*, pp. 217–250. Springer, 2005.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5018–5027, 2017.

Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 79–88, 2018.

Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pp. 5380–5389, 2017.

Shaoan Xie, Lingjing Kong, Mingming Gong, and Kun Zhang. Multi-domain image generation and translation with identifiability guarantees. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Danru Xu, Lauri Parkkonen, Sara Magliacane, and Aapo Hyvarinen. Identifying dependent components from multi-domain linear mixtures. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=OIhwTkpkS1.

Zhengwei Yang, Yuke Li, Qiang Sun, Basura Fernando, Heng Huang, and Zheng Wang. Cross-modal few-shot learning: a generative transfer learning framework, 2025. URL https://arxiv.org/abs/2410.10663.

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

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.

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

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

Yujia Zheng, Yang Liu, Jiaxiong Yao, Yingyao Hu, and Kun Zhang. Nonparametric factor analysis and beyond. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Fei Zhou, Peng Wang, Lei Zhang, Wei Wei, and Yanning Zhang. Revisiting prototypical network for cross domain few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20061–20070, 2023.

A NOTIONS

Table of notions

		Variables	
$\mathbf{x} \in \mathbb{R}^K$	Observations	$\hat{\mathbf{x}} \in \mathbb{R}^K$	Reconstructions
κ	$\|\|L_{\mathbf{x_a}\|\mathbf{z}}\|\|_{op}\|\|L^{-1}_{\mathbf{x_a}\|\mathbf{z}}\|\|_{op}$	ε	Identifiability error
$\mathbf{z} \in \mathbb{R}^N$	Latent variables	$\hat{\mathbf{z}} \in \mathbb{R}^N$	Latent variable estimations
ϵ	True dependent noise term	$\hat{\epsilon}$	Estimation of ϵ
η	Auxiliary variable	$\hat{\eta}$	Estimation of η
		Indices	
$\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$	The indices of partitions of $\mathbf{x}$	$\{A, B, C\}$	The indices of partitions of $\mathbf{x}'$
$n \in [N]$	The indices of $\mathbf{z}$	$\hat{n} \in [N]$	The indices of $\hat{\mathbf{z}}$
i	The index of ρ_Λ and ρ	j	The index of ρ_Λ and ρ
		Operators	
L	Integral linear operator	Λ	Multiplication operator
ρ_Λ	The eigenvalue of$L_{\mathbf{x_a}\|\mathbf{z}} L_{\mathbf{x_b}\|\mathbf{z}} L^{-1}_{\mathbf{x_a}\|\mathbf{z}}$	ρ	The eigenvalue of $L_{\mathbf{x_a},\mathbf{x_b}\|\mathbf{x_c}} L^{-1}_{\mathbf{x_a}\|\mathbf{x_c}}$
Per	Perturbation operator	$\overline{Per}$	The upper bound of $\|\|Per\|\|_{op}$
		True & learned model	
g	True mixing function	$\hat{g}$	Learned mixing function
e	True function of ϵ	$\hat{e}$	Learned function of ϵ
J_Φ	The jacobian of $_\Phi$	$J_{\hat{\Phi}}$	The jacobian of $\hat{g_\Phi}$
		Optimizations	
ψ	Parameters of posterior $q_\psi(\mathbf{z}\|\mathbf{x})$	ϕ	Parameters of posterior $q_\phi(\epsilon\|\mathbf{x})$
δ	Parameters of prior $p_\delta(\mathbf{z})$	γ	Parameters of prior $p_\gamma(\epsilon\|\mathbf{z})$
θ	Parameters of decoder $p_\theta(\mathbf{x}\|\mathbf{z})$	$\|*\|_1$	l_1 norm of $*$

B PROOF OF THEOREMS

B.1 PREVIOUS RESULTS

In this section, we recapitulate and summarize the previous results from (Hu & Schennach, 2008; Zheng et al., 2025).

For $\mathbf{x}' = \{\mathbf{x}'_A, \mathbf{x}'_B, \mathbf{x}'_C\}$, where $\{\mathbf{x}'_A, \mathbf{x}'_B, \mathbf{x}'_C\}$ denote the three disjoint partitions, $p(\mathbf{x}'|\mathbf{z}') = p(\mathbf{x}'_A|\mathbf{z}')p(\mathbf{x}'_B|\mathbf{z}')p(\mathbf{x}'_C|\mathbf{z}')$ [2];

Consider observed variables $\mathbf{x}' \in \mathbb{R}^K$ and the estimated latent variables $\hat{\mathbf{z}}' \in \mathbb{R}^N$, suppose that there exist functions $\hat{g}'$ and $\hat{e}'$, and the following assumptions hold:

1. *The joint distribution of $(\mathbf{x}, \mathbf{z})$ admits a bounded density with respect to a suitable product measure defined on their supports. Furthermore, all marginal and conditional densities derived from this joint distribution are also bounded;*
2. *The operators $L_{\mathbf{x}'_A|\mathbf{z}'}$ and $L_{\mathbf{x}'_A|\mathbf{x}'_B}$ are injective;*
3. *$\forall \underline{\mathbf{z}}' \neq \bar{\mathbf{z}}', p(\mathbf{x}'_C; \underline{\mathbf{z}}') \neq p(\mathbf{x}'_C; \bar{\mathbf{z}}')$;*

[2] **We emphasize that the partition in both previous and our work is an existential structural condition for the identifiability analysis. We do not require the partition labels, nor separate access to $\mathbf{x_a}$, $\mathbf{x_b}$, and $\mathbf{x_c}$, during learning and inference.**

then $\mathbf{z}'$ *must be identified up to an invertible transformation* h'.

Proof: Given Assumption *1*, we can obtain the following:

$$\begin{aligned}
p_{\mathbf{x}'_C\mathbf{x}'_A|\mathbf{x}'_B}(\mathbf{x}'_C,\mathbf{x}'_A|\mathbf{x}'_B) &= \int p_{\mathbf{x}'_C\mathbf{x}'_A\mathbf{z}'|\mathbf{x}'_B}(\mathbf{x}'_C,\mathbf{x}'_A,\mathbf{z}'|\mathbf{x}'_B)d\mathbf{z} \\
&= \int p_{\mathbf{x}'_C|\mathbf{x}'_A\mathbf{z}'\mathbf{x}'_B}(\mathbf{x}'_C|\mathbf{x}'_A,\mathbf{z}',\mathbf{x}'_B)p_{\mathbf{x}'_A\mathbf{z}'|\mathbf{x}'_B}(\mathbf{x}'_A,\mathbf{z}'|\mathbf{x}'_B)d\mathbf{z} \\
&= \int p_{\mathbf{x}'_C|\mathbf{x}'_A\mathbf{z}'}(\mathbf{x}'_C|\mathbf{x}'_A,\mathbf{z}')p_{\mathbf{x}'_A\mathbf{z}'|\mathbf{x}'_B}(\mathbf{x}'_A,\mathbf{z}'|\mathbf{x}'_B)d\mathbf{z} \\
&= \int p_{\mathbf{x}'_C|\mathbf{x}'_A\mathbf{z}'}(\mathbf{x}'_C|\mathbf{x}'_A,\mathbf{z}')p_{\mathbf{x}'_A|\mathbf{z}'\mathbf{x}'_B}(\mathbf{x}'_A|\mathbf{z}',\mathbf{x}'_B)p_{\mathbf{z}'|\mathbf{x}'_B}(\mathbf{z}'|\mathbf{x}'_B)d\mathbf{z} \\
&= \int p_{\mathbf{x}'_C|\mathbf{x}'_A\mathbf{z}'}(\mathbf{x}'_C|\mathbf{x}'_A,\mathbf{z}')p_{\mathbf{x}'_A|\mathbf{z}'}(\mathbf{x}'_A|\mathbf{z}')p_{\mathbf{z}'|\mathbf{x}'_B}(\mathbf{z}'|\mathbf{x}'_B)d\mathbf{z} \\
&= \int p_{\mathbf{x}'_C|\mathbf{z}'}(\mathbf{x}'_C|\mathbf{z}')p_{\mathbf{x}'_A|\mathbf{z}'}(\mathbf{x}'_A|\mathbf{z}')p_{\mathbf{z}'|\mathbf{x}'_B}(\mathbf{z}'|\mathbf{x}'_B)d\mathbf{z} \qquad (19)
\end{aligned}$$

Leveraging this eqation and Definition 4, we can derive the following operator:

$$\begin{aligned}
(L_{\mathbf{x}'_C;\mathbf{x}'_A|\mathbf{x}'_B}f')(\mathbf{x}'_A) &= \int\int p_{\mathbf{x}'_A|\mathbf{z}'}(\mathbf{x}'_A|\mathbf{z}')p_{\mathbf{x}'_C|\mathbf{z}'}(\mathbf{x}'_C|\mathbf{z}')p_{\mathbf{z}'|\mathbf{x}'_B}(\mathbf{z}'|\mathbf{x}'_B)f'(\mathbf{x}'_B)d\mathbf{x}'_Bd\mathbf{z} \\
&= \int p_{\mathbf{x}'_A|\mathbf{z}'}(\mathbf{x}'_A|\mathbf{z}')(\Lambda_{\mathbf{x}'_C;\mathbf{z}'}L_{\mathbf{z}'|\mathbf{x}'_B}f')(\mathbf{z}')d\mathbf{z} \\
&= (L_{\mathbf{x}'_A|\mathbf{z}'}\Lambda_{\mathbf{x}'_C;Z}L_{\mathbf{z}'|\mathbf{x}'_B}f')(\mathbf{x}'_A) \qquad (20)
\end{aligned}$$

The above equation indicates:

$$L_{\mathbf{x}'_C;\mathbf{x}'_A|\mathbf{x}'_B} = L_{\mathbf{x}'_A|\mathbf{z}'}\Lambda_{\mathbf{x}'_C;\mathbf{z}'}L_{\mathbf{z}'|\mathbf{x}'_B} \qquad (21)$$

This equivalence holds over some functions space $\mathcal{G}(\mathcal{Z})$, given the factorization properties of the conditional densities established earlier.

Now, integrating over $\mathbf{x}'_C$, by using the fact that: $\int L_{\mathbf{x}'_C;\mathbf{x}'_A|\mathbf{x}'_B}f'(\mathbf{x}'_C)d\mathbf{x}'_C = L_{\mathbf{x}'_A|\mathbf{x}'_B}f'$ we can obtain:

$$\begin{aligned}
L_{\mathbf{x}'_A|\mathbf{x}'_B}f'(\mathbf{x}'_A) &= \int p_{\mathbf{x}'_A|\mathbf{x}'_B}(\mathbf{x}'_A|\mathbf{x}'_B)f'(\mathbf{x}'_B)d\mathbf{x}'_B \\
&= \int\int p_{\mathbf{x}'_A|\mathbf{z}',\mathbf{x}'_B}(\mathbf{x}'_A|\mathbf{z}',\mathbf{x}'_B)p_{\mathbf{z}'|\mathbf{x}'_B}(\mathbf{z}'|\mathbf{x}'_B)f'(\mathbf{x}'_B)d\mathbf{x}'_Bd\mathbf{z} \\
&= \int p_{\mathbf{x}'_A|\mathbf{z}'}(\mathbf{x}'_A|\mathbf{z}')[L_{\mathbf{z}'|\mathbf{x}'_B}f']d\mathbf{z} \\
&= L_{\mathbf{x}'_A|\mathbf{z}'}L_{\mathbf{z}'|\mathbf{x}'_B}f'](\mathbf{x}'_A) \qquad (22)
\end{aligned}$$

where the second equation leverages $\mathbf{x}'_A \perp \mathbf{x}'_B \mid \mathbf{z}'$. By assuming the inejctivity of $L_{\mathbf{x}'_A|\mathbf{z}'}$ in Assumption *2*, we can arrive at $L_{Z|\mathbf{x}'_B} = L^{-1}_{\mathbf{x}'_A|\mathbf{z}'}L_{\mathbf{x}'_A|\mathbf{x}'_B}$. Substitute this to Eq. 15:

$$L_{\mathbf{x}'_C;\mathbf{x}'_A|\mathbf{x}'_B}L^{-1}_{\mathbf{x}'_A|\mathbf{x}'_B} = L_{\mathbf{x}'_A|\mathbf{z}'}\Lambda_{\mathbf{x}'_C;Z}L^{-1}_{\mathbf{x}'_A|\mathbf{z}'} \qquad (23)$$

where the LHS involves only observable variables, and the RHS explicitly depends on the latent variable $\mathbf{z}'$. $\Lambda_{\mathbf{x}'_C;\mathbf{z}'}$ determines the eigenvalues of $L_{\mathbf{x}'_A|Z}\Lambda_{\mathbf{x}'_C;\mathbf{z}'}L^{-1}_{\mathbf{x}'_A|\mathbf{z}'}$, whose diagonal entries correspond to the conditional distributions $p(\mathbf{x}'_C|\mathbf{z}')$. Each $\mathbf{z}'$ indexing a distinct conditional distribution of $p(\mathbf{x}'_C|\mathbf{z}')$. Under Assumption *3*, where $p(\mathbf{x}'_C \mid \mathbf{z}')$ are distinct for different values of $\mathbf{z}'$, the eigenvalues are distinct. This allows a bijective mapping $h' : \mathcal{Z} \to \mathcal{Z}$ to permute $\mathbf{z}'$ while preserving the values of $p(\mathbf{x}'_C \mid \mathbf{z}')$. Therefore, the latent variable can only be recovered up to such a permutation, i.e., $\hat{\mathbf{z}'} = h'(\mathbf{z}')$, which yields the identifiability up to an invertible transformation h'.

B.2 Proof of Proposition 1 and Theorem 1

Proposition 1: Suppose Assumptions 1 $\sim$ 3 hold, and if both ϱ^i and ϱ^i_Λ include only ρ^i and ρ^i_Λ, respectively. Then, for all $i > 1$ and for all $\mathbf{z}$,

$$\rho^i_\Lambda \in \left(\max\{\rho^i - \kappa\overline{Per}, \rho^{i-1} + \kappa\overline{Per}\}, \rho^i + \kappa\overline{Per}\right) \tag{24}$$

and if $i = 1$, for all $\mathbf{z}$,

$$\rho^i_\Lambda \in (\rho^i - \kappa\overline{Per}, \rho^i + \kappa\overline{Per}) \tag{25}$$

guarantee the distinct spectral values of $L_{\mathbf{x_a}|\mathbf{z}}\Lambda_{\mathbf{x_b}|\mathbf{z}}L^{-1}_{\mathbf{x_a}|\mathbf{z}}$.

Proof:

Recall that we define:

$$\eta = \min_{i\neq j} \frac{|\rho^i - \rho^j|}{2\kappa} - \alpha > 0 \quad \text{and} \quad \overline{Per} < \eta, \tag{26}$$

where $\alpha > 0$ denotes constant, and $\kappa = ||L_{\mathbf{x_a}|\mathbf{z}}||_{op}||L^{-1}_{\mathbf{x_a}|\mathbf{z}}||_{op}$. Notably, $\kappa \geq 1$, since $||L_{\mathbf{x_a}|\mathbf{z}}L^{-1}_{\mathbf{x_a}|\mathbf{z}}||_{op} = ||I||_{op} \leq ||L_{\mathbf{x_a}|\mathbf{z}}||_{op}||L^{-1}_{\mathbf{x_a}|\mathbf{z}}||_{op}$, where I denotes the identity operator.

Suppose that two distinct spectral values, denoted ρ^1_Λ and ρ^2_Λ, fall within the same interval $(\rho^i - \kappa\overline{Per}, \rho^i + \kappa\overline{Per})$. we have:

$$\begin{aligned} |\rho^1_\Lambda - \rho^i| &\leq \kappa\overline{Per} < |\rho^i - \rho^j|, \\ |\rho^2_\Lambda - \rho^i| &\leq \kappa\overline{Per} < |\rho^i - \rho^j|. \end{aligned} \tag{27}$$

Applying the triangle inequality, we obtain:

$$|\rho^1_\Lambda - \rho^2_\Lambda| \leq |\rho^1_\Lambda - \rho^i| + |\rho^2_\Lambda - \rho^i| < |\rho^i - \rho^j| = 2\kappa\eta. \tag{28}$$

However, inequality 28 implies:

$$\frac{|\rho^1 - \rho^2|}{2\kappa} < \eta, \tag{29}$$

which directly contradicts the definition of η. Thus, the uniqueness of each spectral value ρ^i_Λ within its respective interval is proved.

We reorganize the Theorem 1 below:

Theorem 1 *Consider observed variables $\mathbf{x} \in \mathbb{R}^K$ and the estimated latent variables $\hat{\mathbf{z}} \in \mathbb{R}^N$, suppose the following assumptions hold:*

- *i The joint distribution of $(\mathbf{x}, \mathbf{z})$ admits a bounded density with respect to a suitable product measure defined on their supports. Furthermore, all marginal and conditional densities derived from this joint distribution are also bounded;*
- *ii The operators $L_{\mathbf{x_a}|\mathbf{z}}$ and $L_{\mathbf{x_a}|\mathbf{x_c}}$ are injective and bounded-below;*
- *iii $L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}}L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}$ has distinct spectral values with cardinality equal to that of $\Lambda_{\mathbf{x_b}|\mathbf{z}}$. Also, Let $\Theta : \mathcal{Z} \to \mathcal{D}, s.t. p(\mathbf{x_b}|\mathbf{z}) = \Theta(\mathbf{z})$ denotes the map from $\mathcal{Z}$ to $\mathcal{D}$. Θ is injective;*
- *iv There exists an operator M such that $M(L_{\mathbf{x_b}|\mathbf{z}}) = M(L_{\mathbf{x_b}|\tilde{h}(\mathbf{z})}) = t(\mathbf{z})$, where t is a differentiable transformation.*

then for $\tilde{h} \in \tilde{\mathcal{H}}$ and $t \in \mathcal{T}$ ($\tilde{\mathcal{H}} \cap \mathcal{T} \neq \emptyset$), if $h \in \tilde{\mathcal{H}} \cap \mathcal{T} \Rightarrow \hat{\mathbf{z}} = h(\mathbf{z}) = \tilde{h}(\mathbf{z}) = t(\mathbf{z})$. In other words, $\mathbf{z}$ must be subspace identified.

Proof:

Our goal is to demonstrate subspace identifiability of $\mathbf{z}$ from the data generated process in Eq. 1. To such an end, we proceed in the following steps:

Step 1: Operator Construction and Spectral Decomposition.

Our approach follows the previous results in Sec. B.1, and proceeds to decompose the bounded linear operator $L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}}$. The structural misspecification in Definition 3 violates the conditional independence by $p(\mathbf{x}|\mathbf{z}) \neq p(\mathbf{x_a}|\mathbf{z})p(\mathbf{x_b}|\mathbf{z})p(\mathbf{x_c}|\mathbf{z})$, thus introduces a discrepancy between $L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}}$ and $L_{\mathbf{x_a}|\mathbf{z}}L_{\mathbf{x_b}|\mathbf{z}}L_{\mathbf{z}|\mathbf{x_c}}$ according to Eq. 21. Consequently, we can define the difference operators:

$$\begin{aligned}\mathbf{d}(\mathbf{x_b}) &= L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}} - L_{\mathbf{x_a}|\mathbf{z}}L_{\mathbf{x_b}|\mathbf{z}}L_{\mathbf{z}|\mathbf{x_c}},\\ \mathbf{d} &= \int \mathbf{d}(\mathbf{x_b})d\mathbf{x_b} = L_{\mathbf{x_a}|\mathbf{x_c}} - L_{\mathbf{x_a}|\mathbf{z}}L_{\mathbf{z}|\mathbf{x_c}}.\end{aligned} \tag{30}$$

Assumption i guarantees boundedness of each term in Eq. 30. Leveraging these, we rewrite:

$$\begin{aligned}L_{\mathbf{x_a}|\mathbf{z}}\Lambda_{\mathbf{x_b}|\mathbf{z}}L^{-1}_{\mathbf{x_a}|\mathbf{z}} &= (L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}} - \mathbf{d}(\mathbf{x_b}))(L_{\mathbf{x_a}|\mathbf{x_c}} - \mathbf{d})^{-1}\\ &= L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}}L^{-1}_{\mathbf{x_a}|\mathbf{x_c}} + Per,\end{aligned} \tag{31}$$

where Per represents the perturbation term arising from the violation of conditional independence.

Step 2: Spectral resolution of $L_{\mathbf{x_a}|\mathbf{z}}\Lambda_{\mathbf{x_b}|\mathbf{z}}L^{-1}_{\mathbf{x_a}|\mathbf{z}}$

At this point, we aim to understand if $L_{\mathbf{x_a}|\mathbf{z}}\Lambda_{\mathbf{x_b}|\mathbf{z}}L^{-1}_{\mathbf{x_a}|\mathbf{z}}$ is unique spectral decomposition of RHS of Eq. 31 with $Per \neq 0$. We take inspiration from (Dunford & Schwartz, 1971) to proceed with the necessary and sufficient condition by using the projection-valued measure in the following steps.

Assumption i along with the fact that $p(\mathbf{x_b}|\mathbf{z})$ being real-valued suggest that the multiplication operator $\Lambda_{\mathbf{x_b}|\mathbf{z}}$ is bounded self-adjoint, and therefore admits a unique orthogonal projection-valued measure $E_{\mathbf{x_b}}$ such that

$$\Lambda_{\mathbf{x_b}|\mathbf{z}} = \int_{\sigma(\Lambda_{\mathbf{x_b}|\mathbf{z}})} \lambda \, dE_{\mathbf{x_b}}(\lambda) \tag{32}$$

where σ denotes the spectral value, $(E_{\mathbf{x_b}}(\Delta)f)(\mathbf{z}) = \mathbf{1}_{\{p(\mathbf{x_b}|\mathbf{z})\in\Delta\}}f(\mathbf{z})$ for every Borel set $\Delta \subseteq \mathbb{R}$, 1 denotes the indicator and f denotes some function.

For $L_{\mathbf{x_a}|\mathbf{z}}\Lambda_{\mathbf{x_b}|\mathbf{z}}L^{-1}_{\mathbf{x_a}|\mathbf{z}}$, define the projection-valued measure $P_{\mathbf{x_b}}$ by:

$$P_{\mathbf{x_b}}(\Delta) \coloneqq L_{\mathbf{x_a}|\mathbf{z}}E_{\mathbf{x_b}}(\Delta)L^{-1}_{\mathbf{x_a}|\mathbf{z}} \tag{33}$$

Given the properties of projection-valued measure, we can obtain: $P_{x_b}(\mathbb{R}) = I_{\mathcal{K}}$ ($\mathcal{K}$ denotes the range of $L_{\mathbf{x_a}|\mathbf{z}}$). These equations yield

$$L_{\mathbf{x_a}|\mathbf{z}}\Lambda_{\mathbf{x_b}|\mathbf{z}}L^{-1}_{\mathbf{x_a}|\mathbf{z}} = \int_{\sigma(L_{\mathbf{x_a}|\mathbf{z}}\Lambda_{\mathbf{x_b}|\mathbf{z}}L^{-1}_{\mathbf{x_a}|\mathbf{z}})} \lambda \, dP_{\mathbf{x_b}}(\lambda) \tag{34}$$

with $\sigma(L_{\mathbf{x_a}|\mathbf{z}}\Lambda_{\mathbf{x_b}|\mathbf{z}}L^{-1}_{\mathbf{x_a}|\mathbf{z}}) = \sigma(\Lambda_{\mathbf{x_b}|\mathbf{z}})$. Moreover, $\operatorname{Ran} P_{\mathbf{x_b}}(\Delta) = L_{\mathbf{x_a}|\mathbf{z}} \operatorname{Ran} E_{x_b}(\Delta)$ and $\ker P_{x_b}(\Delta) = L_{\mathbf{x_a}|\mathbf{z}} \operatorname{Ran} E_{x_b}(\mathbb{R} \setminus \Delta)$, where Ran denote the range. Hence, $P_{\mathbf{x_b}}(\Delta)$ is uniquely determined by its range and null space.

Since $E_{x_b}(\Delta)$ is a spectral projection, the Hilber space $\mathcal{H}_z$ of $\Lambda_{\mathbf{x_b}|\mathbf{z}}$ is $\mathcal{H}_z = \operatorname{Ran} E_{\mathbf{x_b}}(\Delta) \oplus \operatorname{Ran} E_{\mathbf{x_b}}(\mathbb{R} \setminus \Delta)$. By invertibility of $L_{\mathbf{x}_a|\mathbf{z}}$,

$$\mathcal{K} = \operatorname{Ran} P_{\mathbf{x_b}}(\Delta) \oplus \ker P_{\mathbf{x_b}}(\Delta).$$

Suppose $w \in \operatorname{Ran} P_{x_b}(\Delta) \cap \ker P_{x_b}(\Delta)$. Since $w \in \operatorname{Ran} P_{x_b}(\Delta)$, there exists some $a \in \mathcal{K}$ such that $w = P_{x_b}(\Delta)a$. Hence, $P_{x_b}(\Delta)w = P_{x_b}(\Delta)^2 a = P_{x_b}(\Delta)a = w$. On the other hand, since $w \in \ker P_{x_b}(\Delta)$, $P_{x_b}(\Delta)w = 0$. Therefore $w = 0$, implying

$$\operatorname{Ran} P_{x_b}(\Delta) \cap \ker P_{x_b}(\Delta) = \{0\}.$$

Hence, every $f \in \mathcal{K}$ admits a unique decomposition $f = u + v$, where $u \in \operatorname{Ran} P_{\mathbf{x_b}}(\Delta)$ and $v \in \ker P_{\mathbf{x_b}}(\Delta)$, $\Rightarrow P_{\mathbf{x_b}}(\Delta)f = u$. Therefore, any projection with the same range and null space must map $f = u + v$ to its unique range component u, and is therefore identical to $P_{\mathbf{x_b}}(\Delta)$. Thus, for a fixed data generating process in Eq. 1, Eq. 31 admits an unique spectral projections.

Step 3: Connecting Bijection $\tilde{h}$ with Differentiable Transformation h.

Proposition 1 has proven the distinctness of ρ_Λ, which is determined by $p(\mathbf{x_b}|\mathbf{z})$. In combination of the injectivity map assumption in Assumption iii, $p(\mathbf{x_b}|\mathbf{z})$ corresponds to a particualr $\mathbf{z}$. Since ρ_Λ is the essential range of $p(\mathbf{x_b}|\mathbf{z})$, it remains invariant by re-labeling through permuting $\mathbf{z}$. This can be formally described by $\mathbf{z}' = h'(\mathbf{z})$, where h' denotes a bijection $h' : \mathcal{Z} \to \mathcal{Z}$, and $\mathbf{z}'$ is a permuted version of $\mathbf{z}$.

Definition 1 requires the invertible transformation h to be differentiable. However, $\tilde{h}$ alone may not satisfy this differentiability constraint. To resolve this, we invoke Assumption iv, which guarantees:

$$M(p(\mathbf{x_b}|\mathbf{z})) = M(p(\mathbf{x_b}|h'(\mathbf{z}))) = h(\mathbf{z})$$

where is $h(\mathbf{z}) = \hat{\mathbf{z}}$ iff h is differentiable.. Hence, we conclude the subspace identifiability of $\mathbf{z}$.

B.3 Proof of Theorem 2

We reorganize the Theorem 2 below:

Theorem 2 *Consider observed variables $\mathbf{x} \in \mathbb{R}^K$ and the estimated latent variables $\hat{\mathbf{z}} \in \mathbb{R}^N$, suppose the following assumptions hold:*

- *i The joint distribution of $(\mathbf{x}, \mathbf{z})$ admits a bounded density with respect to a suitable product measure defined on their supports. Furthermore, all marginal and conditional densities derived from this joint distribution are also bounded;*
- *ii The operators $L_{\mathbf{x_a}|\mathbf{z}}$ and $L_{\mathbf{x_a}|\mathbf{x_c}}$ are injective and bounded-below;*
- *iii $L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}} L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}$ has distinct spectral values with cardinality equal to that of $\Lambda_{\mathbf{x_b}|\mathbf{z}}$. Also, Let $\Theta : \mathcal{Z} \to \mathcal{D}, s.t. p(\mathbf{x_b}|\mathbf{z}) = \Theta(\mathbf{z})$ denotes the map from $\mathcal{Z}$ to $\mathcal{D}$. Θ is injective;*
- *iv $\mathcal{Z}$ is compact, and Θ in Assumption iii is continuously differentiable. There exists $c_\Theta > 0$ such that, for an admissible invertible transformation h, Θ satisfies the lower bound of Bi-Lipschitz continuity by:*

$$c_\Theta |\hat{\mathbf{z}} - h(\mathbf{z})| \leq \|\Theta(\mathbf{z}) - \widehat{\Theta}(\hat{\mathbf{z}})\|_{\mathcal{F}}. \tag{35}$$

where $\mathcal{F}$ denotes the function norm.

Then we can obtain $\varepsilon \leq \frac{\kappa\overline{Per} + \kappa\widehat{\overline{Per}}}{c_\Theta}$.

Proof:

Following Steps 1 and 2 in the proof of Theorem 1, for each fixed $\mathbf{x_b}$, the multiplication operator $\Lambda_{\mathbf{x_b}|\mathbf{z}}$ is given by

$$(\Lambda_{\mathbf{x_b}|\mathbf{z}} f)(\mathbf{z}) = p(\mathbf{x_b} \mid \mathbf{z}) f(\mathbf{z}).$$

Since Θ is continuous, $\forall \mathbf{z} \in \mathcal{Z}, p(\mathbf{x_b}|\mathbf{z}) \in ess\, ran_{\mathbf{z}} p(\mathbf{x_b}|\mathbf{z}) = \rho_\Lambda$, where $ess\, ran$ denotes the essential range.

Given $dist(\rho_\Lambda, \rho) \leq \kappa\overline{Per}$ in Eq. 5, we can define a similar bound by $dist(\hat{\rho}_\Lambda, \rho) \leq \hat{\kappa}\widehat{\overline{Per}}$. Therefore, there always exist i, such that:

$$\begin{aligned} |\rho^i_\Lambda - \rho^i_{\hat{\Lambda}|}| &= |\rho^i_\Lambda - \rho^i + \rho^i - \rho^i_{\hat{\Lambda}}| \\ &\leq |\rho^i_\Lambda - \rho^i| + |\rho^i_{\hat{\Lambda}} - \rho^i| \\ &\leq \kappa\overline{Per} + \hat{\kappa}\widehat{\overline{Per}} \end{aligned} \tag{36}$$

where the first $\leq$ is because of the triangle inequality. Eq. 36 suggests that, for almost all $\mathbf{z}, \hat{\mathbf{z}}$:

$$|p^i(\mathbf{x_b}|\mathbf{z}) - p^i(\mathbf{x_b}|\hat{\mathbf{z}})| \leq \kappa\overline{Per} + \hat{\kappa}\widehat{\overline{Per}} \tag{37}$$

Therefore, Taking the essential supremum over $\mathbf{x_b}$ gives

$$\begin{aligned} \|\Theta(\mathbf{z}) - \widehat{\Theta}(\hat{\mathbf{z}})\|_{\mathcal{F}} &= ess\, sup_{\mathbf{x_b}} |p^i(\mathbf{x_b} \mid \mathbf{z}) - p^i(\mathbf{x_b} \mid \hat{\mathbf{z}})| \\ &\leq \kappa\overline{\mathrm{Per}} + \hat{\kappa}\widehat{\overline{\mathrm{Per}}}. \end{aligned} \tag{38}$$

By Assumption iv,

$$c_\Theta|\hat{\mathbf{z}} - h(\mathbf{z})| \leq \|\Theta(\mathbf{z}) - \widehat{\Theta}(\hat{\mathbf{z}})\|_{\mathcal{F}} \tag{39}$$

Combining this inequality with Eq. equation 38 yields

$$\varepsilon = \inf_{h\in\mathcal{H}} |\hat{\mathbf{z}} - h(\mathbf{z})| \leq \frac{\kappa\overline{\mathrm{Per}} + \hat{\kappa}\widehat{\overline{\mathrm{Per}}}}{c_\Theta}$$

This completes the proof.

B.4 Discussions on $\overline{Per}$

In this secion, we discuss the bound of $\overline{Per}$ since it plays a critical role in Theorem 1. We present the following Lemma with the assistance of Assumption 3:

Lemma 1. *Let $\|\cdot\|$ be the operator norm. Consider an operator $L(\cdot)$ subject to the distance operator $\mathbf{d}$ in Eq. 30. If $\|L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}\mathbf{d}\| < 1$, then $(L_{\mathbf{x_a}|\mathbf{x_c}} - \mathbf{d})^{-1} \leq L^{-1}(\mathbf{x_a}|\mathbf{x_c}) + \frac{||L^{-1}(\mathbf{x_a}|\mathbf{x_c})||^2||\mathbf{d}||}{1-||L^{-1}(\mathbf{x_a}|\mathbf{x_c})\times\mathbf{d}||}$*

Proof: Let $(L_{\mathbf{x_a}|\mathbf{x_c}} - \mathbf{d})^{-1} = L^{-1}_{\mathbf{x_a}|\mathbf{x_c}} + B$, we can formalize:

$$\begin{aligned}
L_{\mathbf{x_a}|\mathbf{z}}L_{\mathbf{x_b}|\mathbf{z}}L^{-1}_{\mathbf{x_a}|\mathbf{z}} &= (L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}} - \mathbf{d}(\mathbf{x_b}))(L_{\mathbf{x_a}|\mathbf{x_c}} - \mathbf{d})^{-1} \\
&= (L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}} - \mathbf{d}(\mathbf{x_b}))(L^{-1}_{\mathbf{x_a}|\mathbf{x_c}} + B) \\
&= L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}}L^{-1}_{\mathbf{x_a}|\mathbf{x_c}} + \underbrace{L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}}B - \mathbf{d}(\mathbf{x_b})L^{-1}_{\mathbf{x_a}|\mathbf{x_c}} - \mathbf{d}(\mathbf{x_b})B}_{Per}
\end{aligned} \tag{40}$$

where we can bound B as follows:

$$\begin{aligned}
B &= (L_{\mathbf{x_a}|\mathbf{x_c}} - \mathbf{d})^{-1} - L^{-1}_{\mathbf{x_a}|\mathbf{x_c}} \\
&= (L_{\mathbf{x_a}|\mathbf{x_c}}(I - L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}\mathbf{d}))^{-1} - L^{-1}_{\mathbf{x_a}|\mathbf{x_c}} \\
&= (I - L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}\mathbf{d})^{-1}L^{-1}_{\mathbf{x_a}|\mathbf{x_c}} - L^{-1}_{\mathbf{x_a}|\mathbf{x_c}} \\
&\Rightarrow ||B|| \leq ||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}|| \times ||(I - L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}\mathbf{d})^{-1} - I|| \\
&\Rightarrow ||B|| \leq ||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}||(||(I - L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}\mathbf{d})^{-1}|| + || - I||) \\
&\Rightarrow ||B|| \leq ||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}||(\frac{1}{1 - ||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}\mathbf{d}||} - 1) \\
&\Rightarrow ||B|| \leq ||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}||(\frac{||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}\mathbf{d}||}{1 - ||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}\mathbf{d}||}) \\
&\Rightarrow ||B|| \leq \frac{||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}||^2||\mathbf{d}||}{1 - ||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}\mathbf{d}||}
\end{aligned} \tag{41}$$

Substitute the result from Eq. 41 to Eq. 4, we can obtain that

$$\|Per\| \leq \|L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}}\frac{||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}||^2||\mathbf{d}||}{1 - ||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}\mathbf{d}||} - \mathbf{d}(\mathbf{x_b})L^{-1}_{\mathbf{x_a}|\mathbf{x_c}} - \mathbf{d}(\mathbf{x_b})\frac{||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}||^2||\mathbf{d}||}{1 - ||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}\mathbf{d}||} = \overline{Per} \tag{42}$$

We can adopt the similar strategy to derive that, if $||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}\hat{\mathbf{d}}|| < 1$:

$$\|\widehat{Per}\| \leq \|L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}}\frac{||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}||^2||\hat{\mathbf{d}}||}{1 - ||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}\hat{\mathbf{d}}||} - \hat{\mathbf{d}}(\mathbf{x_b})L^{-1}_{\mathbf{x_a}|\mathbf{x_c}} - \hat{\mathbf{d}}(\mathbf{x_b})\frac{||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}||^2||\hat{\mathbf{d}}||}{1 - ||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}\hat{\mathbf{d}}||} = \widehat{\overline{Per}} \tag{43}$$

Upon Eq. 42, we can derive the sufficient condition to safisfy $||Per||_{op} \leq \overline{Per} < \eta$, which is critical for precise identifiability in Theorem 1:

Lemma 2. *suppose that*

$$||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}||_{op}||\mathbf{d}||_{op} < \eta$$

$$||\mathbf{d}||_{op} < \frac{\eta - ||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}||_{op}||\mathbf{d}||_{op}}{||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}||^2_{op}||\mathbf{d}||(||L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c}}||_{op} + ||\mathbf{d}||_{op}) + ||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}||_{op}(\eta - ||L^{-1}_{\mathbf{x_a}|\mathbf{x_c}}||_{op}||\mathbf{d}||_{op})} \tag{44}$$

Then we can obtain $||Per||_{op} \leq \overline{Per} < \eta$.

B.5 Proof of Proposition 2

Proposition 2 *Suppose Theorem 1 holds, and the following assumptions and regularization conditions are satisfied:*

1. **(Independence)***: For* $n \in [N]$, $\mathbf{z}^n$ *are independent of each other:* $p(\mathbf{z}) = \prod_{n=1}^{N} p(\mathbf{z}^n)$.
2. **(Structural sparsity)***: The map* $\Phi : \mathcal{Z} \to \mathcal{P}$ *is smooth, where* $\mathcal{Z}$ *denotes the support of* $\mathbf{z}$*, and* $\mathcal{P}$ *denotes the space of PDF* $p(\mathbf{x}|\mathbf{z})$ *(* $\mathcal{P} \subset \mathcal{L}^2(\mathcal{X}, \mu)$ *(*μ *denotes the measure)). Let* G *and* $\hat{G}$ *be the binary support matrix of the Jacobian* $J_\Phi(\mathbf{z})$ *and* $J_{\hat{\Phi}}(\hat{\mathbf{z}})$*, respectively, i.e.* $G^{i,j} = 1$ *if* $||\frac{\partial p(\mathbf{x}^i_\mathbf{b}|\mathbf{z})}{\partial \mathbf{z}^j}||_{\mathcal{L}^2} \neq 0$*, and* 0 *for otherwise. For each* $n \in [1, N]$*, there exists a subset of indices* $\mathcal{C}_k$ *satisfying* $\bigcap_{m \in \mathcal{C}_k} G^{m,:} = \{n\}$*;*

Then, $\hat{\mathbf{z}}$ *must correspond component-wise to a permutation of the true latent variables* $\mathbf{z}$*.*

Proof: Theorem 1 guarantees the existence of an invertible trasnformation h such that $\hat{\mathbf{z}} = h(\mathbf{z})$, Our goal is to show that h is a composition of a permutation and component-wise diagonal transformations.

Applying the chain-rule by taking Fréchet derivatives to $\Phi(\mathbf{z}) = \hat{\Phi}(\hat{\mathbf{z}})$, we can obtain:

$$\begin{aligned} &\Phi(\mathbf{z}) = \hat{\Phi}(\hat{\mathbf{z}}) \\ \Rightarrow\ &J_\Phi(\mathbf{z}) = J_{\hat{\Phi}}(\hat{\mathbf{z}}) J_h(\mathbf{z}) \end{aligned} \tag{45}$$

Assumption iii in Theorem 1 assume the injective of Φ. Therefore, Eq. 45 implies

$$\forall \{i, j\} \in \mathcal{G}, \{i, \sigma(j)\} \hat{\mathcal{G}} \tag{46}$$

where σ denotes the permutation transformation in this section.

For simplicity, we denote denote J_h by $\mathbf{H}$. Suppose, for contradiction, that $H(\mathbf{z})$ is not a composition of a diagonal matrix and a permutation matrix, i.e., there exist $j_1 \neq j_2$ such that:

$$\operatorname{supp}(\mathbf{H}^{j_1,:}) \cap \operatorname{supp}(\mathbf{H}^{j_2,:}) \neq \emptyset. \tag{47}$$

Let j_3 be an element in this intersection, so $\sigma(j_3) \in \operatorname{supp}(\mathbf{H}^{j_1,:}) \cap \operatorname{supp}(\mathbf{H}^{j_2,:})$. Without loss of generality, assume $j_3 \neq j_1$. According to Assumption 2, there exists a set $\mathcal{C}_{j_1}$ containing j_1 such that:

$$\bigcap_{i \in \mathcal{C}_{j_1}} G^{i,:} = \{j_1\}. \tag{48}$$

Since $j_3 \neq j_1$, it must be that:

$$j_3 \notin \bigcap_{i \in \mathcal{C}_{j_1}} G^{i,:}, \tag{49}$$

implying there exists some $i_3 \in \mathcal{C}_{j_1}$ such that:

$$j_3 \notin G^{i_3,:}. \tag{50}$$

However, since $j_1 \in G^{i_3,:}$, we have $(i_3, j_1) \in G$. Given $\mathbf{H}$ is a permuted transformation, Eq. 46 implies:

$$(i_3, \sigma(j_3)) \in \hat{G}. \tag{51}$$

However, this means $(i_3, j_3) \in G$, which contradicts Eq. 50. This contradiction implies $\mathbf{H}$ must be a composition of a permutation matrix and a diagonal matrix.

Together with the equation $J_g = J_{\hat{g}}\mathbf{H}$, we achieve the desired result that t is composed of a permutation and component-wise invertible functions.

B.6 Additional Proposition

Assumption 1 in Proposition 2 claims the independence between each dimension of $\mathbf{z}$. To relax such a constraint, we can alternatively allow latent variables $\mathbf{z}$ to exhibit dependence via a known auxiliary domain variable $\mathbf{u}$. In other words, we assume conditional independence across dimensions of $\mathbf{z}$ given $\mathbf{u}$, i.e., $p(\mathbf{z}|\mathbf{u}) = \prod_{n=1}^{N} p(\mathbf{z}^n|\mathbf{u})$. Specifically, we modify the original data-generating process in Eq. 1 to incorporate the domain index $\mathbf{u}$ explicitly:

$$\mathbf{x} = g(\mathbf{z}, \epsilon), \quad \epsilon = e(\mathbf{z}, \mathbf{u}, \eta) \tag{52}$$

where $\mathbf{u}$ denotes the domain index. Under these conditions, we establish the following proposition:

Proposition 3. *Suppose the subspace identifiability condition in Theorem 1 holds. Additionally, assume the following conditions:*

1. **(Conditional independence)***: Latent variables are conditionally independent given domain* $\mathbf{u}$*:* $p(\mathbf{z}|\mathbf{u}) = \prod_{n=1}^{N} p(\mathbf{z}^n|\mathbf{u})$*. Also,* $p(\mathbf{z}|\mathbf{u})$ *is positive and twice differentiable.*
2. **(Sufficient variability)***: There exist* $2N+1$ *distinct domain values* $\mathbf{u} \in [1, 2N+1]$*, such that the* $2N$ *vectors* $\mathbf{w}(\mathbf{z}, \mathbf{u}) - \mathbf{w}(\mathbf{z}, \mathbf{u}_0)$ *(with* $\mathbf{u} \neq \mathbf{u}_0$*) are linearly independent, where the vector* $\mathbf{w}(\mathbf{z}, \mathbf{u})$ *is defined as:* $\mathbf{w}(\mathbf{z}, \mathbf{u}) = \{\mathbf{v}(\mathbf{z}, \mathbf{u}), \mathbf{v}'(\mathbf{z}, \mathbf{u})\}$ *with*

$$\mathbf{v}(\mathbf{z}, \mathbf{u}) = \left(\frac{\partial \log p(\mathbf{z}^1|\mathbf{u})}{\partial \mathbf{z}^1}, \ldots, \frac{\partial \log p(\mathbf{z}^N|\mathbf{u})}{\partial \mathbf{z}^N} \right)$$
$$\mathbf{v}'(\mathbf{z}, \mathbf{u}) = \left(\frac{\partial^2 \log p(\mathbf{z}^1|\mathbf{u})}{(\partial \mathbf{z}^1)^2}, \ldots, \frac{\partial^2 \log p(\mathbf{z}^N|\mathbf{u})}{(\partial \mathbf{z}^N)^2} \right)$$

Then $\{\hat{\mathbf{z}}^{\hat{n}}|\hat{n} \in [1, N]\}$ *must be a component-wise transformation of a permuted version of true* $\{\mathbf{z}^n|n \in [1, N]\}$

Insight: Proposition 3 shows the identifiability results on non-i.i.d. scenario. Domain-dependent changes in $p(\mathbf{z} \mid \mathbf{u})$ provide sufficient variation to distinguish the latent coordinates, while Theorem 1 first removes the ambiguity introduced by the misspecified observation structure and ϵ. Thus, non-i.i.d. variation can replace marginal latent independence as the additional information needed to reduce the remaining invertible ambiguity to component-wise transformations and permutation.

Proof: By Theorem 1 there exists an invertible reparameterization $h : \mathcal{Z} \to \mathcal{Z}$ such that $\hat{\mathbf{z}} = h(\mathbf{z})$ and $\mathbf{z} = h^{-1}(\hat{\mathbf{z}})$. Applying the change-of-variables formula to the conditional densities (for any fixed $\mathbf{u}$) gives:

$$p_{\hat{\mathbf{z}}|\mathbf{u}}(\hat{\mathbf{z}} \mid \mathbf{u}) = p_{\mathbf{z}|\mathbf{u}}\big(h^{-1}(\hat{\mathbf{z}}) \mid \mathbf{u}\big) \big|\det J_{h^{-1}}(\hat{\mathbf{z}})\big|. \tag{53}$$

Taking logarithms yields

$$\log p_{\hat{\mathbf{z}}|\mathbf{u}}(\hat{\mathbf{z}} \mid \mathbf{u}) = \log p_{\mathbf{z}|\mathbf{u}}(\mathbf{z} \mid \mathbf{u}) + \log\big|\det J_{h^{-1}}(\hat{\mathbf{z}})\big|, \tag{54}$$

Under Assumption 1, we have

$$\sum_{i=1}^{n} \log p_{\hat{\mathbf{z}}^i|\mathbf{u}}(\hat{\mathbf{z}}^i \mid \mathbf{u}) = \sum_{i=1}^{n} \log p_{\mathbf{z}^i|\mathbf{u}}(\mathbf{z}^i \mid \mathbf{u}) + \log\big|\det J_{h^{-1}}(\hat{\mathbf{z}})\big|. \tag{55}$$

We then take second derivatives with respect to $\hat{\mathbf{z}}^k$ and $\hat{\mathbf{z}}^v$ for $k \neq v$. Since each term on the left-hand side of Eq. 55 depends only on a single coordinate $\hat{\mathbf{z}}^i$, we have $\partial \log p_{\hat{\mathbf{z}}^i|\mathbf{u}}(\hat{\mathbf{z}}^i \mid \mathbf{u})/\partial \hat{\mathbf{z}}^k = 0$ for $i \neq k$, which implies

$$\frac{\partial^2}{\partial \hat{\mathbf{z}}^k \partial \hat{\mathbf{z}}^v} \sum_{i=1}^{n} \log p_{\hat{\mathbf{z}}^i|\mathbf{u}}(\hat{\mathbf{z}}^i \mid \mathbf{u}) = 0. \tag{56}$$

For the right-hand side, define for $i = 1, \ldots, n$

$$\tilde{h}^{i,(k)} := \frac{\partial \mathbf{z}^i}{\partial \hat{\mathbf{z}}^k}, \qquad \tilde{h}^{i,(k,v)'} := \frac{\partial^2 \mathbf{z}^i}{\partial \hat{\mathbf{z}}^k \partial \hat{\mathbf{z}}^v}, \tag{57}$$

$$\eta_i'(\mathbf{z}^i, \mathbf{u}) := \frac{\partial}{\partial \mathbf{z}^i} \log p_{\mathbf{z}^i|\mathbf{u}}(\mathbf{z}^i \mid \mathbf{u}), \qquad \eta_i''(\mathbf{z}^i, \mathbf{u}) := \frac{\partial^2}{\partial \mathbf{z}^{i2}} \log p_{\mathbf{z}^i|\mathbf{u}}(\mathbf{z}^i \mid \mathbf{u}) \tag{58}$$

A direct application of the chain rule gives

$$\sum_{i=1}^{n}\Big(\eta_i''(\mathbf{z}^i,\mathbf{u})\tilde{h}^{i,(k)}\tilde{h}^{i,(v)}+\eta_i'(\mathbf{z}^i,\mathbf{u})\tilde{h}^{i,(k,v)\prime}\Big)+\frac{\partial^2}{\partial\hat{\mathbf{z}}^k\partial\hat{\mathbf{z}}^v}\log\big|\det J_{h^{-1}}(\hat{\mathbf{z}})\big|=0. \tag{59}$$

Fix (k,v) with $k\neq v$ and evaluate this identity at $2n{+}1$ distinct values of the conditioning variable, $\mathbf{u}^{(j)}$ for $j\in\{0,1,\dots,2n\}$. Subtracting the equation at $\mathbf{u}^{(0)}$ from that at $\mathbf{u}^{(j)}$ cancels the log-determinant term (which does not depend on $\mathbf{u}$) and yields, for $j=1,\dots,2n$,

$$\sum_{i=1}^{n}\Big(\big[\eta_i''(\mathbf{z}^i,\mathbf{u}^{(j)})-\eta_i''(\mathbf{z}^i,\mathbf{u}^{(0)})\big]\tilde{h}^{i,(k)}\tilde{h}^{i,(v)}+\big[\eta_i'(\mathbf{z}^i,\mathbf{u}^{(j)})-\eta_i'(\mathbf{z}^i,\mathbf{u}^{(0)})\big]\tilde{h}^{i,(k,v)\prime}\Big)=0. \tag{60}$$

Let

$$\mathbf{w}(\mathbf{z},\mathbf{u}):=\big(\eta_1''(\mathbf{z}^1,\mathbf{u}),\dots,\eta_n''(\mathbf{z}^n,\mathbf{u}),\eta_1'(\mathbf{z}^1,\mathbf{u}),\dots,\eta_n'(\mathbf{z}^n,\mathbf{u})\big)^\top. \tag{61}$$

Under Assumption 2, the $2n$ vectors $\mathbf{w}(\mathbf{z},\mathbf{u}^{(j)})-\mathbf{w}(\mathbf{z},\mathbf{u}^{(0)})$ for $j=1,\dots,2n$ are linearly independent, so the only solution to the homogeneous linear system Eq. 60 is

$$\tilde{h}^{i,(k)}\tilde{h}^{i,(v)}=0\quad\text{and}\quad\tilde{h}^{i,(k,v)\prime}=0\qquad\text{for all } i\in\{1,\dots,n\}\text{ and all } k\neq v. \tag{62}$$

Hence each row of the Jacobian $J_{h^{-1}}(\hat{\mathbf{z}})$ has at most one nonzero entry, and all mixed second derivatives vanish. Since h^{-1} is invertible, each row must in fact have exactly one nonzero entry; moreover, two distinct rows cannot share the same nonzero column (otherwise $\det J_{h^{-1}}(\hat{\mathbf{z}})=0$), so there exists a permutation π such that

$$\hat{\mathbf{z}}^{\pi(i)}=h^i(\mathbf{z}^i)\qquad\text{for } i=1,\dots,n, \tag{63}$$

which shows that $\hat{\mathbf{z}}$ is obtained from $\mathbf{z}$ by a permutation of component-wise invertible transformations.

Assumption Intuitions:
• Assumption 1 relaxes the marginal independence of $\mathbf{z}$, which allows that $\mathbf{z}$ can be statistically dependent.
• Assumption 2 requires the domain variable to induce sufficiently rich changes in the first- and second-order derivatives of the latent conditional log-densities. These variations provide enough constraints to distinguish different latent variable.

C Synthetic Experiment Details

C.1 Synthetic data-generating process

We construct a synthetic data-generating process that realizes the latent-dependent noise and misspecified conditional structure considered in this work.

Let

$$\mathbf{z}=(\mathbf{z}^1,\mathbf{z}^2,\mathbf{z}^3)^\top,\qquad\mathbf{z}\sim\mathcal{N}(0,I_3), \tag{64}$$

so that the latent coordinates are mutually independent. We independently sample

$$\eta^0,\eta^{\mathbf{a}},\eta^{\mathbf{b}},\eta^{\mathbf{c}}\overset{\text{ind.}}{\sim}\mathcal{N}(0,I_3), \tag{65}$$

where η^0 is shared across the three observed blocks and $\eta^{\mathbf{a}},\eta^{\mathbf{b}},\eta^{\mathbf{c}}$ are block-specific exogenous noises.

For each $v\in\{\mathbf{a},\mathbf{b},\mathbf{c}\}$, define

$$\begin{aligned}\epsilon^v&=C^v\mathbf{z}+\tau\eta^0+s_v\eta^v,\qquad\tau>0,\quad s_v>0,\\ \mathbf{x}_v&=\Psi(B^v\mathbf{z}+\epsilon^v)=\Psi(M^v\mathbf{z}+\tau\eta^0+s_v\eta^v),\qquad M^v:=B^v+C^v,\end{aligned} \tag{66}$$

where where the existence of η^0 guarantees the dependence of $\mathbf{x}^v$ conditioning on $\mathbf{z}$, Ψ is applied coordinate-wise and is defined by

$$\Psi(t)=t+\alpha\tanh(t),\qquad 0<\alpha<1. \tag{67}$$

Since

$$\Psi'(t) = 1 + \alpha \operatorname{sech}^2(t) \in [1, 1 + \alpha], \tag{68}$$

Ψ is a smooth bijection with a smooth inverse.

We choose

$$M^{\mathbf{a}} = M^{\mathbf{c}} = I_3, \qquad M^{\mathbf{b}} = \begin{pmatrix} 1 & r & 0 \\ r & 0 & 1 \\ 0 & r & 1 \end{pmatrix}, \qquad 0 < r < 1. \tag{69}$$

Since

$$\det(M^{\mathbf{b}}) = -r(1 + r) \neq 0, \tag{70}$$

$M^{\mathbf{b}}$ is invertible. We further set

$$B^v = C^v = \frac{1}{2} M^v, \tag{71}$$

so that

$$\epsilon^v = \frac{1}{2} M^v \mathbf{z} + \tau \eta^0 + s_v \eta^v. \tag{72}$$

Hence the complete noise variable $\epsilon = (\epsilon^{\mathbf{a}}, \epsilon^{\mathbf{b}}, \epsilon^{\mathbf{c}})$ explicitly depends on $\mathbf{z}$, as required by Eq. equation 1. The resulting observation

$$\mathbf{x} = (\mathbf{x_a}, \mathbf{x_b}, \mathbf{x_c}) \in \mathbb{R}^9 \tag{73}$$

contains three disjoint three-dimensional blocks.

In practice, we sample We sample 10,000 points from the data synthesizing process above mentioned. We leverage 90% among them for training, and the rest 10% for test.

C.2 Assumption Justifications:

Below we justify if all assumptions from Theorem 1 and Proposition 2 hold for our data synthesizing:

Misspecified conditional structure. Conditioning on $\mathbf{z}$ fixes all deterministic terms involving $\mathbf{z}$, while the three blocks continue to share η^0. Let

$$\mathbf{y}_v := M^v \mathbf{z} + \tau \eta^0 + s_v \eta^v. \tag{74}$$

Then, for distinct $v, v' \in \{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$,

$$\operatorname{Cov}(\mathbf{y}_v, \mathbf{y}_{v'} \mid \mathbf{z}) = \tau^2 I_3. \tag{75}$$

Therefore, for every $\tau > 0$,

$$p(\mathbf{x_a}, \mathbf{x_b}, \mathbf{x_c} \mid \mathbf{z}) \neq p(\mathbf{x_a} \mid \mathbf{z}) p(\mathbf{x_b} \mid \mathbf{z}) p(\mathbf{x_c} \mid \mathbf{z}). \tag{76}$$

Because Ψ is invertible, the conditional dependence is preserved after the transformation. The conditionally independent setting is recovered when $\tau = 0$.

Constructing both the precise and approximate identifiability regime. The shared-noise strength τ directly controls the amount of structural misspecification. Let $\mathbf{d}_\tau(\mathbf{x_b})$, $\mathbf{d}_\tau$, and Per_τ denote the corresponding difference and perturbation operators generated with τ in Eq. 66.

For the Gaussian construction in Eq. 66, the conditional covariance induced by the shared disturbance is $\tau^2 I_3$. Hence the corresponding conditional densities vary smoothly with τ^2. On the operator spaces considered here, there exist finite constants $C_d, C_b > 0$ and let $\tau_{\max} > 0$ such that, for $0 < \tau \leq \tau_{\max}$,

$$\begin{aligned} \|\mathbf{d}_\tau\|_{\mathrm{op}} &< C_d \tau^2, \\ \operatorname*{ess\,sup}_{\mathbf{x_b}} \|\mathbf{d}_\tau(\mathbf{x_b})\|_{\mathrm{op}} &< C_b \tau^2. \end{aligned} \tag{77}$$

Let

$$M_0 := \sup_{0<\tau\le\tau_{\max}} \|L^{-1}_{\mathbf{x_a}|\mathbf{x_c},\tau}\|_{\mathrm{op}},$$
$$M_1 := \sup_{0<\tau\le\tau_{\max}} \|L_{\mathbf{x_a},\mathbf{x_b}|\mathbf{x_c},\tau}\|_{\mathrm{op}}, \tag{78}$$

which are finite under Assumptions 1 and 2. Using Lemma 1, whenever $M_0 C_d \tau^2 < 1$,

$$\|Per_\tau\|_{\mathrm{op}} \le \left(M_1 + C_b\tau^2\right) \frac{M_0^2 C_d \tau^2}{1 - M_0 C_d \tau^2} + M_0 C_b \tau^2$$
$$=: \overline{Per}. \tag{79}$$

Let

$$\eta = \frac{\delta}{2\kappa} - \alpha > 0 \tag{80}$$

be defined as in Section 3.1. By continuity of $\overline{Per}(\tau)$ at $\tau = 0$, there exists $\tau_0 > 0$ such that

$$0 < \tau < \tau_0 \quad \Longrightarrow \quad 0 < \overline{Per}(\tau) < \eta. \tag{81}$$

We choose the shared-noise strength $\tau \in (0, \tau_0)$ in the precise-identifiability experiments. Thus, the synthetic data simultaneously exhibit structural misspecification ($\tau > 0$) and satisfy the perturbation condition required by Proposition 1.

To evaluate approximate identifiability, we use the same data-generating process but increase τ to generate stronger structural misspecification. We select values of τ for which

$$\overline{Per}(\tau) \ge \eta, \tag{82}$$

so that the sufficient perturbation condition for precise identifiability is no longer satisfied. All remaining assumptions of Theorem 2 are kept unchanged. Therefore, the same synthetic model produces both regimes: small nonzero τ yields precise identifiability, whereas larger perturbations are used to evaluate the approximate-identifiability guarantee.

Bounded densities. Before applying Ψ, the joint vector $(\mathbf{z}, \mathbf{y_a}, \mathbf{y_b}, \mathbf{y_c})$ is a nondegenerate multivariate Gaussian because $s_v > 0$ for every block. Hence its joint, marginal, and conditional densities are bounded. Since Ψ' is uniformly bounded above and bounded away from zero, the change-of-variables formula preserves boundedness of the corresponding densities after transformation. Thus the bounded-density regularity condition in Assumption 1 is satisfied.

Injectivity of Θ. For the $\mathbf{x_b}$ block,

$$\Psi^{-1}(\mathbf{x_b}) \mid \mathbf{z} \sim \mathcal{N}\left(M^{\mathbf{b}}\mathbf{z}, (\tau^2 + s_{\mathbf{b}}^2) I_3\right). \tag{83}$$

Suppose

$$p(\cdot \mid \mathbf{z}) = p(\cdot \mid \mathbf{z}'). \tag{84}$$

Equation equation 83 then implies

$$M^{\mathbf{b}}\mathbf{z} = M^{\mathbf{b}}\mathbf{z}'. \tag{85}$$

Since $M^{\mathbf{b}}$ is invertible,

$$\mathbf{z} = \mathbf{z}'. \tag{86}$$

Therefore the conditional-distribution map

$$\Theta : \mathcal{Z} \to \mathcal{D}, \qquad \Theta(\mathbf{z}) = p(\mathbf{x_b} \mid \mathbf{z}), \tag{87}$$

is injective.

Structural sparsity for Proposition 2. For each observed coordinate $\mathbf{x}^j$, define

$$\Phi^j(\mathbf{z}) := p(\mathbf{x}^j \mid \mathbf{z}). \tag{88}$$

Since the conditional mean of each pre-transformed coordinate depends on $\mathbf{z}$ through the corresponding row of M^v, the support of $\partial\Phi^j/\partial\mathbf{z}^i$ coincides with the support of that row. Ordering the observations as

$$(\mathbf{x}^1_\mathbf{a}, \mathbf{x}^2_\mathbf{a}, \mathbf{x}^3_\mathbf{a}, \mathbf{x}^1_\mathbf{b}, \mathbf{x}^2_\mathbf{b}, \mathbf{x}^3_\mathbf{b}, \mathbf{x}^1_\mathbf{c}, \mathbf{x}^2_\mathbf{c}, \mathbf{x}^3_\mathbf{c}), \tag{89}$$

the structural-support matrix is

$$G = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{90}$$

Thus the structural-sparsity condition of Proposition 2 is satisfied. Furthermore,

$$p(\mathbf{z}) = \prod_{i=1}^{3} p(\mathbf{z}^i) \tag{91}$$

by construction, satisfying its latent-independence condition.

C.3 Implementations & Training Details.

In our synthetic experiments, we set the dimensions $N = 3$ and $K = 9$. The encoders, decoder, and normalizing flow modules were each implemented using single-layer multilayer perceptrons (MLPs) followed by Leaky ReLU activations.

Our implementation utilized PyTorch 1.11.0. For optimization, we adopted the AdamW optimizer Loshchilov & Hutter (2019), which is known for enhancing generalization in deep learning models. The hyperparameters were configured as follows: a learning rate of 1×10^{-3} and a batch size of 64. To guarantee robustness and statistical reliability, each model was trained using 10 different random seeds. We report the overall performance as the mean $\pm$ standard deviation computed across these runs. The loss function employed balances the reconstruction error and the KL-divergence, with weighting coefficients set to $\beta_1 = \beta_2 = 0.02$. All experiments were performed on a single NVIDIA GeForce RTX 2080 Ti GPU equipped with 11GB of memory.

C.4 The objective of IndVAE

In this section, we explain our trained objective for IndVAE, which is designed by taking inspirations from (Hu & Schennach, 2008) To incorporate with the conditional independence assumption, we consider the k-th observed variable $\mathbf{x}^k$ is generated as $\mathbf{x}^k = g^k(\mathbf{z}, \epsilon)$. Accordingly, the log-likelihood of the data generating process of Eq. 1 can be transformed as follows:

$$\begin{aligned} \log p(\mathbf{z}, \epsilon, \mathbf{x}) &= \log p_\theta(\mathbf{x}|\mathbf{z}, \epsilon) + \log p_\gamma(\epsilon|\mathbf{z}) + \log p_\delta(\mathbf{z}) \\ &= \sum_{k=1}^{K} \log p_\theta(\mathbf{x}^k|\mathbf{z}, \epsilon) + \log p_\gamma(\epsilon|\mathbf{z}) + \log p_\delta(\mathbf{z}) \end{aligned} \tag{92}$$

Accordingly, the loss function becomes:

$$\begin{aligned} \mathcal{L}_{\text{ELBO}} =& \mathbb{E}_{\hat{\mathbf{z}}\sim q_\psi, \hat{\epsilon}\sim q_\phi} \left[\sum_{k=1}^{K} \log p_\theta(\hat{\mathbf{x}}^k|\hat{\mathbf{z}}, \hat{\epsilon}) \right] \\ &- \beta_1 \mathbb{E}_{\hat{\mathbf{z}}\sim q_\psi} \big(\log q\left(\hat{\mathbf{z}}|\mathbf{x}\right) - \log p_\delta(\mathbf{z}) \big) - \beta_2 \mathbb{E}_{\hat{\mathbf{z}}\sim q_\psi, \hat{\epsilon}\sim q_\phi} \big(\log q\left(\hat{\epsilon}|\mathbf{x}\right) - \log p_\gamma(\hat{\epsilon}|\hat{\mathbf{z}}) \big) \end{aligned} \tag{93}$$

Table 1: The details of our network architectures for the experiment on MSTM-17 dataset, where BS means batch size, $N = 32$, $M = 32$ and $K = 1280$.

Configuration	Description	Output dimensions
Encoder $q\psi$		
Input: $\mathbf{x}$		BS $\times K$
Dense	128 neurons, LeakyReLU	BS $\times$ 128
Dense	128 neurons, LeakyReLU	BS $\times$ 128
Dense	Output embeddings	BS $\times 2N$
Bottleneck	Compute mean and variance of posterior	$\mu_{\mathbf{z}}$, $\sigma_{\mathbf{z}}$
Reparameterization	Sequential sampling	$\hat{\mathbf{z}}$
Encoder q_ϕ		
Input: $\mathbf{x}$		BS $\times K$
Dense	128 neurons, LeakyReLU	BS $\times$ 128
Dense	128 neurons, LeakyReLU	BS $\times$ 128
Dense	Output embeddings	BS $\times 2M$
Bottleneck	Compute mean and variance of posterior	μ_ϵ, σ_ϵ
Reparameterization	Sequential sampling	$\hat{\epsilon}$
Decoder		
Input: $\hat{\mathbf{z}}, \hat{\epsilon}$		BS $\times (N + M)$
Dense	128 neurons, LeakyReLU	BS $\times$ 128
Dense	128 neurons, LeakyReLU	BS $\times$ 128
Dense	input embeddings	BS $\times K$
Prior module		
Input	$\hat{\mathbf{z}}, \hat{\epsilon}$	BS $\times (N + M)$
InverseTransformation	$\hat{\eta}$	BS $\times M$
JacobianCompute	$\log \lvert \det J_{\hat{\epsilon}} \rvert$	BS
Classifier		
Input: $\hat{\mathbf{z}}$		BS $\times N$
Dense	256 neurons, LeakyReLU	BS $\times$ 256
Dense	256 neurons, LeakyReLU	BS $\times$ 256
Dense	output one-hot embeddings	BS $\times$ 4101

D Additional Experiments

D.1 Real-world Experiment for Justifying Proposition 2

Task Setup: To validate our proposed identifiability theories in realistic and complex scenarios, we apply them to the task of Person Classification. In Person Classification, the goal is to assign a unique identity index to each individual, based on input images. This setup aligns well with our generalized dependency structure setting, as each image of an individual inherently contains noise, such as varying poses, gaits, or clothes, making it challenging to disentangle these factors from the underlying identity. Also, different body parts cannot be independent conditioning upon the latent person identify index. Consequently, this task serves as a solid playground for evaluating the robustness and efficacy of our theoretical framework in addressing real-world complexities.

In our implementation, we first employ a pretrained feature extractor to derive feature representations of each individual person, denoted as $\mathbf{x} \in \mathbb{R}^K$. We consider $\mathbf{x}$ are generated by the latent variable $\mathbf{z} \in \mathbb{R}^N$, directly associated with each person's identity index, along with a dependent noise variable $\epsilon \in \mathbb{R}^M$, capturing variations such as pose, gait, or clothes. Inspired by the two-phase training pipeline proposed by (Li et al., 2024b; 2025a), we adapt our approach to Person Classification task as follows. Figure 5 suggests the data generating process of the Person Classification task. First, we train our approach by optimizing the objective function detailed in Eq. 16. Subsequently, we introduce a classifier $\hat{c}$, implemented by a multilayer perceptron (MLP), to predict the one-hot encoded index label $\hat{y}$ from the inferred latent representation $\hat{\mathbf{z}}$: $\hat{y} = \text{MLP}(\hat{\mathbf{z}})$ The classifier

is optimized using a cross-entropy loss given by: $\mathcal{L}_{\text{cls}}^{\text{CE}} = -\mathbb{E}_{\hat{y}}\left[\text{one-hot}(y) \cdot \log(\text{softmax}(\hat{y}))\right]$ where one-hot(y) denotes the one-hot embedding of the true person index label.

Data and Comparing Approaches We conduct our experiments on the MSMT17 dataset (Wei et al., 2018), which comprises images of 4,101 unique individuals. Each individual in the dataset has more than 10 images, resulting in a total of over 120,000 images. We partition the dataset into three parts: 60% for training, 20% for validation, and the remaining 20% for test.

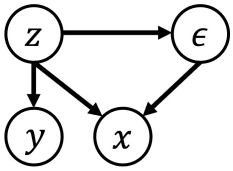


Figure 5: Visualization of the data generating process for Person Classification task

For performance comparison, we select several state-of-the-art methods on the task of Person Classification,including GTL (Yang et al., 2025), AGW (Ye et al., 2021), TransReID (He et al., 2021), and CLIPReID (Li et al., 2023). We also benchmark against MCRL (Sun et al., 2025) and IndVAE (Hu & Schennach, 2008) to evaluate the efficacy of our identifiability results under generalized dependency structure.

Table 2: Comparison of Top-1 Accuracy on MSMT17 dataset

Methods	Acc
AGW (Ye et al., 2021)	85.5 ± 1.2
TransReID (He et al., 2021)	87.8 ± 0.5
CLIPReID (Li et al., 2023)	90.1 ± 0.3
GTL Yang et al. (2025)	91.5 ± 1.2
MCRL (Sun et al., 2025)	92.6 ± 0.9
IndVAE (Hu & Schennach, 2008)	$\underline{93.1 \pm 0.5}$
Ours	$\mathbf{96.2 \pm 0.6}$

Results & Discussions: Table 2 reports the comparison of Top-1 Accuracy (Acc) among state-of-the-art methods on the MSMT17 dataset. Our method achieves a superior performance, substantially surpassing approaches that do not explicitly handle dependent noise, such as IndVAE (Hu & Schennach, 2008) and MCRL (Sun et al., 2025). Furthermore, our proposed method demonstrates notable improvements over leading methods including GTL (Yang et al., 2025), CLIPReID (Li et al., 2023), TransReID (He et al., 2021), and AGW (Ye et al., 2021), outperforming them by significant margins. These results clearly highlight the effectiveness and robustness of our proposed framework in accurately addressing generalized dependency structures in complex real-world scenarios.

Implementations:

To obtain a fair comparison, we adopt the approach outlined by (Yang et al., 2025), employing the pretrained CLIP model (Radford et al., 2021b) as the visual encoder to generate 1280-dimensional representations for $\mathbf{x}$. Table 1 summarizes the specific network architectures implemented for our experiments on the real-world MSTM17 dataset.

To train our framework, we utilize the AdamW optimizer combined with a cosine annealing learning rate schedule. The initial learning rate is set to 2×10^{-3}, with a weight decay parameter of 1×10^{-2} to prevent overfitting. The ELBO loss function incorporates equal weighting coefficients $\beta_1 = \beta_2 = 0.02$. We use a batch size of 128, chosen to balance computational efficiency with optimization stability. The framework is implemented in PyTorch. Training is done for 80 epochs on a multi-GPU configuration comprising four NVIDIA GeForce RTX 2080 Ti GPUs, collectively providing 44GB of memory.

D.2 Additional Ablation Studies

Ablations for higher dimensional z

we additionally evaluated the scalability of our method to higher latent dimensions on the synthetic dataset by varying the latent dimensionality $N \in \{8, 12, 18\}$, while keeping the network architecture, training protocol, and all other hyperparameters fixed. Table 3 reports the MCC and compares with IndVAE Hu & Schennach (2008): Even as the latent dimension increases, our method consis-

N	IndVAE	Ours
8	0.64 ± 0.06	0.80 ± 0.02
12	0.51 ± 0.03	0.68 ± 0.05
18	0.47 ± 0.04	0.61 ± 0.05

Table 3: MCC on the synthetic dataset for increasing latent dimensionality N.

Table 4: Sensitivity of MCC to regularization hyperparameters.

Hyperparameter	Value	MCC
λ	1	0.87 ± 0.04
	0.1	0.82 ± 0.03
	0.01	0.70 ± 0.07
	10	0.68 ± 0.05
β_1	0.02	0.87 ± 0.04
	1	0.73 ± 0.02
	0.001	0.65 ± 0.04
β_2	0.02	0.87 ± 0.04
	1	0.82 ± 0.01
	0.001	0.78 ± 0.06

tently achieves higher MCC than IndVAE, indicating that the Jacobian-based sparsity regularization remains effective and that our approach scales well to higher-dimensional latent spaces within the considered regime.

Ablations for hyperparameter sensitivity

In our implementation on the synthetic dataset, we set the weight of the sparsity regularizer to 1 and the KL weights to $\beta_1 = \beta_2 = 0.02$. In this section, we conducted a sensitivity analysis in which we vary *one* of these three hyperparameters at a time while keeping the others fixed at their default values. For all runs we use the same network architecture, batch size, number of epochs, learning rate, and training protocol as in the main experiments. The numbers reported below are MCC scores (mean $\pm$ std. over multiple runs) on the synthetic dataset. These results show the default setting $(\lambda, \beta_1, \beta_2) = (1, 0.02, 0.02)$ obtains the best results.

Additional real-world experiments

we expand our real-world evaluation beyond the original dataset and consider two additional person identity classification benchmarks. Specifically, we use the SYSU-MM01 dataset Wu et al. (2017) and the ROBOTPKU dataset Liu et al. (2017). We only use the RGB modality, since our focus is not on cross-modal person re-identification. SYSU-MM01 contains RGB images of 491 identities from 6 cameras, with a total of 30,071 images. ROBOTPKU contains more than 16,000 RGB images of 180 identities, captured under dynamic robotic viewpoints. These datasets thus provide solid playgrounds for our experiments.

For performance comparison, we follow the same Person Classification protocol as in our main experiment and compare against several state-of-the-art methods, including GTL (Yang et al., 2025), AGW (Ye et al., 2021), TransReID (He et al., 2021), CLIPReID (Li et al., 2023), LDP-net Zhou et al. (2023), Style Fu et al. (2023), as well as MCRL (Sun et al., 2025) and IndVAE (Hu & Schennach, 2008). Tables 5 and 6 report the Top-1 classification accuracy (mean $\pm$ std. over multiple runs).

To further examine the effect of the architecture choice, we replace the MLP in our framework on the MSTM-17 dataset with a single-layer Gated Recurrent Unit (GRU) Goodfellow et al. (2016) using the same hidden dimension (the classifier architecture and all training and evaluation protocols remain unchanged, and we set $N = M = 32$ for fair comparison). The resulting Top-1 accuracies are: GRU: 96.9 ± 0.4 versus our original MLP-based model: 96.2 ± 0.6. The GRU improves the classification results slightly against the MLP architecture. Overall, our method consistently outperforms strong baselines across three real-world person identity datasets.

Table 5: Comparison of Top-1 Accuracy on the ROBOTPKU dataset.

Methods	Acc
AGW (Ye et al., 2021)	87.6 ± 0.8
TransReID (He et al., 2021)	90.2 ± 0.9
CLIPReID (Li et al., 2023)	91.7 ± 1.1
GTL Yang et al. (2025)	93.9 ± 0.5
MCRL (Sun et al., 2025)	94.5 ± 1.0
IndVAE (Hu & Schennach, 2008)	$\underline{95.8 \pm 0.8}$
Ours	98.4 ± 0.5

Table 6: Comparison of Top-1 Accuracy on the SYSU-MM01 dataset.

Methods	Acc
LDP-net Zhou et al. (2023)	91.7 ± 1.1
Style Fu et al. (2023)	92.8 ± 0.8
CLIPReID (Li et al., 2023)	94.1 ± 1.0
GTL Yang et al. (2025)	95.7 ± 0.4
MCRL (Sun et al., 2025)	96.4 ± 0.8
IndVAE (Hu & Schennach, 2008)	$\underline{96.8 \pm 0.5}$
Ours	99.1 ± 0.5

D.3 ADDITIONAL REAL-WORLD EXPERIMENTS FOR PROPOSITION 3

Training Objective: Figure 6 visualizes the data-generating process described in Eq. 52. Accordingly, the likelihood for this process, given the known auxiliary variable $\mathbf{u}$, is expressed as:

$$p(\mathbf{z}, \epsilon, \mathbf{x}|\mathbf{u}) = p_\theta(\mathbf{x}|\mathbf{z}, \epsilon)p_\gamma(\epsilon|\mathbf{z}, \mathbf{u})p_\delta(\mathbf{z}|\mathbf{u}) \tag{94}$$

As a result, we redesign the encoder and prior module to learn the distribution $p_\gamma(\epsilon|\mathbf{z}, \mathbf{u})$ and $p_\delta(\mathbf{z}|\mathbf{u})$, as shown in Eq. 94, while keeping the decoder unchanged. Accordingly, the ELBO is:

$$\begin{aligned}\mathcal{L}_{\text{ELBO}} =& \mathbb{E}_{\hat{\mathbf{z}}\sim q_\psi, \hat{\epsilon}\sim q_\phi}\left[\log p_\theta(\hat{\mathbf{x}}^k|\hat{\mathbf{z}}, \hat{\epsilon}^k)\right] \\ &- \beta_1 \mathbb{E}_{\hat{\mathbf{z}}\sim q_\psi}\big(\log q\left(\hat{\mathbf{z}}|\mathbf{x}\right) - \log p_\delta(\mathbf{z}|\mathbf{u})\big) \\ &- \beta_2 \mathbb{E}_{\hat{\mathbf{z}}\sim q_\psi, \hat{\epsilon}\sim q_\phi}\big(\log q\left(\hat{\epsilon}|\mathbf{x}\right) - \log p_\gamma(\hat{\epsilon}|\hat{\mathbf{z}}, \mathbf{u})\big)\end{aligned} \tag{95}$$

We further evaluate our method on two standard domain adaptation benchmarks, Office-Home (Venkateswara et al., 2017) and VisDA-2017 (Peng et al., 2018). As shown in Table 7, our method achieves the best performance on all 12 transfer directions of Office-Home. In particular, it improves over UniMoS by 4.6 percentage points on $A \to C$ and 3.3 points on $A \to P$. Averaged over all transfer directions, our method achieves an accuracy of 80.17%, compared with 77.88% for the strongest competing method, UniMoS.

Table 8 shows a similar trend on VisDA-2017. Our method achieves the best performance on 10 of the 12 categories and ties for the best result on Motorcycle. Notable improvements are observed

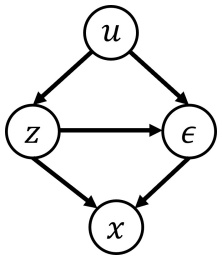


Figure 6: Visualization of the data generations of Eq. 52.

Table 7: Results on Office-Home.

Method	$A \to C$	$A \to P$	$A \to R$	$C \to A$	$C \to P$	$C \to R$	$P \to A$	$P \to C$	$P \to R$	$R \to A$	$R \to C$	$R \to P$
CLIP (Radford et al., 2021a)	51.7	81.5	82.3	71.7	81.5	82.3	71.7	51.7	82.3	71.7	51.7	81.5
DAPrompt (Ge et al., 2023)	54.1	84.3	84.8	74.4	83.7	85.0	74.5	54.6	84.8	75.2	54.7	83.8
ADCLIP (Singha et al., 2023)	55.4	85.2	85.6	76.1	85.8	86.2	76.7	56.1	85.4	76.8	56.1	85.5
UniMoS (Li et al., 2024a)	59.5	89.4	86.9	75.2	89.6	86.8	75.4	58.4	87.2	76.9	59.5	89.7
Ours	**64.1**	**92.7**	**88.5**	**79.2**	**91.0**	**89.5**	**77.9**	**60.2**	**89.6**	**78.4**	**59.9**	**91.0**

Table 8: Results on VisDA-2017.

Method	Plane	Bicycle	Bus	Car	Horse	Knife	Motorcycle	Person	Plant	Skateboard	Train	Truck
CLIP (Radford et al., 2021a)	98.2	83.9	90.5	73.5	97.2	84.0	95.3	65.7	79.4	89.9	91.8	63.3
DAPrompt (Ge et al., 2023)	97.8	83.1	88.8	77.9	97.4	91.5	94.2	79.7	88.6	89.3	92.5	62.0
ADCLIP (Singha et al., 2023)	98.1	83.6	91.2	76.6	98.1	93.4	96.0	81.4	86.4	91.5	92.1	64.2
UniMoS (Li et al., 2024a)	97.7	88.2	90.1	74.6	96.8	95.8	92.4	84.1	90.8	89.0	91.8	65.3
Ours	**98.5**	**90.1**	**93.7**	**76.8**	**98.7**	**97.4**	**96.0**	**87.2**	**92.0**	**93.4**	**94.2**	**67.5**

on Person (87.2%), Bus (93.7%), and Truck (67.5%). The average accuracy reaches 90.46%, outperforming UniMoS at 88.05%. The only category on which our method does not achieve the best result is Car, where DAPrompt obtains 77.9% compared with our 76.8%.

Overall, the consistent gains across different source–target pairs and object categories indicate that the learned representation transfers more effectively across domains. Together with the person-index classification results, these findings are consistent with our hypothesis that explicitly separating the latent representation $\mathbf{z}$ from the dependent noise ϵ improves not only in-domain discrimination, but also representation robustness and transferability.

E Conclusion & Limitations

We study identifiability under a misspecified structure where observations may remain dependent given the latent variables and the noise may itself depend on the latents. We characterize structural misspecification as a perturbed factor analysis problem. Under controlled perturbation and spectral separation, we establish precise subspace identifiability and further obtain component-wise identifiability under structural sparsity. Beyond this precise regime, we establish approximate subspace identifiability by bounding the identification error of the recovered representation from its admissible equivalence class. We further develop an unsupervised variational estimator and evaluate the resulting framework on synthetic and real-world tasks.

A limitation of the current theory is that the approximate guarantee is restricted to the subspace level. Extending it to approximate component-wise identifiability requires additional information to understand the component-level ambiguity under perturbation, beyond the conditions used for approximate subspace recovery. Characterizing the minimal structural assumptions or auxiliary information sufficient for such a guarantee is an important direction for future work.