CAUSAL REPRESENTATION LEARNING FROM MULTIMODAL MEDICAL OBSERVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Prevalent in biological applications (e.g., human phenotype measurements), multimodal datasets can provide valuable insights into the underlying biological mechanisms. However, current machine learning models designed to analyze such datasets still lack interpretability and theoretical guarantees, which are essential to biological applications. Recent advances in causal representation learning have shown promise in uncovering the interpretable latent causal variables with formal theoretical certificates. Unfortunately, existing works for multimodal distributions either rely on restrictive parametric assumptions or provide rather coarse identification results, limiting their applicability to biological research which favors a detailed understanding of the mechanisms.

In this work, we aim to develop flexible identification conditions for multimodal data and principled methods to facilitate the understanding of biological datasets. Theoretically, we consider a flexible nonparametric latent distribution (c.f., parametric assumptions in prior work) permitting causal relationships across potentially different modalities. We establish identifiability guarantees for each latent component, extending the subspace identification results from prior work. Our key theoretical ingredient is the structural sparsity of the causal connections among distinct modalities, which, as we will discuss, is natural for a large collection of biological systems. Empirically, we propose a practical framework to instantiate our theoretical insights. We demonstrate the effectiveness of our approach through extensive experiments on both numerical and synthetic datasets. Results on a real-world human phenotype dataset are consistent with established medical research, validating our theoretical and methodological framework.

1 INTRODUCTION

Multimodal datasets provide rich and comprehensive insights into complex biological systems, holding potential in providing a deeper understanding of biological mechanisms. For example, the human phenotype dataset (Levine et al., 2024) includes measurements from multiple modalities, such as anthropometrics, sleep monitoring, and genetics. Proper analysis of such data can potentially uncover the underlying mechanisms that drive phenotypic diversity and disease susceptibility, such as the discovery of novel molecular markers and the development of predictive models for disease. Recent advances in large-scale models have made it possible to exploit large biological datasets for various tasks such as protein structure prediction (Jumper et al., 2021; Lin et al., 2023), gene-disease association identification (Diaz Gonzalez et al., 2023; Zagirova et al., 2023), and novel drug candidate discovery (Pal et al., 2023; Zheng et al., 2024b).

Despite the impressive performance of these models, their trustworthiness remains a matter of debate (Zheng et al., 2023). A major concern is the lack of interpretability, which poses serious challenges in biological research, limiting the safe and ethical application of these models. For example, in clinical decision-making (Hager et al., 2024), if the model recommends a specific treatment plan for a patient based on genomic data, clinicians need to understand the rationale behind the model's recommendation. Without such transparency, it is difficult to trust the model's results and integrate these systems into critical decision-making processes. While many explainable models have been developed for multimodal datasets (Tang et al., 2023), this aspect is still largely under-explored.

Fortunately, recent advances in causal representation learning (CRL) (Schölkopf et al., 2021) have shown promise in identifying latent causal structures from raw observations, which is well-suited

for biological applications. For example, a plethora of CRL works (Hyvarinen et al., 2019; Khemakhem et al., 2020a; Zhang et al., 2024b; Buchholz et al., 2024; von Kügelgen et al., 2023; Zhang et al., 2024a; Li et al., 2024; Ahuja et al., 2023) can naturally utilize the temporal information or domain indices in fMRI data for identifying the latent causal model. Recently, a line of CRL works has arisen to investigate multimodal distributions (Yao et al., 2023; Morioka & Hyvarinen, 2023; 2024; Daunhawer et al., 2023; Sturma et al., 2023; Gresele et al., 2020). Leveraging the shared information over modalities, these works have developed identifiability guarantees for latent variables despite potentially complex nonlinear causal relations (Yao et al., 2023; Morioka & Hyvarinen, 2024; Daunhawer et al., 2023). Nevertheless, some aspects of these works are still limited. For instance, Von Kügelgen et al. (2021); Daunhawer et al. (2023); Yao et al. (2023) only identify latent subspaces that are directly shared by multiple modalities. However, in practice, many informative latent variables may influence the multiple modalities indirectly through intermediate latent variables. Moreover, such subspace identifiability loses track of the intricate causal influences among individual causal components, resulting in a limited view of the latent mechanism. Morioka & Hyvarinen (2024); Gresele et al. (2020); Morioka & Hyvarinen (2023) relies on specific forms of latent variable distributions (e.g., independence or exponential family). These constraints restrict their applicability for biological datasets that involve complex interactions among latent factors.

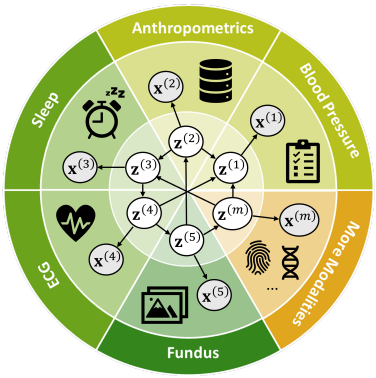


Figure 1: Multimodal data with causal latent variables.

In this work, we aim to develop identification theory with *multimodal biological* datasets in mind, and design *principled and interpretable* models to facilitate analyzing such datasets. We assume that observations $\mathbf{x}^{(m)}$ in any modality m are generated by a specific set of latent components $\{z_i^{(m)}\}_i$ and permit flexible causal relations among latent components from potentially distinct modalities $z_i^{(m)} \to z_j^{(n)}$ for $m \neq n$, $i \neq j$ (Figure 1). **Theoretically**, we provide identifiability guarantees for each latent component $z_i^{(m)}$, thus generalizing the subspace identification results in Yao et al. (2023); Daunhawer et al. (2023) while avoiding independence or parametric assumptions on the latent distribution $p(\{\mathbf{z}^{(m)}\}_m)$ as in Morioka & Hyvarinen (2023; 2024); Gresele et al. (2020). In particular, we first show that any latent subspace $\mathbf{z}^{(m)}$ can be identified as long as $\mathbf{z}^{(m)}$ exerts sufficient influences on other modalities, which is weaker than assuming $\mathbf{z}^{(m)}$ is directly shared over multiple modalities as in Daunhawer et al. (2023); Yao et al. (2023). Based on this subspace identification, we leverage the sparsity of the causal connections among modalities to further identify each latent component $\{z_i^{(m)}\}_i$. This notion of causal sparsity has been explored in recent work (Lachapelle et al., 2023; Xu et al., 2024; Zheng et al., 2022) in other causal identification settings and has been shown realistic in many biological systems (Busiello et al., 2017; Milo et al., 2002; Babu et al., 2004; West et al., 2002; Banavar et al., 1999) as we will discuss in Section 4.

Empirically, we develop a theoretically grounded estimation framework to recover the latent components in each modality. Our model implements our theoretical conditions (in particular, conditional independence and sparsity constraints) on top of normalizing flows (Huang et al., 2018; Dinh et al., 2016) and variational auto-encoders (VAE) (Kingma & Welling, 2013). Extensive experiments on both numerical and synthetic datasets demonstrate its effectiveness. Most notably, our framework enables the discovery of latent causal variables that reflect complex biological interactions and analyze potential causal mechanisms between different modalities, which are important for clinical decision-making. The evaluation results on a real-world human phenotype dataset provide novel insights into relationships between modalities, and the discovered causal relationships align with findings from medical research, highlighting our contributions to the biological domain.

2 Related Work

Machine learning (ML) models for biology. For biological applications, ML models are designed to extract informative representations to facilitate downstream tasks, including DNA sequence modeling (Zhou et al., 2024; Nguyen et al., 2023; Dalla-Torre et al., 2023), protein structure prediction (Jumper et al., 2021; Lin et al., 2023), and disease detection (Zhou et al., 2023; Jang et al., 2024). While sequence modeling of DNA, RNA, and proteins has been well developed, thanks to

the success of large language models (LLMs) (Celaj et al., 2023; Shulgina et al., 2024; Nguyen et al., 2024; Li et al., 2023; Chen et al., 2023; Lin et al., 2023), this approach focuses on a single modality, thus constraining its applicability to multi-model datasets, which is often the case for biological measures. Although some efforts have been made toward integrating multi-modal biological data (Garau-Luis et al., 2024; Pei et al., 2024; Taylor et al., 2022), these approaches often lack theoretical guarantees, limiting the trustworthiness of the model output. In this paper, we leverage causal principles to develop theoretically-sound ML models for multi-modal biological data, toward providing reliable and interpretable insights into the underlying biological structure.

CRL and multimodality. CRL aims to discover the high-level causal variables from low-level observations, regarded as a combined field of machine learning and causality (Schölkopf et al., 2021). The methods of CRL with identifiability conditions can be categorized according to the additional assumed structures, including functional constraints (Xu et al., 2024; Zheng et al., 2022; Zheng & Zhang, 2023; Buchholz et al., 2022), interventional/multi-distribution (Hyvarinen et al., 2019; Khemakhem et al., 2020a; Zhang et al., 2024b; Kong et al., 2023; Buchholz et al., 2024; von Kügelgen et al., 2023; Zhang et al., 2024a; Li et al., 2024; Varici et al., 2023; Ahuja et al., 2023; Jiang & Aragam, 2023; Brehmer et al., 2022; Lachapelle et al., 2024), and of our particular interest, multimodality (Yao et al., 2023; Morioka & Hyvarinen, 2023; 2024; Daunhawer et al., 2023; Sturma et al., 2023; Gresele et al., 2020). For clarity, in Table 1, we summarize representative works in the multimodality category and compare them with our work.

Empirical CRL for multimodal applications. Unlike aforementioned works that pay significant attention to identifiability, a line of CRL works instead focuses more on the practical applications in various fields, without considering theoretical identifiability. Mao et al. (2022) assume independent latent variables and propose a two-module amortized variational algorithm to learn representations from medical images and other biological data. Zheng et al. (2024a) develop a contrastive learning-based approach to extract modality-specific and modality-invariant representations from time-series tabular data and language text data for root cause analysis. Rawls et al. (2021) leverage behavioral and psychiatric phenotyping and high-resolution neuroimaging data from Human Connectome Project (Van Essen et al., 2013) and run existing Greedy Fast Causal Inference (Ogarrio et al., 2016) to analyze the causal relations for alcohol use disorder. Differently, we provide formal identification theory and further bake the derived insights into our estimation model.

General multimodal learning. In a general ML context, multimodal learning involves learning a representation from multimodal data (e.g., text, image, audio) towards specific tasks (Manzoor et al., 2023; Zhang et al., 2020). Among these representation learning methods for weakly supervised data, contrastive learning (Daunhawer et al., 2023; Wang et al., 2022; Peng et al., 2022; Radford et al., 2021; Khosla et al., 2020; Oord et al., 2018) stands out for its effectiveness, scalability, and robustness, with a prominent example being `CLIP` (Radford et al., 2021). In contrast with these works, our work focuses on discovering the causal relationships among the learned underlying factors over multiple modalities, with the goal of generating novel insights into the biological system.

Table 1: **Related work on multi-modal causal representation learning.** This table considers whether more than two modalities can be handled, whether the latent-variable distribution is nonparametric, whether the mixing function can be nonlinear, and whether the identifiability is component-wise.

Related work	> 2 Modalities	Nonparametric Prior	Nonlinear Mixing	Component-wise Iden.
Von Kügelgen et al. (2021)	×	✓	✓	×
Daunhawer et al. (2023)	×	✓	✓	×
Morioka & Hyvarinen (2024)	✓	×	✓	✓
Yao et al. (2023)	✓	✓	✓	×
Ours	✓	✓	✓	✓

3 Latent Multimodal Causal Models

Real-world biological datasets are often curated with multiple modalities, each characterizing a distinct yet interrelated aspect of the subject. For instance, human phenotype datasets (Levine et al., 2024) consist of tabular data, time series, images, and text capturing distinct biological measurements including anthropometrics, sleep monitoring, and genetics. Understanding the latent factors behind each modality and their interplay can provide valuable insights into underlying biological

mechanisms, which in turn facilitates the advancement of medical technologies. With this goal in mind, we formalize such multimodal data-generating processes as follows.

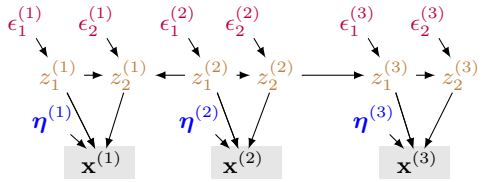


Figure 2: Illustrative examples of the hypothesis space underlying the biology system.

Data-generating processes. Let $\mathbf{x} := [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}]$ be a set of observations/measurements from M modalities, where $\mathbf{x}^{(m)} \in \mathbb{R}^{d(\mathbf{x}^{(m)})}$ represents the observation from modality m with dimensionality $d(\mathbf{x}^{(m)})$. Let $\mathbf{z} = [\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(M)}]$ be the set of causally related latent variables underlying M modalities. Specifically, the data generation process (Figure 2) can be formulated as

$$z_i^{(m)} := g_{z_i^{(m)}}(\mathrm{Pa}(z_i^{(m)}), \epsilon_i^{(m)}), \quad \text{(latent causal relations)} \tag{1}$$

$$\mathbf{x}^{(m)} := g_{\mathbf{x}^{(m)}}(\mathbf{z}^{(m)}, \boldsymbol{\eta}^{(m)}), \quad \text{(generating functions)} \tag{2}$$

where we denote the parents of a variable with $\mathrm{Pa}(\cdot)$. Since we allow for causal relations within each modality and across multiple modalities, $\mathrm{Pa}(\cdot)$ potentially returns latent variables across multiple modalities. The differentiable function $g_{\mathbf{z}}$ encodes the latent causal ~~directed acyclic~~ graph connecting latent components and its Jacobian matrix $\boldsymbol{J}_{g_{\mathbf{z}}}$ can be permuted into a strictly triangular matrix. We use ϵ_i^m to denote the exogenous variable for $z_i^{(m)}$ and exogenous variables are mutually independent. We use $\boldsymbol{\eta}^{(m)}$ to denote domain-specific information independent of other components.

Example. In healthcare, a chest X-ray $\mathbf{x}^{(m)}$ may reflect latent factors such as lung functions, heart sizes, and bone structures, represented by $\mathbf{z}^{(m)}$. These latent variables can causally influence those in other modalities $\mathbf{z}^{(n)}$ such as demographic factors like age and sex which may be reflected in heart rhythm and electrical conduction in an ECG represented by $\mathbf{x}^{(n)}$.

Goal. As outlined previously, we aim to learn the latent variables underlying each modality and their causal relations. Formally, for two specifications $\boldsymbol{\theta} := \{g_{\mathbf{x}^{(m)}}, g_{\mathbf{z}^{(m)}}, p(\boldsymbol{\epsilon}^{(m)})\}_{m=1}^M$ and $\hat{\boldsymbol{\theta}} := \{\hat{g}_{\mathbf{x}^{(m)}}, \hat{g}_{\mathbf{z}^{(m)}}, \hat{p}(\boldsymbol{\epsilon}^{(m)})\}_{m=1}^M$ of the data-generating process Eq. (1) and Eq. (2) that fit the marginal distribution $p(\mathbf{x})$, we would like to show that, *given the same* $\mathbf{x}$ *value*, each latent component $\hat{z}_i^{(m)}$ is equivalent to its counterpart $z_i^{(m)}$ up to an invertible map $h_i^{(m)}$, i.e., $\hat{z}_i^{(m)} = h_i^{(m)}(z_i^{(m)})$. This *component-wise identifiability* disentangles latent components (e.g., gene types, nutrient levels) from the measurements $\mathbf{x}$ and preserves their original information. Once component-wise identifiability is achieved, one can readily apply classical causal learning algorithms (e.g., PC (Spirtes et al., 2001)) to identified components $\hat{z}_i^{(m)}$ to infer the graphical structures. One can choose structural learning algorithms suitable to the assumed graph class (e.g., potentially non-DAGs) and this step is orthogonal to our contribution. These structures characterize the interactions among all latent components across modalities as we desire for the biological applications.

4 Identification Theory

As motivated in Section 3, we address the component-wise identifiability of latent components $z_i^{(m)}$.

Remarks on the problem. Identification for multimodal distributions often leverages the structure among available modalities. However, component-wise identification, especially in the general nonparametric setting, is challenging. Daunhawer et al. (2023); Von Kügelgen et al. (2021); Yao et al. (2023) require certain information redundancy: the information of the latent variables should be fully shared and preserved by at least two modalities' observations – that is, we can express $\mathbf{z}^{(m)}$ as functions of $\mathbf{x}^{(m_1)}$ and $\mathbf{x}^{(m_2)}$ individually. Moreover, the identification can only be achieved up to *subspaces* (i.e., groups of latent components) determined by the sharing pattern. However, often the latent components may not be fully shared by multiple modalities. For example, in health monitoring, while sleep patterns do not directly reveal genetic predispositions, genetic factors can influence sleep disorders. In that case, the subspace identification can fall short of providing detailed interpretations of biological systems and the mechanisms encoded in the graphical structures

over individual causal components. For work that achieves component-wise identifiability, Morioka & Hyvarinen (2023; 2024) assume that the latent distribution $p(\{\mathbf{z}^{(m)}\}_{m=1}^{M})$ follows an exponential family form with additive causal influences from multiple parents, which may be restrictive in general cases. For instance, in brain imaging studies, fMRI data and EEG data capture different neural activities, and the interactions between brain regions are often highly nonlinear. Clearly, for general multimodal distributions (Figure 2), we cannot access the information redundancy assumed in Daunhawer et al. (2023); Von Kügelgen et al. (2021); Yao et al. (2023) and the nicely-behaved latent causal models in parametric assumptions (Morioka & Hyvarinen, 2023; 2024).

Our high-level approach. We decompose the problem into two stages: we first identify latent subspaces $\mathbf{z}^{(m)}$ (Section 4.1) and further disentangle identified subspaces into components $z_i^{(m)}$ (Section 4.2). For the subspace identification, we only assume the information of subspace $\mathbf{z}^{(m)}$ is preserved in its corresponding observation $\mathbf{x}^{(m)}$ and exerts sufficient influence on other modalities' observations $\mathbf{x}^{(-m)}$, thus weakening the redundancy assumption in prior work (Daunhawer et al., 2023; Yao et al., 2023). For the component-wise identification, we leverage a natural notion of structural sparsity in the literature (Zheng et al., 2022; Lachapelle et al., 2024) – the dependency among all the modalities should be explained with a minimal number of causal edges among latent subspaces $\{\mathbf{z}^{(m)}\}_{m=1}^{M}$. This enables us to further disentangle each subspace into components, without resorting to parametric assumptions (Morioka & Hyvarinen, 2023; 2024).

Notations. We adopt the notation $d(\cdot)$ and $I(\cdot)$ to indicate the dimensionality and the component indices of the argument, respectively. We use $-m$ to indicate the complement of the modality m and bracketed (m) in superscripts and subscripts to index modality m directly. We denote sub-matrices $[\cdot]_{R,C}$ where R and C are index sets. Under this notation, setting R (or C) as : indicates all indices along that dimension.

4.1 Identifying Latent Subspaces

As previously discussed, we now provide the subspace identifiability. Formally, we would like to show that the estimated latent subspace $\hat{\mathbf{z}}^{(m)}$ for any modality m and its true counterpart $\mathbf{z}^{(m)}$ are equivalent up to an invertible map $h^{(m)}(\cdot)$, i.e., $\hat{\mathbf{z}}^{(m)} = h^{(m)}(\mathbf{z}^{(m)})$.

Given the data-generating process Eq. (2), the task amounts to removing modality-specific information $\boldsymbol{\eta}^{(m)}$ from the observation data $\mathbf{x}^{(m)}$ while retaining the latent variables $\mathbf{z}^{(m)}$ causally related with other modalities. In light of this, we express the relations between latent variables $\mathbf{z}^{(m)}$ and the observation of its own modality $\mathbf{x}^{(m)}$ and other modalities $\mathbf{x}^{(-m)}$ as Eq. (3).

$$\mathbf{x}^{(m)} = g_{\mathbf{x}^{(m)}}(\mathbf{z}^{(m)}, \boldsymbol{\eta}^{(m)}), \quad \mathbf{x}^{(-m)} = \tilde{g}_{\mathbf{x}^{(-m)}}(\mathbf{z}^{(m)}, \tilde{\boldsymbol{\eta}}^{(-m)}), \tag{3}$$

where $\tilde{\boldsymbol{\eta}}^{(-m)}$ denotes all the information necessary to generate the complement group $\mathbf{x}^{(-m)}$ beyond $\mathbf{z}^{(m)}$. Consequently, $\tilde{\boldsymbol{\eta}}^{(-m)}$ may admit causal/statistic relationships with $\mathbf{z}^{(m)}$.[1] We denote the joint map of $g_{\mathbf{x}^{(m)}}$ and $\tilde{g}_{\mathbf{x}^{(-m)}}$ as $\tilde{g}^{(m)} : (\mathbf{z}^{(m)}, \boldsymbol{\eta}^{(m)}, \tilde{\boldsymbol{\eta}}^{(-m)}) \mapsto \mathbf{x}$.

Condition 4.1 (Subspace Identifiability Conditions)**.**

A1 [Smoothness & Invertibility]: The generating functions $g_{\mathbf{x}^{(m)}}$ and $\tilde{g}^{(m)}$ are smooth and have smooth inverse functions.

A2 [Linear Independence]: The generating function $\tilde{g}_{\mathbf{x}^{(-m)}}$ is smooth and its Jacobian columns corresponding to $\mathbf{z}^{(m)}$ (i.e., $[\boldsymbol{J}_{\tilde{g}_{\mathbf{x}^{(-m)}}}]_{:,I(\mathbf{z}^{(m)})}$) are linearly independent almost anywhere.

Discussion on the conditions. Condition 4.1-A1 requires that the latent variables $\mathbf{z}^{(m)}$ information is preserved in its observation $\mathbf{x}^{(m)}$, so that the identification of latent variables is well-defined (Hyvarinen et al., 2019; Khemakhem et al., 2020a; Von Kügelgen et al., 2021; Kong et al., 2023; Yao et al., 2023; Daunhawer et al., 2023). Since this holds for any modality m, the observations $\mathbf{x}^{(-m)}$ should collectively preserve the information of other modality $\mathbf{z}^{(-m)}$. Condition 4.1-A2

[1] We use $\tilde{\cdot}$ to differentiate $\tilde{\boldsymbol{\eta}}^{(-m)}$ from a collection of modality-specific variables $\boldsymbol{\eta}$ defined in Eq. (2).

formalizes the notation of a minimal connectivity over modalities: $\mathbf{z}^{(m)}$ should also exert sufficient influence over other modalities $\mathbf{z}^{(-m)}$, so that the other modality observations $\mathbf{x}^{(-m)}$ could be informative to identify $\mathbf{z}^{(m)}$. This condition excludes degenerate scenarios in which the causal influences among modalities are nearly negligible and is equivalent to local invertibility of $\mathbf{z}^{(m)}$ strictly weaker than the global invertibility assumption in prior work (Daunhawer et al., 2023; Von Kügelgen et al., 2021; Yao et al., 2023) (e.g., $y = x^2$ is locally invertible but not globally so), as discussed earlier.

Theorem 4.2 (Subspace Identifiability). *Let $\boldsymbol{\theta} := \{g_{\mathbf{x}^{(m)}}, \tilde{g}_{\mathbf{z}^{(-m)}}, p(\boldsymbol{\epsilon}^{(m)}), p(\tilde{\boldsymbol{\epsilon}}^{(-m)})\}_{m=1}^{M}$ and $\hat{\boldsymbol{\theta}} := \{\hat{g}_{\mathbf{x}^{(m)}}, \hat{\tilde{g}}_{\mathbf{z}^{(-m)}}, p(\hat{\boldsymbol{\epsilon}}^{(m)}), p(\hat{\tilde{\boldsymbol{\epsilon}}}^{(-m)})\}_{m=1}^{M}$ be two specifications of the data-generating process in Eq. (3). Suppose that they generate identical observational distributions (i.e., $p(\mathbf{x}) = \hat{p}(\mathbf{x})$), $\boldsymbol{\theta}$ satisfies Condition 4.1, and $\hat{\boldsymbol{\theta}}$ satisfies Condition 4.1-A1. The latent subspace $\hat{\mathbf{z}}^{(m)}$ for any group m and its counterpart $\mathbf{z}^{(m)}$ are equivalent up to an invertible map $h^{(m)}(\cdot)$, i.e., $\hat{\mathbf{z}}^{(m)} = h^{(m)}(\mathbf{z}^{(m)})$.*

Interpretation and proof sketch. Theorem 4.2 states that one can disentangle the modality-specific information $\boldsymbol{\eta}^{(m)}$ and the latent variables $\mathbf{z}^{(m)}$ contained in the observation $\mathbf{x}^{(m)}$ (which is a mixture of both). To attain this, we leverage the fact that $\boldsymbol{\eta}^{(m)}$ doesn't have influence over other modalities $\mathbf{x}^{(-m)}$ whereas $\mathbf{z}^{(m)}$ has a nontrivial influence over $\mathbf{x}^{(-m)}$ as characterized in Condition 4.3-A2. This crucial distinction gives sufficient footprints to disentangle these two subspaces for each modality, yielding the intended result.

4.2 Identifying Latent Components

Proceeding from the subspace identifiability (Theorem 4.2), we now further disentangle each subspace into individual components $z_i^{(m)}$ as defined in Section 3. As foreshadowed, our key condition entails the sparsity of the graphical structures among modalities. Such dependency structures are captured in the generating function $g_{\mathbf{z}}$ defined component-wise in Eq. (1), in particular its partial derivatives. We now introduce Condition 4.3 that facilitates component-wise identification.

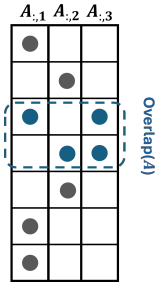


Figure 3: **Sparse A.**

Additional notations. We denote latent components in modality m that are parents (resp. children) to latent components in other modalities as upstream variables $U^{(m)}$ (resp. downstream variables $D^{(m)}$). We denote the nonzero matrix entries' indices with $\text{Supp}(\cdot)$. We use (m) and $(-m)$ in matrix subscripts to index dimensions of modality m and all modalities other than m respectively. We denote the collection of partial derivatives among all latent components $\frac{\partial z_i^{(m)}}{\partial z_j^{(n)}}$ as a matrix function $\boldsymbol{G}(\mathbf{z}, \boldsymbol{\epsilon}) \in \mathbb{R}^{d(\mathbf{z}) \times d(\mathbf{z})}$. We denote the sub-matrix consisting of matrix $\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_2}$' rows with more than one non-zero entries as $\text{Overlap}(\boldsymbol{A}) := \boldsymbol{A}_{S,:}$ where $S := \{i \in [d_1] : \|\boldsymbol{A}_{i,:}\|_0 > 1\}$. We adopt $\text{diag}(\cdot)$ to denote matrices consisting of equally-sized square matrices on its diagonal. Further, for any $\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_2}$, define $d^*(\boldsymbol{A}) := \max\{|R| : R \subset [d_1],\ \text{rank}\left([\text{Overlap}(\boldsymbol{A})]_{R,:}\right) < d_2\}$.

Condition 4.3 (Component Identifiability Conditions). Over the domain of $(\mathbf{z}, \boldsymbol{\epsilon})$, for any modality m, nonempty $C^{(m)} \subset U^{(m)}$, nonempty $R^{(m)} \subset D^{(m)}$, and $\boldsymbol{T} = \text{diag}(\boldsymbol{T}_1, \ldots, \boldsymbol{T}_{M-1})$ with invertible $\boldsymbol{T}_i \in \mathbb{R}^{d(\mathbf{z}^{(i)}) \times d(\mathbf{z}^{(i)})}$, we have

$$\begin{aligned} &\left| \bigcup_{j \in I(C^{(m)})} \text{Supp}([\boldsymbol{TG}]_{(-m),j}) \right| - d^*\left([\boldsymbol{TG}]_{(-m),I(C^{(m)})}\right) > \max_{j \in I(C^{(m)})} \left\| [\boldsymbol{G}]_{(-m),j} \right\|_0; \\ &\left| \bigcup_{j \in I(R^{(m)})} \text{Supp}(([\boldsymbol{GT}^{-1}]_{j,(-m)})^\top) \right| - d^*\left(([\boldsymbol{GT}^{-1}]_{I(R^{(m)}),(-m)})^\top\right) > \max_{j \in I(R^{(m)})} \left\| [\boldsymbol{G}]_{j,(-m)} \right\|_0. \end{aligned} \tag{4}$$

Discussion on the conditions. Overall, Condition 4.3 necessitates sparse causal connections among different modalities $\{\mathbf{z}^{(m)}\}_{m=1}^{M}$. This allows each component $z_i^{(m)}$ to connect to other modalities' components $z_j^{(n)}$ $(m \neq n)$ in an ideally distinct manner, leaving a causal footprint for identification. $\boldsymbol{G}$ denotes the graphical connectivity in the model $\boldsymbol{\theta}$ and $\boldsymbol{T}$ denotes potential mixings of latent variables in the model $\hat{\boldsymbol{\theta}}$. Thus, sub-matrices $[\boldsymbol{TG}]_{(-m),I(C^{(m)})}$ and

$([\boldsymbol{G}\boldsymbol{T}^{-1}]_{I(R^{(m)}),(-m)})^\top$ represent the cross-modality connectivity between modality m and other modalities $-m$ in the model $\hat{\boldsymbol{\theta}}$. Condition 4.3 effectively imposes a sparsity constraint on these edges. Since all involved components $z_i^{(m)}$ are latent, causal directions are unknown. Moreover, any subset of variables on both sides of the causal edges could potentially be entangled. We introduce upstream variables $U^{(m)}$ and downstream variables $D^{(m)}$ to indicate the causal directions (note that $U^{(m)} \cap D^{(m)}$ could be nonempty) and utilize $\boldsymbol{T}$ to account for potential entanglements in other modalities when we look at modality m. We employ d^* to characterize the maximal amount of sparsity due to accidentally canceled out nonzero entries if a subset of $C^{(m)}$ or $R^{(m)}$ are entangled. Thus, Eq. (4) intuitively indicates that any entanglement would lead to a denser Jacobian matrix $\hat{\boldsymbol{G}}$ than its counterpart $\boldsymbol{G}$, rendering sparsity a faithful signal for entanglement (more intuition in Appendix C.2). We give one simple example for sparse $\boldsymbol{A}$ in Figure 3. The availability of multiple modalities greatly enhances the feasibility of such sparsity conditions, especially with a large number of modalities, because the entanglement is limited within a single modality (thanks to Theorem 4.1) and all other modalities can be leveraged to provide space for sparse connections.

Sparsity conditions have been embraced by the causal representation learning community (Lachapelle et al., 2024; Moran et al., 2022; Fumero et al., 2023; Xu et al., 2024). Especially related to our work is Zheng et al. (2022). As discussed above, we are obliged to deal with causal structures among all latent variables. In contrast, Zheng et al. (2022) assume the sparsity of the causal connections between the latent variables and the observed variables – the directions (from the latent to the observed variables) are given and the children are directly observed. Further, we propose an exact characterization of the potential cancellation dimensions d^*, leading to a tight bound in Eq. (4). Notably, sparse properties manifest in biological systems of our interest, including gene-regulatory networks (Milo et al., 2002; Babu et al., 2004; Nacher & Akutsu, 2013; Liu et al., 2011), metabolic systems (West et al., 2002; Banavar et al., 1999), and other living systems (Busiello et al., 2017), evidencing the plausibility of Condition 4.3 for biological applications.

Theorem 4.4 (Component-wise Identifiability). *Let $\boldsymbol{\theta} := (\{g_{\mathbf{x}^{(m)}}, g_{\mathbf{z}^{(m)}}, p(\boldsymbol{\epsilon}^{(m)})\}_{m=1}^M)$ and $\hat{\boldsymbol{\theta}} := (\{\hat{g}_{\mathbf{x}^{(m)}}, \hat{g}_{\mathbf{z}^{(m)}}, \hat{p}(\boldsymbol{\epsilon}^{(m)})\}_{m=1}^M)$ be two specifications of the data-generating process in Eq. (1) and Eq. (2). Suppose that they generate identical observational distributions (i.e., $p(\mathbf{x}) = \hat{p}(\mathbf{x})$) and $\boldsymbol{\theta}$ satisfies Condition 4.1 and Condition 4.3. If $\hat{\boldsymbol{\theta}}$ satisfies the following condition:*

$$\sum_{m \neq n \in [M]} \left\| [\boldsymbol{J}_{\hat{g}_{\mathbf{z}}}]_{(m),(n)} \right\|_0 \leq \sum_{m \neq n \in [M]} \left\| [\boldsymbol{J}_{g_{\mathbf{z}}}]_{(m),(n)} \right\|_0, \tag{5}$$

each component $z_i^{(m)}$ and its counterpart $\hat{z}_{\pi(i)}^{(m)}$ are equivalent up to an invertible map $h(\cdot)$, i.e., $\hat{z}_{\pi(i)}^{(m)} = h(z_i^{(m)})$ under a permutation π over $[d(\mathbf{z}^{(m)})]$.

Interpretation and proof sketch. The key idea of Theorem 4.4 is that for sparse causal graphs (as characterized in Condition 4.3), the mixing of latent components in any modality would introduce unnecessary causal edges connecting the other latent subspaces. We give a simple example to aid intuition: for a true causal graph $z_1^{(1)} \to z_1^{(2)}$ and $z_2^{(1)} \to z_2^{(2)}$, suppose that $\hat{z}_1^{(1)}$ is a nontrivial mixture of $z_1^{(1)}$ and $z_2^{(1)}$ and other estimates are correctly identified, i.e., $[\hat{z}_1^{(1)}, \hat{z}_2^{(1)}, \hat{z}_1^{(2)}, \hat{z}_2^{(2)}] = [h(z_1^{(1)}, z_2^{(1)}), z_2^{(1)}, z_1^{(2)}, z_2^{(2)}]$. As a consequence of the mixing, the estimated causal graph would include an additional edge $\hat{z}_1^{(1)} \to \hat{z}_2^{(2)}$, forbidden by the sparsity constraint in Eq. (**??**).

Implications. In the context of biological applications, Theorem 4.4 indicates that under proper constraints, each component $\hat{z}_i^{(m)}$ in our estimation uniquely captures the information of an intrinsic biological factor behind the medical measurements (e.g., genetic predisposition). Therefore, the learned representation enjoys strong interpretability under theoretical guarantees, which is often lacking in existing biological models as noted in Section 2. Theorem 4.2 and Theorem 4.4 offer insights for practical model design, which we employ in our architecture in Section 5.

Generality of our framework. Theorem 4.2 and Theorem 4.4 can be straightforwardly extended to identify causal models with directly shared latent variables as those in Yao et al. (2023); Daunhawer et al. (2023); Von Kügelgen et al. (2021), thus strictly more general than prior work. In

particular, we can identify such shared latent variables block-wise through incorporating contrastive learning objectives into Theorem 4.1 (Yao et al., 2023; Daunhawer et al., 2023; Von Kügelgen et al., 2021). Therefore, we can treat such blocks of shared latent variables as separate modalities and directly apply Theorem 4.4 to attain the component-wise identifiability.

5 Estimation Model Architectures

Given identifiability results, we further propose an estimation framework that enforces the proposed assumptions as constraints to identify the latent variables in each modality, as shown in Figure 4.

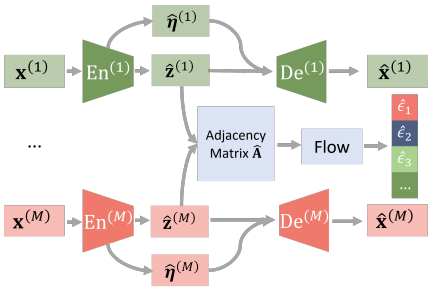


Figure 4: **Estimation framework**. Given multi-modal observations $(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)})$, the latent variables and exogenous variables in modality m are inferred as $\hat{\mathbf{z}}^{(m)}$ and $\hat{\eta}^{(m)}$ by individual encoders. The observations are then reconstructed with corresponding decoders as $\hat{\mathbf{x}}^{(m)}$. We enforce independence conditions by minimizing the KL divergence term $\mathrm{KL}\left([\{\hat{\eta}^{(m)}\}_{m=1}^{M}, \{\hat{\epsilon}_i\}_{i=1}^{d_z}]; \mathcal{N}(\mathbf{0}, \boldsymbol{I})\right)$. We enforce the sparsity constraint by minimizing the $\mathcal{L}_1$ norm in the inferred adjacency matrix $\hat{\mathbf{A}}$.

Encoder and decoder. Each modality $\mathbf{x}^{(m)}$ is given as an input to the corresponding encoder and outputs the estimated latent $\hat{\mathbf{z}}^{(m)}$ and exogenous variables $\hat{\eta}^{(m)}$. They are then concatenated and passed to the corresponding decoder to reconstruct the observations as $\hat{\mathbf{x}}^{(m)}$. The reconstruction loss is calculated using the mean squared error (MSE) as $\mathcal{L}_{\text{Recon}} = \sum_{m=1}^{M} ||\mathbf{x}^{(m)} - \hat{\mathbf{x}}^{(m)}||_2^2$.

Conditional independence constraints. We enforce the conditional independence condition $\mathbf{x}^{(m)} \perp\!\!\!\perp \mathbf{x}^{(n)} \mid \mathbf{z}^{(m)}$ and the independence condition on $\eta^{(m)} \perp\!\!\!\perp \mathbf{z}^{(m)}$ by enforcing independence among components in $\gamma = [\{\hat{\eta}^{(m)}\}_{m=1}^{M}, \{\hat{\epsilon}_i\}_{i=1}^{d_z}]$. Such equivalence is shown in Proposition B.1 and B.2, and proofs are provided in Appendix B. Specifically, we minimize the KL divergence loss between the posterior and a Gaussian prior distribution: $\mathcal{L}_{\text{Ind}} = \mathrm{KL}(p(\gamma)||\mathcal{N}(\mathbf{0}, \boldsymbol{I}))$.

Proposition 5.1. *[Conditional Independence Condition] Denote $\mathbf{x}^{(m)}$ and $\mathbf{x}^{(n)}$ are two different multimodal observations. $\mathbf{z}^{(m)} \subset \mathbf{z}$ are the set of block-identify latent variables, and $\eta^{(m)} \subset \eta$ are exogenous variables in modality m. We have $\mathbf{x}^{(m)} \perp\!\!\!\perp \mathbf{x}^{(n)} \mid \mathbf{z}^{(m)} \iff \eta^{(m)} \perp\!\!\!\perp \eta^{(n)}$.*

Proposition 5.2. *[Independent Noise Condition] Denote $\mathbf{z}$ and η as the block-identified latent variables and exogenous variables across all modalities. ϵ's are the causally-related noise terms. We have $\eta \perp\!\!\!\perp \mathbf{z} \iff \eta \perp\!\!\!\perp \epsilon$.*

Sparsity regularization. We use normalization flow (Dinh et al., 2016; Huang et al., 2018) to estimate the exogenous variables $\boldsymbol{\epsilon}$ in Eq. (1) and implement the causal relations through a learnable adjacency matrix $\hat{\mathbf{A}}$. The binary values in $\hat{\mathbf{A}}$ represent the causal generation process between latent variables, e.g. $\hat{A}_{i,j} = 1$ indicates $\hat{z}_j$ is the parent of $\hat{z}_i$, while $\hat{A}_{i,j} = 0$ means $\hat{z}_j$ dose not contribute to the generation of $\hat{z}_i$. For each component $\hat{z}_i$, we select its parents $\mathrm{Pa}(\hat{z}_i)$ based on the adjacency matrix, and apply the flow transformation from $\mathrm{Pa}(\hat{z}_i)$ to $\hat{\epsilon}_i$.

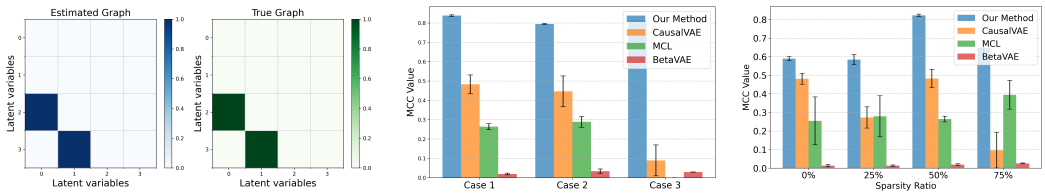
To encourage sparsity among the latent variables $\hat{\mathbf{z}}$, we introduce a regularization term on the learned adjacency matrix. Based on the sparsity assumption, the optimal causal graph should be the minimal one that still allows the model to accurately match the ground truth generative distribution. To achieve this, we reduce the dependencies between different components of $\hat{\mathbf{z}}$ by adding a $\mathcal{L}_1$ penalty on the adjacency matrix, s.t., $\mathcal{L}_{\text{Sp}} = ||\hat{\mathbf{A}}||_1$.

Optimization. The model parameters are optimized using the combination objective:

$$\mathcal{L} = \alpha_{\text{Ind}}\mathcal{L}_{\text{Ind}} + \alpha_{\text{Sp}}\mathcal{L}_{\text{Sp}} + \alpha_{\text{Recon}}\mathcal{L}_{\text{Recon}}. \tag{6}$$

6 EXPERIMENT RESULTS

To evaluate the efficacy of our proposed method, we conduct extensive experiments on (1) numerical, (2) synthetic and (3) real-world datasets. In terms of the baselines, we compare our method with: (1) BetaVAE (Higgins et al., 2017), which does not consider causal relationships in the latent space. (2) CausalVAE (Yang et al., 2020), which considers the causally related latent variables with a single modality. (3) Multimodal contrastive learning (MCL) (Daunhawer et al., 2023), which recovers the latent factors from multimodality through contrastive learning. Throughout the experiments, we consider the following evaluation metrics: (1) Mean Correlation Coefficient (MCC) measures how well the estimated latent variables match the true ones, with an MCC of 1 indicating perfect identifiability up to permutation and invertible transformations. (2) R2 measures the proportion of variance in the ground truth latent that is explained by the estimated latent, with a value of 1 indicating perfect reconstruction. (3) Structural Hamming Distance (SHD) quantifies the difference between the estimated and true causal skeletons, where a lower SHD indicates better recovery.



(a) Causal comparison between estimated and true graphs (SHD=0).

(b) Comparison of identifiability result under different cases.

(c) Identifiability result under different sparsity ratios.

Figure 5: Numerical experiment results. (a) Successful recovery of the inter-modal causal graph. (b) Baseline comparisons in different cases. (c) Sparsity ablation result.

6.1 NUMERICAL DATASET

Setup. In the numerical simulations, we consider three cases with different numbers of modalities and inter-modal causal relations. Case 1: 15-dimensional observations across two modalities, each with two latent and one exogenous variable. Case 2: 20-dimensional observations across two modalities, each with three latent and one exogenous variable. Case 3: 15-dimensional observations across four modalities, each with two latent and one exogenous variable. The nonparametric mixing function is simulated by a random MLP with LeakyReLU units, and the inter-modal latent variables are sparse causally related. The detailed data generation process is provided in Appendix D.1.

Results and ablation. Figure 5 shows the identifiability results in different cases, where the high MCC indicates the successful recovery of the latent variables. The inter-modal causal relations are successfully recovered (SHD=0) and the comparison result in case 1 is shown in Figure 5(a). The comparisons with the baselines are shown in Figure 5(b) (MCL is not applicable in case 3 due to the two modality limitation). CausalVAE requires additional supervision signals to establish identifiability, and MCL assumes content invariance and can only block identify latent variables. In general, these baselines neither account for the multimodal setting nor for the modality-specific latent variables, and therefore do not recover the latent variables.

As an ablation study, we further show the consequences of violating the sparsity assumptions to validate our theorem. Based on case 1, we create four types of datasets with different sparsity ratios and report the MCC in each scenario in Figure 5(c). The sparsity ratio represents the proportion of existing causal links to all possible causal links between modality-specific latent variables. A value of 0 indicates that the latent variables between modalities are fully connected, while higher values correspond to sparser connections. The result shows that identifiability can be better achieved with a higher sparsity ratio, and our framework outperforms other baselines in all scenarios.

6.2 SYNTHETIC DATASET: VARIANT MNIST

Setup. We use the colored MNIST (Arjovsky et al., 2019) and fashion MNIST (Xiao et al., 2017) as two modalities of image observations derived from variants of the MNIST dataset (LeCun, 1998). The two latent variables for colored MNIST are the class label (cause) and the image color (effect),

while for fashion MNIST, they are the class label (cause) and the image rotation angle (effect). The class labels from the colored MNIST serve as the cause for the class labels in the fashion MNIST in a non-deterministic manner. More data descriptions are provided in Appendix D.2.

Results. Table 2 presents the identifiability comparison results, where the highest MCC and R2 indicate the strong performance of our method. BetaVAE does not account for latent variables, and CausalVAE, which requires additional supervision, fails to recover the latent variables effectively.

Table 2: The results of MNIST dataset.

	MCL	BetaVAE	CausalVAE	Ours
R2	0.48 $\pm$ 0.01	0.22 $\pm$ 0.00	0.02 $\pm$ 0.01	**0.89** $\pm$ 0.05
MCC	0.82 $\pm$ 0.02	0.03 $\pm$ 0.00	0.14 $\pm$ 0.01	**0.87** $\pm$ 0.02

6.3 Real-world Dataset: Human Phenotype

The human phenotype dataset (Shilo et al., 2021) is a large-scale, longitudinal collection of phenotypic profiles from a diverse global population. It includes comprehensive human health data and provides a comprehensive view of health and disease drivers. The dataset contains various types of participant information, categorized into tabular, time series, and image data. Specifically, it includes health information across 30 modalities, such as blood tests, anthropometry, fundus imaging, etc. Detailed data descriptions can be found in Appendix D.3.

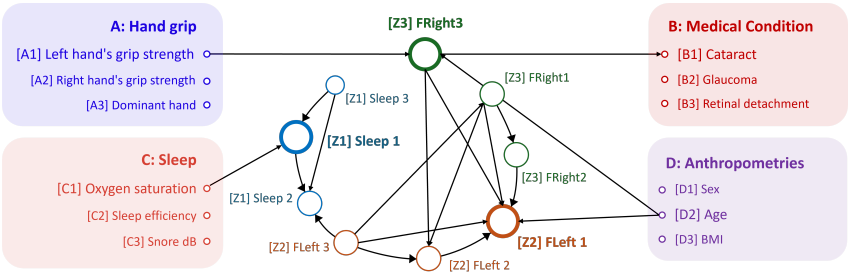


Figure 6: Causal analysis results across different modalities, including hand grip, medical conditions, sleep, and anthropometries. We ran the causal algorithm on all variables but reported only the causal relations that have direct connections to the estimated latent variables for clarity.

In this work, we focus on the fundus imaging dataset for both right and left eyes (*FRight* and *FLeft*) and the sleep monitoring dataset (*Sleep*) to estimate the latent factors underlying each modality. To validate our result, we applied the PC algorithm (Spirtes et al., 2001) to discover causal relationships between the estimated latent variables (Z1, Z2, Z3) and other four additional tabular modalities (A, B, C, D), providing an implicit evaluation on the effectiveness. The result with direct causal relations is shown in Figure 6, with variables from the same modality sharing the same color and different modalities in distinct colors.

A key finding is that the causal relationships discovered are consistent with findings from medical research. For example, *Sleep 1* shows a direct causal relationship with *Oxygen saturation*, suggesting that sleep conditions may influence blood oxygen levels. This observation is consistent with previous studies (Wali et al., 2020). In addition, the fundus-related latent variables *FRight 1* and *FLeft 1* have a direct causal relationship with *Age*, suggesting that aging plays an important role in changes in retinal health (Ege et al., 2002; Einbock et al., 2005). Interestingly, the fundus image of the right eye has a direct causal relationship with the grip strength of the left hand, as recently demonstrated in biological research (Bikbov et al., 2023; Qiu et al., 2020).

7 Conclusion and Limitations

In this work, we develop a theoretically grounded framework for recovering latent causal variables from multi-modal observations. Extensive experimental results on synthetic and real-world datasets demonstrate the practical effectiveness of our approach. **Limitations:** Empirically, our framework assumes prior knowledge of the number of latent variables in each modality, which may be unrealistic in real-world scenarios. Additionally, a detailed evaluation against the quantitative benchmarks used in biological models remains an area for future exploration.

BIBLIOGRAPHY

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International conference on machine learning*, pp. 372–407. PMLR, 2023.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

M Madan Babu, Nicholas M Luscombe, L Aravind, Mark Gerstein, and Sarah A Teichmann. Structure and evolution of transcriptional regulatory networks. *Current opinion in structural biology*, 14(3):283–291, 2004.

Jayanth R Banavar, Amos Maritan, and Andrea Rinaldo. Size and form in efficient transportation networks. *Nature*, 399(6732):130–132, 1999.

Mukharram M Bikbov, Rinat M Zainullin, Timur R Gilmanshin, Ellina M Iakupova, Gyulli M Kazakbaeva, Songhomitra Panda-Jonas, Azaliia M Tuliakova, Albina A Fakhretdinova, Leisan I Gilemzianova, and Jost B Jonas. Hand grip strength and ocular associations: the ural eye and medical study. *British Journal of Ophthalmology*, 107(10):1567–1574, 2023.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35:38319–38331, 2022.

Simon Buchholz, Michel Besserve, and Bernhard Schölkopf. Function classes for identifiable nonlinear independent component analysis. *Advances in Neural Information Processing Systems*, 35: 16946–16961, 2022.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *Advances in Neural Information Processing Systems*, 36, 2024.

Daniel M Busiello, Samir Suweis, Jorge Hidalgo, and Amos Maritan. Explorability and the origin of network sparsity in living systems. *Scientific reports*, 7(1):12323, 2017.

Albi Celaj, Alice Jiexin Gao, Tammy TY Lau, Erle M Holgersen, Alston Lo, Varun Lodaya, Christopher B Cole, Robert E Denroche, Carl Spickett, Omar Wagih, et al. An rna foundation model enables discovery of disease mechanisms and candidate therapeutics. *bioRxiv*, pp. 2023–09, 2023.

Ken Chen, Yue Zhou, Maolin Ding, Yu Wang, Zhixiang Ren, and Yuedong Yang. Self-supervised learning on millions of pre-mrna sequences improves sequence-based rna splicing prediction. *bioRxiv*, pp. 2023–01, 2023.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023. doi: 10.1101/2023.01.11.523679. URL https://www.biorxiv.org/content/early/2023/01/15/2023.01.11.523679.

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E Vogt. Identifiability results for multimodal contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Armando D Diaz Gonzalez, Kevin S Hughes, Songhui Yue, and Sean T Hayes. Applying biobert to extract germline gene-disease associations for building a knowledge graph from the biomedical literature. In *Proceedings of the 2023 7th International Conference on Information System and Data Mining*, pp. 37–42, 2023.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Bernhard M Ege, Ole K Hejlesen, Ole V Larsen, and Toke Bek. The relationship between age and colour content in fundus images. *Acta Ophthalmologica Scandinavica*, 80(5):485–489, 2002.

Wilma Einbock, Andreas Moessner, Ute EK Schnurrbusch, Frank G Holz, Sebastian Wolf, and FAM Study Group. Changes in fundus autofluorescence in patients with age-related maculopathy. correlation to visual function: a prospective study. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 243:300–305, 2005.

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

Juan Jose Garau-Luis, Patrick Bordes, Liam Gonzalez, Masa Roller, Bernardo P. de Almeida, Lorenz Hexemer, Christopher Blum, Stefan Laurent, Jan Grzegorzewski, Maren Lang, Thomas Pierrot, and Guillaume Richard. Multi-modal transfer learning between biological foundation models, 2024. URL https://arxiv.org/abs/2406.14150.

Luigi Gresele, Paul K Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ica. In *Uncertainty in Artificial Intelligence*, pp. 217–227. PMLR, 2020.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622, 2024.

Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017.

Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural autoregressive flows. In *International conference on machine learning*, pp. 2078–2087. PMLR, 2018.

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

Boa Jang, Youngbin Ahn, Eun Kyung Choe, Chang Ki Yoon, Hyuk Jin Choi, and Young-Gon Kim. A disease-specific foundation model using over 100k fundus images: Release and validation for abnormality and multi-disease classification on downstream tasks. *arXiv preprint arXiv:2408.08790*, 2024.

Yibo Jiang and Bryon Aragam. Learning nonparametric latent causal graphs with unknown interventions. *Advances in Neural Information Processing Systems*, 36:60468–60513, 2023.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

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, 2020a.

Ilyes Khemakhem, Ricardo Monti, Diederik Kingma, and Aapo Hyvarinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. *Advances in Neural Information Processing Systems*, 33:12768–12778, 2020b.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial identifiability for domain adaptation. *arXiv preprint arXiv:2306.06510*, 2023.

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*, 2024.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Zachary Levine, Iris Kalka, Dmitry Kolobkov, Hagai Rossman, Anastasia Godneva, Smadar Shilo, Ayya Keshet, Daphna Weissglas-Volkov, Tal Shor, Alon Diament, et al. Genome-wide association studies and polygenic risk score phenome-wide association studies across complex phenotypes in the human phenotype project. *Med*, 5(1):90–101, 2024.

Sizhen Li, Saeed Moayedpour, Ruijiang Li, Michael Bailey, Saleh Riahi, Lorenzo Kogler-Anele, Milad Miladi, Jacob Miner, Dinghai Zheng, Jun Wang, et al. Codonbert: Large language models for mrna design and optimization. *bioRxiv*, pp. 2023–09, 2023.

Xiutian Li, Siqi Sun, and Rui Feng. Causal representation learning via counterfactual intervention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3234–3242, 2024.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Yang-Yu Liu, Jean-Jacques Slotine, and Albert-László Barabási. Controllability of complex networks. *nature*, 473(7346):167–173, 2011.

Muhammad Arslan Manzoor, Sarah Albarri, Ziting Xian, Zaiqiao Meng, Preslav Nakov, and Shangsong Liang. Multimodality representation learning: A survey on evolution, pretraining and its applications. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–34, 2023.

Haiyi Mao, Hongfu Liu, Jason Xiaotian Dou, and Panayiotis V Benos. Towards cross-modal causal structure and representation learning. In *Machine Learning for Health*, pp. 120–140. PMLR, 2022.

Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

G Moran, D Sridhar, Y Wang, and D Blei. Identifiable deep generative models via sparse decoding. *Transactions on machine learning research*, 2022.

Hiroshi Morioka and Aapo Hyvarinen. Connectivity-contrastive learning: Combining causal discovery and representation learning for multimodal data. In *International conference on artificial intelligence and statistics*, pp. 3399–3426. PMLR, 2023.

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

Jose C Nacher and Tatsuya Akutsu. Structural controllability of unidirectional bipartite networks. *Scientific reports*, 3(1):1647, 2013.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A. Baccus, and Chris Ré. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution, 2023. URL https://arxiv.org/abs/2306.15794.

Eric Nguyen, Michael Poli, Matthew G Durrant, Armin W Thomas, Brian Kang, Jeremy Sullivan, Madelena Y Ng, Ashley Lewis, Aman Patel, Aaron Lou, et al. Sequence modeling and design from molecular to genome scale with evo. *BioRxiv*, pp. 2024–02, 2024.

Juan Miguel Ogarrio, Peter Spirtes, and Joe Ramsey. A hybrid causal search algorithm for latent variable models. In *Conference on probabilistic graphical models*, pp. 368–379. PMLR, 2016.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Soumen Pal, Manojit Bhattacharya, Md Aminul Islam, and Chiranjib Chakraborty. Chatgpt or llm in next-generation drug discovery and development: pharmaceutical and biotechnology companies can make use of the artificial intelligence-based device for a faster way of drug discovery and development. *International Journal of Surgery*, 109(12):4382–4384, 2023.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Xiaozhuan Liang, Yin Fang, Jinhua Zhu, Shufang Xie, Tao Qin, and Rui Yan. Biot5+: Towards generalized biological understanding with iupac integration and multi-task tuning. *arXiv preprint arXiv:2402.17810*, 2024.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8238–8247, 2022.

Alejandro Pérez-Castilla, Amador García-Ramos, Beatríz Redondo, Fernández-Revelles Andrés, Raimundo Jiménez, and Jesús Vera. Determinant factors of intraocular pressure responses to a maximal isometric handgrip test: hand dominance, handgrip strength and sex. *Current Eye Research*, 46(1):64–70, 2021.

Zihan Qiu, Wei Wang, Yan Tan, Miao He, Langhua Wang, Yuting Li, Xia Gong, and Wenyong Huang. Associations of grip strength with retinal and choroidal thickness in patients with type 2 diabetes mellitus without retinopathy: a cross-sectional study. *BMJ open*, 10(7):e036782, 2020.

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

Eric Rawls, Erich Kummerfeld, and Anna Zilverstand. An integrated multimodal model of alcohol use disorder generated by data-driven causal discovery analysis. *Communications biology*, 4(1): 435, 2021.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Smadar Shilo, Noam Bar, Ayya Keshet, Yeela Talmor-Barkan, Hagai Rossman, Anastasia Godneva, Yaron Aviv, Yochai Edlitz, Lee Reicher, Dmitry Kolobkov, et al. 10 k: a large-scale prospective longitudinal study in israel. *European journal of epidemiology*, 36(11):1187–1194, 2021.

Yekaterina Shulgina, Marena I Trinidad, Conner J Langeberg, Hunter Nisonoff, Seyone Chithrananda, Petr Skopintsev, Amos J Nissley, Jaymin Patel, Ron S Boger, Honglue Shi, et al. Rna language models predict mutations that improve rna function. *bioRxiv*, 2024.

M Slabaugh, P Chen, B Smit, and Glaucoma Today. Cataract surgery and iop. *Glaucoma Today*, pp. 17–8, 2013.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2001.

Nils Sturma, Chandler Squires, Mathias Drton, and Caroline Uhler. Unpaired multi-domain causal representation learning. *Advances in Neural Information Processing Systems*, 36, 2023.

Xin Tang, Jiawei Zhang, Yichun He, Xinhe Zhang, Zuwan Lin, Sebastian Partarrieu, Emma Bou Hanna, Zhaolin Ren, Hao Shen, Yuhong Yang, et al. Explainable multi-task learning for multi-modality biological data analysis. *Nature communications*, 14(1):2546, 2023.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science, 2022. URL https://arxiv.org/abs/2211.09085.

David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

Burak Varici, Emre Acarturk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning with interventions. *arXiv preprint arXiv:2301.08230*, 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. *Advances in neural information processing systems*, 34:16451–16467, 2021.

Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36, 2023.

Siraj Omar Wali, Bahaa Abaalkhail, Ibrahim AlQassas, Faris Alhejaili, David W Spence, and Seithikurippu R Pandi-Perumal. The correlation between oxygen saturation indices and the standard obstructive sleep apnea severity. *Annals of thoracic medicine*, 15(2):70–75, 2020.

Teng Wang, Wenhao Jiang, Zhichao Lu, Feng Zheng, Ran Cheng, Chengguo Yin, and Ping Luo. Vlmixer: Unpaired vision-language pre-training via cross-modal cutmix. In *International Conference on Machine Learning*, pp. 22680–22690. PMLR, 2022.

Geoffrey B West, Brian J Enquist, and James H Brown. Modelling universality and scaling. *Nature*, 420(6916):626–627, 2002.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

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.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Structured causal disentanglement in variational autoencoder. *arXiv preprint arXiv:2004.08697*, 2020.

Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. *arXiv preprint arXiv:2311.04056*, 2023.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. *arXiv preprint arXiv:2110.05428*, 2021.

Diana Zagirova, Stefan Pushkov, Geoffrey Ho Duen Leung, Bonnie Hei Man Liu, Anatoly Urban, Denis Sidorenko, Aleksandr Kalashnikov, Ekaterina Kozlova, Vladimir Naumov, Frank W Pun, et al. Biomedical generative pre-trained based transformer language model for age-related disease target discovery. *Aging (Albany NY)*, 15(18):9293, 2023.

Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020.

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, 36, 2024a.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. *arXiv preprint arXiv:2402.05052*, 2024b.

Lecheng Zheng, Zhengzhang Chen, Jingrui He, and Haifeng Chen. Multi-modal causal structure learning and root cause analysis. *arXiv preprint arXiv:2402.02357*, 2024a.

Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh TN Nguyen, Lauren T May, Geoffrey I Webb, and Shirui Pan. Large language models for scientific synthesis, inference and explanation. *arXiv preprint arXiv:2310.07984*, 2023.

Yizhen Zheng, Huan Yee Koh, Maddie Yang, Li Li, Lauren T May, Geoffrey I Webb, Shirui Pan, and George Church. Large language models in drug discovery and development: From disease mechanisms to clinical trials. *arXiv preprint arXiv:2409.04481*, 2024b.

Yujia Zheng and Kun Zhang. Generalizing nonlinear ica beyond structural sparsity. *Advances in Neural Information Processing Systems*, 36:13326–13355, 2023.

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.

Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome, 2024. URL https://arxiv.org/abs/2306.15006.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pp. 12979–12990. PMLR, 2021.

CONTENTS

A Notation and Terminology 18

B Conditional Independence 18

C Identifiability Theory 19

D Experimental Details 29

E Extended Experiment 32

F Implementation Details 33

G Algorithm Pseudocode 34

A NOTATION AND TERMINOLOGY

We summarize the notations used throughout the paper in Table 3.

Index	
m, n	Group index, latent index
i, j	Variable element index
d_c	Dimensionality of the shared latent variables
d_c^m	Dimensionality of the shared latent variables in group m
d_s^m	Dimensionality of the group-specific latent variables in group m
Variable	
$\mathbf{x}^{(m)}$	Observations in each group
$\mathbf{z}^{(m)}$	Latent variables in each group
$\mathbf{x}^{(m)}, \mathbf{x}^{(-m)}$	One specific observation in group m, and the rest of others
$\mathbf{z}^{(m)}, \mathbf{z}^{(-m)}$	One specific latent variables in group m, and the rest of others
$\hat{\mathbf{x}}^{(m)}$	Reconstructed observation in modality m
$\hat{z}_i$	Estimated latent variables over z_i
η	Exogenous variables
ϵ	Mutually independent noise term
$\text{Pa}(z_n)$	Set of direct cause nodes/parents of variable z_n
Function and Hyperparameter	
$g_x^{(m)}$	Nonparametric mixing function in group m
g_z	Causal function among latent variables
p	Distribution function (e.g., p_{z_i} is the distribution of z_i.)
α	Weights in the augmented ELBO objective

Table 3: List of notations.

B CONDITIONAL INDEPENDENCE

Here we provide the proofs for the constraints utilized in the estimation framework.

Proposition B.1. *[Conditional Independence Condition] Denote $\mathbf{x}^{(m)}$ and $\mathbf{x}^{(n)}$ are two different multimodal observations. $\mathbf{z}^{(m)} \subset \mathbf{z}$ are the set of block-identify latent variables, and $\eta^{(m)} \subset \eta$ are exogenous variables in modality m. We have*

$$\mathbf{x}^{(m)} \perp\!\!\!\perp \mathbf{x}^{(n)} \mid \mathbf{z}^{(m)} \Longleftrightarrow \eta^{(m)} \perp\!\!\!\perp \eta^{(n)}. \tag{7}$$

Proof. Given the data generation process in Eq. (2), the following assumptions hold true for any $m, n \in [M]$: (1) $\mathbf{z}^{(m)} \perp\!\!\!\perp \eta^{(m)}$; (2) $\mathbf{z}^{(m)} \perp\!\!\!\perp \eta^{(n)}$; (3) $\eta^{(m)} \perp\!\!\!\perp \mathbf{x}^{(n)}$.

Sufficient condition. Given LHS of Eq. (7), we have

$$p(\mathbf{x}^{(m)}, \mathbf{x}^{(n)} \mid \mathbf{z}^{(m)}) = p(\mathbf{x}^{(m)} \mid \mathbf{z}^{(m)}) p(\mathbf{x}^{(n)} \mid \mathbf{z}^{(m)}).$$

$$\begin{aligned}
\xRightarrow{RHS} p(\mathbf{x}^{(m)}, \mathbf{x}^{(n)} \mid \mathbf{z}^{(m)}) &= \frac{p(\mathbf{x}^{(m)}, \mathbf{x}^{(n)}, \mathbf{z}^{(m)})}{p(\mathbf{z}^{(m)})} = \frac{p(\eta^{(m)}, \eta^{(n)}, \mathbf{z}^{(m)})}{p(\mathbf{z}^{(m)})} |\det \frac{\partial \eta^{(m)}}{\partial \mathbf{x}^{(m)}}| |\det \frac{\partial \eta^{(n)}}{\partial \mathbf{x}^{(n)}}| \\
&= p(\eta^{(m)}, \eta^{(n)} \mid \mathbf{z}^{(m)}) |\det \frac{\partial \eta^{(m)}}{\partial \mathbf{x}^{(m)}}| |\det \frac{\partial \eta^{(n)}}{\partial \mathbf{x}^{(n)}}| \\
\xRightarrow{LHS} p(\mathbf{x}^{(m)} \mid \mathbf{z}^{(m)}) p(\mathbf{x}^{(n)} \mid \mathbf{z}^{(m)}) &= \frac{p(\mathbf{x}^{(m)}, \mathbf{z}^{(m)})}{p(\mathbf{z}^{(m)})} \frac{p(\mathbf{x}^{(n)}, \mathbf{z}^{(m)})}{p(\mathbf{z}^{(m)})} \\
&= \frac{p(\eta^{(m)}, \mathbf{z}^{(m)})}{p(\mathbf{z}^{(m)})} |\det \frac{\partial \eta^{(m)}}{\partial \mathbf{x}^{(m)}}| \frac{p(\eta^{(n)}, \mathbf{z}^{(m)})}{p(\mathbf{z}^{(m)})} |\det \frac{\partial \eta^{(n)}}{\partial \mathbf{x}^{(n)}}| \\
&= p(\eta^{(m)} \mid \mathbf{z}^{(m)}) p(\eta^{(n)} \mid \mathbf{z}^{(n)}) |\det \frac{\partial \eta^{(m)}}{\partial \mathbf{x}^{(m)}}| |\det \frac{\partial \eta^{(n)}}{\partial \mathbf{x}^{(n)}}|
\end{aligned}$$

Thus we have

$$p(\eta^{(m)}, \eta^{(n)}|\mathbf{z}^{(m)}) = p(\eta^{(m)}|\mathbf{z}^{(m)})p(\eta^{(n)}|\mathbf{z}^{(n)}) \Rightarrow p(\eta^{(m)}, \eta^{(n)}) = p(\eta^{(m)})p(\eta^{(n)}) \Rightarrow \eta^{(m)} \perp\!\!\!\perp \eta^{(n)}$$

Necessary condition. Given RHS of Eq. (7) and above conclusion, we have

$$p(\mathbf{x}^{(m)} \mid \mathbf{z}^{(m)}) = p(\eta^{(m)})|\det\frac{\partial \eta^{(m)}}{\partial \mathbf{x}^{(m)}}|, \quad p(\mathbf{x}^{(n)} \mid \mathbf{z}^{(m)}) = p(\eta^{(n)})|\det\frac{\partial \eta^{(n)}}{\partial \mathbf{x}^{(n)}}|$$

$$\xrightarrow{Multiplication} p(\mathbf{x}^{(m)} \mid \mathbf{z}^{(m)})p(\mathbf{x}^{(n)} \mid \mathbf{z}^{(m)}) = p(\eta^{(m)})p(\eta^{(n)})|\det\frac{\partial \eta^{(m)}}{\partial \mathbf{x}^{(m)}}||\det\frac{\partial \eta^{(n)}}{\partial \mathbf{x}^{(n)}}|$$

$$= p(\eta^{(m)}, \eta^{(n)})|\det\frac{\partial \eta^{(m)}}{\partial \mathbf{x}^{(m)}}||\det\frac{\partial \eta^{(n)}}{\partial \mathbf{x}^{(n)}}| = \frac{p(\eta^{(m)}, \eta^{(n)}, \mathbf{z}^{(m)})}{p(\mathbf{z}^{(m)})}|\det\frac{\partial \eta^{(m)}}{\partial \mathbf{x}^{(m)}}||\det\frac{\partial \eta^{(n)}}{\partial \mathbf{x}^{(n)}}| = \frac{p(\mathbf{x}^{(m)}, \mathbf{x}^{(n)}, \mathbf{z}^{(m)})}{p(\mathbf{z}^{(m)})}$$

$$\Rightarrow p(\mathbf{x}^{(m)} \mid \mathbf{z}^{(m)})p(\mathbf{x}^{(n)} \mid \mathbf{z}^{(m)}) = p(\mathbf{x}^{(m)}, \mathbf{x}^{(n)} \mid \mathbf{z}^{(m)}) \Rightarrow \mathbf{x}^{(m)} \perp\!\!\!\perp \mathbf{x}^{(n)} \mid \mathbf{z}^{(m)} \tag{8}$$

□

Proposition B.2. *[Independent Noise Condition] Denote* $\mathbf{z}$ *and* η *as the block-identified latent variables and exogenous variables across all modalities.* ϵ*'s are the causally-related noise terms. We have*

$$\eta \perp\!\!\!\perp \mathbf{z} \Longleftrightarrow \eta \perp\!\!\!\perp \epsilon. \tag{9}$$

Proof. Given the causal function in Eq. (1), we have $p(\mathbf{z}) = p(\epsilon)|\det\frac{\partial \epsilon}{\partial \mathbf{z}}|$.

Sufficient condition. Suppose $(\mathbf{z}, \eta) = h(\epsilon, \eta)$ and $\eta \perp\!\!\!\perp \mathbf{z}$, we have

$$p(\mathbf{z}, \eta) = p(\epsilon, \eta)|\text{def}\frac{\partial \epsilon}{\partial \mathbf{z}}| \Rightarrow p(\mathbf{z})p(\eta) = p(\epsilon, \eta)|\text{def}\frac{\partial \epsilon}{\partial \mathbf{z}}| \Rightarrow p(\epsilon)p(\eta)|\det\frac{\partial \epsilon}{\partial \mathbf{z}}| = p(\epsilon, \eta)|\text{def}\frac{\partial \epsilon}{\partial \mathbf{z}}|$$
$$\Rightarrow p(\epsilon)p(\eta) = p(\epsilon, \eta) \Rightarrow \eta \perp\!\!\!\perp \epsilon \tag{10}$$

Necessary condition. Suppose $(\mathbf{z}, \eta) = h(\epsilon, \eta)$ and $\eta \perp\!\!\!\perp \epsilon$, we have

$$p(\mathbf{z}, \eta) = p(\epsilon, \eta)|\text{def}\frac{\partial \epsilon}{\partial \mathbf{z}}| \Rightarrow p(\mathbf{z}, \eta) = p(\epsilon)|\text{def}\frac{\partial \epsilon}{\partial \mathbf{z}}|p(\eta) \Rightarrow p(\mathbf{z}, \eta) = p(\mathbf{z})p(\eta) \Rightarrow \eta \perp\!\!\!\perp \mathbf{z} \tag{11}$$

□

C Identifiability Theory

C.1 Proof for Theorem 4.2

We present the proof for Theorem 4.2. For ease of reference, we duplicate Condition 4.1 and Theorem 4.2 below.

Condition C.1 (Subspace Identifiability Conditions)**.**

A1 [Smoothness & Invertibility]: The generating functions $g_{\mathbf{x}^{(m)}}$ and $\tilde{g}^{(m)}$ are smooth and have smooth inverse functions.

A2 [Linear Independence]: The generating function $\tilde{g}_{\mathbf{x}^{(-m)}}$ is smooth and its Jacobian columns corresponding to $\mathbf{z}^{(m)}$ (i.e., $[\boldsymbol{J}_{\tilde{g}_{\mathbf{x}^{(-m)}}}]_{:, I(\mathbf{z}^{(m)})}$) are linearly independent almost anywhere.

Theorem 4.2 (Subspace Identifiability)**.** *Let* $\boldsymbol{\theta} := \{g_{\mathbf{x}^{(m)}}, \tilde{g}_{\mathbf{z}^{(-m)}}, p(\boldsymbol{\epsilon}^{(m)}), p(\tilde{\boldsymbol{\epsilon}}^{(-m)})\}_{m=1}^{M}$ *and* $\hat{\boldsymbol{\theta}} := \{\hat{g}_{\mathbf{x}^{(m)}}, \hat{\tilde{g}}_{\mathbf{z}^{(-m)}}, p(\hat{\boldsymbol{\epsilon}}^{(m)}), p(\hat{\tilde{\boldsymbol{\epsilon}}}^{(-m)})\}_{m=1}^{M}$ *be two specifications of the data-generating process in Eq. (3). Suppose that they generate identical observational distributions (i.e.,* $p(\mathbf{x}) = \hat{p}(\mathbf{x})$*),* $\boldsymbol{\theta}$ *satisfies Condition 4.1, and* $\hat{\boldsymbol{\theta}}$ *satisfies Condition 4.1-A1. The latent subspace* $\hat{\mathbf{z}}^{(m)}$ *for any group* m *and its counterpart* $\mathbf{z}^{(m)}$ *are equivalent up to an invertible map* $h^{(m)}(\cdot)$*, i.e.,* $\hat{\mathbf{z}}^{(m)} = h^{(m)}(\mathbf{z}^{(m)})$*.*

Proof. Given the generating processes in Eq. (2) and Eq. (1), we can express any observed group $\mathbf{x}^{(m)}$ and its complement $\mathbf{x}^{(-m)} := \mathbf{x} \setminus \mathbf{x}^{(m)}$ as two views of the latent variables of group m:

$$\mathbf{x}^{(m)} := g^{(m)}(\mathbf{z}^{(m)}, \boldsymbol{\eta}^{(m)}), \tag{12}$$

$$\mathbf{x}^{(-m)} := g^{(-m)}(\mathbf{z}^{(m)}, \tilde{\boldsymbol{\eta}}^{(-m)}), , \tag{13}$$

where $\boldsymbol{\eta}^{(m)}$ stands for exogenous variables for the group $\mathbf{x}^{(m)}$ and $\tilde{\boldsymbol{\eta}}^{(-m)}$ represents all the information necessary to generate the complement group $\mathbf{x}^{(-m)}$ beyond $\mathbf{z}^{(m)}$.

Following the classic definition of identifiability, we define two specifications $\boldsymbol{\theta} = \{g_{\mathbf{x}^{(m)}}, g_{\mathbf{z}^{(m)}}, p(\boldsymbol{\epsilon}^{(m)})\}_{m=1}^{M}$ and $\hat{\boldsymbol{\theta}} := \{\hat{g}_{\mathbf{x}^{(m)}}, \hat{g}_{\mathbf{z}^{(m)}}, \hat{p}(\boldsymbol{\epsilon}^{(m)})\}_{m=1}^{M}$ that fit the observation distribution $p(\mathbf{x})$. To show the identifiability in terms of the functions in $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$, we show that given the same $\mathbf{x}^{(m)}$ value the identifiability between $\mathbf{z}^{(m)}$ and $\hat{\mathbf{z}}^{(m)}$.

Thus, the subspace identification is equivalent to show that for each group m, the estimated latent variable $\hat{\mathbf{z}}^{(m)}$ and the true counterpart are related via an invertible map h, i.e., $\hat{\mathbf{z}}^{(m)} = h(\mathbf{z}^{(m)})$.

Eq. (12) and the invertibility of the map $(\mathbf{z}, \boldsymbol{\eta}^{(m)}, \tilde{\boldsymbol{\eta}}^{(-m)}) \mapsto (\mathbf{x}^{(m)}, \mathbf{x}^{(-m)})$ (Condition 4.1-A1) give rise to an invertible map $\tilde{h} : (\hat{\mathbf{z}}^{(m)}, \hat{\boldsymbol{\eta}}^{(m)}, \hat{\tilde{\boldsymbol{\eta}}}^{(-m)}) \mapsto (\mathbf{z}^{(m)}, \boldsymbol{\eta}^{m}, \tilde{\boldsymbol{\eta}}^{(-m)})$.

The matched observed distribution between the true and the estimated models for the generating process Eq. (13) yields that

$$g^{(-m)}(\mathbf{z}^{(m)}, \tilde{\boldsymbol{\eta}}^{(-m)}) = \hat{g}^{(-m)}(\hat{\mathbf{z}}^{(m)}, \hat{\tilde{\boldsymbol{\eta}}}^{(-m)}). \tag{14}$$

Plugging in $\tilde{h}$ gives

$$\hat{g}^{(-m)}(\hat{\mathbf{z}}^{(m)}, \hat{\tilde{\boldsymbol{\eta}}}^{(-m)}) = g^{(-m)}\left(\left[\tilde{h}\left(\hat{\mathbf{z}}^{(m)}, \hat{\boldsymbol{\eta}}^{(m)}, \hat{\tilde{\boldsymbol{\eta}}}^{(-m)}\right)\right]_{I(\mathbf{z}^{(m)}), I(\tilde{\boldsymbol{\eta}}^{(-m)})}\right). \tag{15}$$

where we adopt $I(\cdot)$ to indicate the indices of its argument.

For any $i \in [d(\mathbf{x}^{(m)})]$ and $j \in [d(\hat{\boldsymbol{\eta}}^{(m)})]$, we take partial derivative w.r.t. $\hat{\eta}_j^{(m)}$ on both sides of Eq. (15):

$$\underbrace{\frac{\partial[\hat{g}^{(-m)}]_i}{\partial[\hat{\eta}^{(m)}]_j}}_{=0} = \frac{\partial[g^{(-m)}]_i}{\partial[\hat{\eta}^{(m)}]_j}. \tag{16}$$

The left-hand side of Eq. (15) equals to zero because $\hat{g}^{(-m)}$ is not a function of $\hat{\boldsymbol{\eta}}^{(m)}$.

Therefore, expanding the right-hand side of Eq. (15) gives:

$$\sum_{k \in I(\mathbf{z}^{(-m)}) \cup I(\tilde{\boldsymbol{\eta}}^{(-m)})} \frac{\partial[g^{(-m)}]_i}{\partial[\tilde{h}]_k} \cdot \frac{\partial[\tilde{h}]_k}{\partial[\hat{\eta}^{(m)}]_j} = \sum_{k \in I(\mathbf{z}^{(-m)})} \frac{\partial[g^{(-m)}]_i}{\partial[\tilde{h}]_k} \cdot \frac{\partial[\tilde{h}]_k}{\partial[\hat{\eta}^{(m)}]_j} = 0. \tag{17}$$

The first equality in Eq. (17) is due to the fact that $\tilde{\boldsymbol{\eta}}^{(-m)}$ is a function of $\mathbf{x}^{(-m)}$ and varying $\hat{\boldsymbol{\eta}}^{(m)}$ doesn't vary $\mathbf{x}^{(-m)}$ ($\hat{\boldsymbol{\eta}}^{(m)}$ is a function of $\mathbf{x}^{(m)}$ thanks to the invertibility of $\hat{g}^{(m)}$), i.e., $\frac{\partial[\tilde{\eta}^{(-m)}]_k}{\partial[\hat{\eta}^{(m)}]_j} = 0$.

Condition 4.1-A2 implies that the matrix $\left(\frac{\partial[g^{(-m)}]_i}{\partial[\tilde{h}]_k}\right)_{i,k}$ has a full column rank. Therefore, its null space contains only a zero vector, which, together with Eq. (17), implies that $\frac{\partial[z^{(m)}]_k}{\partial[\hat{\eta}^{(m)}]_j} = 0$. Consequently, given the generating process Eq. (12) and the invertibility of $g^{(m)}$ and $\hat{g}^{(m)}$ (Condition 4.1-A1), the estimated latent variable $\hat{\mathbf{z}}^{(m)}$ and the true latent variable $\mathbf{z}^{(m)}$ are related via an invertible map, as desired.

□

C.2 Proof for Theorem 4.4

We present the proof for Theorem 4.4. For ease of reference, we duplicate Condition 4.3 and Theorem 4.4.

Condition C.2 (Component Identifiability Conditions)**.** Over the domain of $(\mathbf{z}, \boldsymbol{\epsilon})$, for any modality m, nonempty $C^{(m)} \subset U^{(m)}$, nonempty $R^{(m)} \subset D^{(m)}$, and $\boldsymbol{T} = \operatorname{diag}(\boldsymbol{T}_1, \dots, \boldsymbol{T}_{M-1})$ with invertible $\boldsymbol{T}_i \in \mathbb{R}^{d(\mathbf{z}^{(i)}) \times d(\mathbf{z}^{(i)})}$, we have

$$\begin{aligned} \left| \bigcup_{j \in I(C^{(m)})} \operatorname{Supp}([\boldsymbol{T}\boldsymbol{G}]_{(-m),j}) \right| - d^* \left([\boldsymbol{T}\boldsymbol{G}]_{(-m),I(C^{(m)})} \right) &> \max_{j \in I(C^{(m)})} \left\| [\boldsymbol{G}]_{(-m),j} \right\|_0; \\ \left| \bigcup_{j \in I(R^{(m)})} \operatorname{Supp}(([\boldsymbol{G}\boldsymbol{T}^{-1}]_{j,(-m)})^\top) \right| - d^* \left(([\boldsymbol{G}\boldsymbol{T}^{-1}]_{I(R^{(m)}),(-m)})^\top \right) &> \max_{j \in I(R^{(m)})} \left\| [\boldsymbol{G}]_{j,(-m)} \right\|_0. \end{aligned} \tag{4}$$

Theorem 4.4 (Component-wise Identifiability)**.** *Let $\boldsymbol{\theta} := (\{g_{\mathbf{x}^{(m)}}, g_{\mathbf{z}^{(m)}}, p(\boldsymbol{\epsilon}^{(m)})\}_{m=1}^M)$ and $\hat{\boldsymbol{\theta}} := (\{\hat{g}_{\mathbf{x}^{(m)}}, \hat{g}_{\mathbf{z}^{(m)}}, \hat{p}(\boldsymbol{\epsilon}^{(m)})\}_{m=1}^M)$ be two specifications of the data-generating process in Eq. (1) and Eq. (2). Suppose that they generate identical observational distributions (i.e., $p(\mathbf{x}) = \hat{p}(\mathbf{x})$) and $\boldsymbol{\theta}$ satisfies Condition 4.1 and Condition 4.3. If $\hat{\boldsymbol{\theta}}$ satisfies the following condition:*

$$\sum_{m \neq n \in [M]} \left\| [\boldsymbol{J}_{\hat{g}_{\mathbf{z}}}]_{(m),(n)} \right\|_0 \leq \sum_{m \neq n \in [M]} \left\| [\boldsymbol{J}_{g_{\mathbf{z}}}]_{(m),(n)} \right\|_0, \tag{5}$$

each component $z_i^{(m)}$ and its counterpart $\hat{z}_{\pi(i)}^{(m)}$ are equivalent up to an invertible map $h(\cdot)$, i.e., $\hat{z}_{\pi(i)}^{(m)} = h(z_i^{(m)})$ under a permutation π over $[d(\mathbf{z}^{(m)})]$.

Condition 4.3 stipulates sparse cross-modality causal connections among latent components $\mathbf{z}$. Under this condition, when one latent component $\hat{z}_i^{(m)}$ is a function of two components $\hat{z}_j^{(m)}$ and $\hat{z}_k^{(m)}$ (when component-wise identification breaks), the cross-modality causal connections in $\boldsymbol{G}$ are guaranteed to be denser than those in $\hat{\boldsymbol{G}}$. Therefore, the sparsity control enforces us to select the sparest estimated models, in which one latent component $\hat{z}_i^{(m)}$ is a function of a unique component $z_j^{(m)}$, yielding the desired component-wise identifiability.

Proof. Given Theorem 4.2, Condition 4.1 implies that the estimated group-wise latent variable $\hat{\mathbf{z}}^{(m)}$ is related to the true variable $\mathbf{z}^{(m)}$ through an invertible transformation $h^{(m)}$, i.e.,

$$\hat{\mathbf{z}}^{(m)} = h^{(m)}(\mathbf{z}^{(m)}). \tag{18}$$

It follows that the Jacobian matrix $\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}$ can be arranged into a block-diagonal matrix, in which diagonal block m corresponds to a Jacobian matrix $\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}^{(m)}}{\partial \mathbf{z}^{(m)}}}$. Then, the goal is to prove that these diagonal blocks are actually generalized permutation matrices, whose each column only contains one nonzero entry.

We divide the proof into several steps for the sake of exposition. At step 1, we derive an equivalence relation between the estimation model $(\hat{g}_z, \hat{g}_x)$ and the true model (g_z, g_x). At step 2, we apply Theorem 4.2 to the equivalence to characterize the relation between the true and the estimated graph structure. At step 3 and 4, we leverage the sparsity condition (Condition 4.3) to reason about the identifiability of each component $z_i^{(m)}$ for $m \in [M]$ and $i \in [d(z^{(m)})]$.

Step 1. The generating process in Eq. (1) and the subspace identification Eq. (18) imply

$$\hat{g}_z(\hat{\mathbf{z}}, \hat{\boldsymbol{\epsilon}}) = h \circ g_z(\mathbf{z}, \boldsymbol{\epsilon}), \tag{19}$$

where h is defined as the Cartesian product of individual $h^{(m)}$ functions.

Taking partial derivatives w.r.t, z_i of both sides of Eq. (19) yields:

$$\begin{bmatrix} \boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}} & \boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\boldsymbol{\epsilon}}}} \end{bmatrix} \begin{bmatrix} \boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}} \\ \boldsymbol{T}_{\frac{\partial \hat{\boldsymbol{\epsilon}}}{\partial \mathbf{z}}} \end{bmatrix} = \boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}} \boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}. \tag{20}$$

Each $\boldsymbol{T}$ matrix is the Jacobian matrix consisting of the corresponding partial derivatives. We use $\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}$ to denote the derivatives from the function g_z which encodes the dependence structure among z components. The same applies to $\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}$. As discussed above, the matrix $\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}$ has a block-diagonal structure (after proper permutations) with block m corresponding to the Jacobian matrix of $h^{(m)}$. Moreover, the matrix $\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \epsilon}}$ is strictly diagonal due to the generating function Eq. (1).

Step 2. In this step, we simplify Eq. (20) to derive the relation between the estimated graph structures and true graph structures encoded in $\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}$ and $\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}$ respectively.

First, we note that the $\boldsymbol{T}_{\frac{\partial \hat{\epsilon}}{\partial \mathbf{z}}}$ is also block-diagonal w.r.t. the groups. To see this, we compute the partial derivatives therein as follows: $\frac{\partial \hat{\epsilon}_i^{(m)}}{\partial z_j^{(n)}} = \frac{\partial \hat{\epsilon}_i^{(m)}}{\partial \hat{\tilde{z}}_i^{(m)}} \frac{\partial \hat{\tilde{z}}_i^{(m)}}{\partial z_j^{(n)}}$, where we denote that output of $\hat{g}_z$ with $\hat{\tilde{z}}$ in the derivative. Due to the equivalent relation $\mathbf{z} = \tilde{\mathbf{z}}$ (Eq. (1)), we have $\frac{\partial \hat{\tilde{z}}_i^{(m)}}{\partial z_j^{(n)}} = \frac{\partial \hat{z}_i^{(m)}}{\partial z_j^{(n)}}$ which is zero for distinct groups $m \neq n$ (Eq. (18)). It follows that

$$\frac{\partial \hat{\epsilon}_i^{(m)}}{\partial z_j^{(n)}} = 0,\ m \neq n. \tag{21}$$

Therefore, we have shown that $\boldsymbol{T}_{\frac{\partial \hat{\epsilon}}{\partial \mathbf{z}}}$ is block-diagonal w.r.t. the groups.

This structure allows us to simplify Eq. (20) to directly characterize the relation between the two graphical structures $\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}$ and $\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}$. In particular, since $\boldsymbol{T}_{\frac{\partial \hat{\epsilon}}{\partial \mathbf{z}}}$ is block-diagonal and $\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \epsilon}}$ is diagonal, the off-diagonal blocks on the left-hand side of Eq. (20) are determined by $\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}} \boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}$. Therefore, it follows from Eq. (20):

$$\left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}} \boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(m),(n)} = \left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}} \boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(m),(n)},\ m \neq n, \tag{22}$$

where we adopt subscripts (m) to denote the block for group m.

On account of the block-diagonal structure of $\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}$, the left-hand side of Eq. (22) can be expressed as follows:

$$\left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}} \boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(m),(n)} = \left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}\right]_{(m),:} \left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{:,(n)} = \left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}\right]_{(m),(n)} \left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(n),(n)}. \tag{23}$$

Analogously, the right-hand side of Eq. (22) can be expressed as:

$$\left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}} \boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(m),(n)} = \left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(m),:} \left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{:,(n)} = \left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(m),(m)} \left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(m),(n)}. \tag{24}$$

It follows from Eq. (22), Eq. (23), and Eq. (24) that

$$\begin{aligned} &\left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}\right]_{(m),(n)} \left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(n),(n)} = \left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(m),(m)} \left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(m),(n)} \\ &\Longrightarrow \\ &\left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}\right]_{(m),(n)} = \left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(m),(m)} \left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(m),(n)} \left[\boldsymbol{T}_{\frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}}\right]_{(n),(n)}. \end{aligned} \tag{25}$$

Eq. (25) relates the true off-diagonal ($m \neq n$) structure $\left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(m),(n)}$ and its estimated counterpart $\left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}\right]_{(m),(n)}$.

Step 3. We now reason about the component-wise identifiability within each group through the sparsity of the off-diagonal regions. Following Eq. (25) and the block-diagonal structures of $\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}$ and $\boldsymbol{T}_{\frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}}$, we can express the functional influence from any group (m) to the other groups $(-m)$ as

$$\left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}\right]_{(-m),m} = \left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(-m),:} \left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{:,(m)} \left[\boldsymbol{T}_{\frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}}\right]_{(m),(m)} \tag{26}$$

$$= \left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(-m),(-m)} \left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(-m),(m)} \left[\boldsymbol{T}_{\frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}}\right]_{(m),(m)} = \left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}} \boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(-m),(m)} \left[\boldsymbol{T}_{\frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}}\right]_{(m),(m)}. \tag{27}$$

We note that the matrix $\left[\boldsymbol{T}_{\frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}}\right]_{(m),(m)}$ is the Jacobian matrix of the inverse of the invertible map $h^{(m)}$ defined in Eq. (18). Its invertibility implies that there exists a permutation $\sigma : I(m) \to I(m)$ over group m's component indices $I(m) \subset [d(\mathbf{z})]$, such that for any $i \in [d(\mathbf{z}^{(m)})]$, we have $\left[\boldsymbol{T}_{\frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}}\right]_{i,\hat{i}} \neq 0$ where $\hat{i} = \sigma(i)$.

To show the component-wise identifiability, for any component $\hat{i} \in I(m)$, we would like to show that $\left[\boldsymbol{T}_{\frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}}\right]_{i,\hat{i}}$ is the only nonzero entry in the column $\left[\boldsymbol{T}_{\frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}}\right]_{(m),\hat{i}}$.

We denote components in modality m that have children in other modalities $-m$ as upstream variables $U^{(m)}$, i.e., $\left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(-m),i(z)} \neq \mathbf{0}$ for any $z \in U^{(m)}$, where we denote the index of the component z with $i(z) \in [d(\mathbf{z})]$. Let $C_i^{(m)}$ be the largest subset of $U^{(m)}$ such that $\left[\boldsymbol{T}_{\frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}}\right]_{C_i^{(m)},\hat{i}}$ is component-wise nonzero and we would like to show that $C_i^{(m)}$ does not contain other components than $z_i^{(m)}$.

We proceed by contradiction. Suppose that $C_i^{(m)}$ contains components other than $z_i^{(m)}$. For any $\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_2}$, define $d^*(\boldsymbol{A}) := \max\{|R| : R \subset [d_1], \text{rank}\left([\text{Overlap}(\boldsymbol{A})]_{R,:}\right) < |d_2|\}$, where $d^*(\boldsymbol{A})$ is the maximal number of non-zero entries that can be canceled out by linearly combining its columns. It follows from Equation (27) that

$$\begin{aligned}
\left\|\left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}\right]_{(-m),\hat{i}}\right\|_0 &= \left\|\left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(-m),(m)}\left[\boldsymbol{T}_{\frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}}\right]_{(m),\hat{i}}\right\|_0 \\
&= \left\|\left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(-m),I(C_i^{(m)})\cup\{i\}}\left[\boldsymbol{T}_{\frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}}\right]_{I(C_i^{(m)})\cup\{i\},\hat{i}}\right\|_0 \\
&\geq \left|\bigcup_{j\in I(C_i^{(m)})\cup\{i\}} \text{Supp}\left(\left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(-m),j}\right)\right| - d^*\left(\left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(-m),I(C_i^{(m)})\cup\{i\}}\right).
\end{aligned} \tag{28}$$

If $i \in U^{(m)}$, we have $I(C_i^{(m)}) \cup \{i\} = I(C_i^{(m)})$ and $\left|C_i^{(m)}\right| > 1$. It follow from Eq. 28 and Condition 4.3 that

$$\begin{aligned}
\left\|\left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}\right]_{(-m),\hat{i}}\right\|_0 &\geq \left|\bigcup_{j\in I(C_i^{(m)})} \text{Supp}\left(\left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(-m),j}\right)\right| - d^*\left(\left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(-m),I(C_i^{(m)})}\right). \\
&\underbrace{>}_{\text{Condition 4.3}} \max_{j\in I(C_i^{(m)})}\left\|\left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(-m),j}\right\|_0 \geq \left\|\left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(-m),i}\right\|_0.
\end{aligned} \tag{29}$$

If $i \notin U^{(m)}$, the strict inequality in Eq. (29) holds true trivially, as $\left|C_i^{(m)}\right| \geq 1$ and $\left\|\left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(-m),i}\right\|_0 = 0$ by definition.

Note that the above reasoning holds for all $z_i^{(m)} \in \mathbf{z}^{(m)}$ for all $m \in [M]$. That is, if any $C_i^{(m)}$ contained components other than $z_i^{(m)}$, we would have

$$\left\|\left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}\right]_{(-m),\hat{i}}\right\|_0 > \left\|\left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(-m),i}\right\|_0. \tag{30}$$

Also note that in general we have a lower bound $\left\| \left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}} \right]_{(-m),\hat{i}} \right\|_0 \geq \left\| \left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}} \right]_{(-m),i} \right\|_0$ and the equality is attained when $\boldsymbol{T}_{\frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}}$ is a generalized permutation matrix (each row/column contain only one nonzero element).

Therefore, the strict inequality in Eq. (29) implies that if any $C_i^{(m)}$ contained components other than $z_i^{(m)}$, we would have

$$\sum_{m \neq n \in [M]} \left\| \left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}} \right]_{(m),(n)} \right\|_0 > \sum_{m \neq n \in [M]} \left\| \left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}} \right]_{(m),(n)} \right\|_0, \tag{31}$$

which would violate the sparsity constraint Eq. (5).

Therefore, we have shown that $C_i^{(m)}$ cannot contain components other than $z_i^{(m)}$. We conclude the element $(i, \hat{i})$ is the unique nonzero element in column $\left[\boldsymbol{T}_{\frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}} \right]_{I(U^{(m)}),\hat{i}}$. That is, the component $\hat{z}_{\hat{i}}^{(m)}$ cannot functionally influence components in $U^{(m)}$ other than $\mathbf{z}_i^{(m)}$.

Analogously, define the set of downstream variables $D^{(m)} \subseteq \mathbf{z}^{(m)}$ that possess parents in other groups $\left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}} \right]_{i(z),(-m)} \neq \mathbf{0}$ for $z \in D^{(m)}$. The same argument yields that the component $\hat{z}_{\hat{i}}^{(m)}$ cannot be functionally influenced by components in $D^{(m)}$ other than $\{z_i^{(m)}\}$.

Step 4. Suppose that $\hat{z}_{\hat{i}}^{(m)}$ is a function of component $z_j \in U^{(m)} \setminus D^{(m)}$ that influences other groups but is not influenced by others (i.e., a source variable). That is, we have that $\frac{\partial \hat{z}_{\hat{i}}^{(m)}}{\partial z_j^{(m)}}(\mathbf{z}_*^{(m)}) \neq 0$ for some $\mathbf{z}_*^{(m)}$. Without loss of generality, we suppose that $\frac{\partial \hat{z}_{\hat{i}}^{(m)}}{\partial z_j^{(m)}}(\mathbf{z}_*^{(m)}) \neq 0$ for some $\mathbf{z}_*^{(m)}$. Due to the smoothness of $h^{(m)}$, there exists an open line segment $([z_1^{(m)}, \ldots, z_j^{(m)} - l, \ldots, z_{d(\mathbf{z}^{(m)})}], [z_1^{(m)}, \ldots, z_j^{(m)} + l, \ldots, z_{d(\mathbf{z}^{(m)})}])$ for some $l > 0$, over which $\frac{\partial \hat{z}_{\hat{i}}^{(m)}}{\partial z_j^{(m)}} > 0$ and thus the map $z_j^{(m)} \mapsto \hat{z}_{\hat{i}}^{(m)}$ is monotonic and invertible over $(z_j^{(m)} - l, z_j^{(m)} + l)$. Therefore, there exists a monotonic map $\hat{z}_{\hat{i}}^{(m)} \mapsto z_j^{(m)}$ over the image of $(z_j^{(m)} - l, z_j^{(m)} + l)$. That is, we have the partial derivative $\frac{\partial z_j^{(m)}}{\partial \hat{z}_{\hat{i}}^{(m)}} > 0$. This is impossible because the component $\hat{z}_{\hat{i}}^{(m)}$ cannot functionally influence components in $U^{(m)}$. This contradiction implies that $\hat{z}_{\hat{i}}^{(m)}$ cannot be a function of components $U^{(m)} \setminus D^{(m)}$ either. Overall, we have derived that $\hat{z}_{\hat{i}}^{(m)}$ cannot be a function of components $U^{(m)} \cup D^{(m)} = \mathbf{z}^{(m)} \setminus \{z_i^{(m)}\}$. Therefore, we have a bijection $\hat{z}_{\hat{i}}^{(m)} = h_{\hat{i}}^{(m)}(z_i^{(m)})$. Since this holds for any group m and any component i, we have arrived at the desired conclusion.

□

C.3 Extended Theorem 4.2 and its Proof

We restate Theorem C.7 from Yao et al. (2023), which we invoke in our Theorem C.9. We drop the entropy regularization term in Yao et al. (2023), since we assume the invertibility of estimated functions $\hat{g}^{(m)}$ directly.

Definition C.3 (View-Specific Encoders)**.** The *view-specific encoders* $R := \{r_k : \mathcal{X}_k \to \mathcal{Z}_{S_k}\}_{k \in V}$ consist of smooth functions mapping from the respective observation spaces to the view-specific latent space, where the dimension of the kth latent space $|S_k|$ is assumed known for all $k \in V$.

Definition C.4 (Selection)**.** A selection $\oslash$ operates between two vectors $a \in \{0,1\}^d, b \in \mathbb{R}^d$ s.t.

$$a \oslash b := [b_j : a_j = 1, j \in [d]]$$

Definition C.5 (Content Selectors)**.** The content selectors $\Phi := \{\phi(i,k)\}_{V_i \in \mathcal{V}, k \in V_i}$ with $\phi^{(i,k)} \in \{0,1\}^{|d(\mathbf{z}^{(m)})|}$ perform selection C.4 on the encoded information: for any subset $V_i \subset [M]$ and view

$k \in V_i$ we have the selected representation: $\phi(i,m) \oslash \hat{\mathbf{z}}^{(m)}$ with $\left\|\phi^{(i,k)}\right\|_0 = \left\|\phi^{(i,k')}\right\|_0$ for all $V_i \in \mathcal{V}, k, k' \in V_i$.

Definition C.6 (Information-Sharing Regularizer)**.** The following regularizer penalizes the ℓ_0-norm $\|\cdot\|_0$ of the content selectors Φ: $\mathrm{Reg}(\Phi) := -\sum_{V_i \in \mathcal{V}} \sum_{k \in V_i} \left\|\phi^{(i,k)}\right\|_0$.

Theorem C.7 (View-Specific Encoder for Identifiability (Yao et al., 2023))**.** *Let $R := \{\hat{g}_{(m)}\}_{m=1}^{M}$ and Φ respectively be the generating functions and content selectors (Definition C.5) that solve the following constrained optimization problem:*

$$\min \mathrm{Reg}(\Phi) \qquad \textit{subject to:} \qquad R, \Phi \in \arg\min \mathcal{L}_{\textit{alignment}}(R, \Phi), \tag{32}$$

where

$$\mathcal{L}_{\textit{alignment}}(R, \Phi) = \sum_{V_i \in \mathcal{V}} \sum_{\substack{m_1, m_2 \in V_i \\ k < k'}} \mathbb{E}\left[\left\|\phi(i, m_1) \oslash [\hat{g}^{(m_1)}]^{-1}(\mathbf{x}_k) - \phi(i, m_2) \oslash [\hat{g}^{m_2}]^{-1}(\mathbf{x}_{m_2})\right\|_2\right] \tag{33}$$

Then for any subset of modalities $V_i \subset [M]$ and any modality $m \in V_i$, $\phi(i,m) \oslash [\hat{g}^{(m)}]^{-1}$ identifies the shared subspace $\mathbf{z}^{\left(\cap_{m \in V_i} m\right)}$.

Definition C.8 (Reconstruction Loss)**.** The following loss penalizes the deviation of the estimate $\hat{\mathbf{x}}$ and its corresponding true counterpart $\mathbf{x}$ in ℓ_2 $L_{\text{recons}} := \mathbb{E}_{\mathbf{x}}(\mathbf{x} - \hat{\mathbf{x}})$.

Theorem C.9 (Generalized Subspace Identifiability)**.** *We estimate the generating process in Eq. (3) with model $\{(\hat{g}_{\mathbf{x}^{(m)}}, \hat{\tilde{g}}_{\mathbf{x}^{(-m)}})\}_{m=1}^{M}$ under additional terms in Eq. (32).*

$$\min \mathrm{Reg}(\Phi) \qquad \textit{subject to:} \qquad R, \Phi \in \arg\min \mathcal{L}_{\textit{alignment}} + \mathcal{L}_{\textit{generation}}. \tag{34}$$

Under Condition 4.1, the estimated latent variable $\hat{\mathbf{z}}^{(m)}$ for any group m and its true counterpart $\mathbf{z}^{(m)}$ are equivalent up to an invertible map $h^{(m)}(\cdot)$, i.e., $\hat{\mathbf{z}}^{(m)} = h^{(m)}(\mathbf{z}^{(m)})$.

Proof. We note that the latent model with shared latent variables across modalities can still be cast into Equation (3) and satisfies Condition 4.1. As a consequence, Theorem 4.2 gives us the subspace identification for each modality as in the disjoint case. Moreover, we can identify any blocks among modalities thanks to Theorem C.7. This concludes the proof.

□

C.4 Extended Theorem 4.4 and its Proof

Additional notations and discussion. We slightly abuse the notation to denote both sets and vectors with bold symbols $\mathbf{z}$. Let $\mathbf{z}^{(m \cap n)}$ be the set of latent components shared by modality m and n, i.e., $\mathbf{z}^{(m \cap n)} := \mathbf{z}^{(m)} \cap \mathbf{z}^{(n)}$. Analogously, let $\mathbf{z}^{(m \setminus n)}$ be the set of latent components in modality m that are not shared by n, i.e., $\mathbf{z}^{(m \setminus n)} := \mathbf{z}^{(m)} \setminus \mathbf{z}^{(n)}$.

The participation of multiple modalities requires a new definition of the shared blocks in $\mathbf{z}$ since the sharing structure could be nested and various numbers of modalities could share one partition. We partition the entire latent space $\mathbf{z}$ into disjoint blocks $\{\mathbf{z}^{(b)}\}_{b \in B}$, whose components z have exactly the same modality membership $\mathcal{M}(z) := \{m \in [M] : z \in \mathbf{z}^{(m)}\}$. We define the $\mathbf{z}^{(H(b))}$ as the smallest (the least components) identified partition in $\mathbf{z}$ that contains $\mathbf{z}^{(b)}$. In the two-modal case, we have $B = \{(m \cap n), (m \setminus n), (n \setminus m)\}$ and $\mathbf{z}^{(H(m \cap n))} = \mathbf{1}_{|\mathbf{z}^{(m)}| \leq |\mathbf{z}^{(n)}|} \mathbf{z}^{(m)} + \mathbf{1}_{|\mathbf{z}^{(m)}| > |\mathbf{z}^{(n)}|} \mathbf{z}^{(n)}$.

We denote $\mathbf{z}^{(b_1)} \prec \mathbf{z}^{(b_2)}$ if block $\mathbf{z}^{(b_1)}$ is shared by a strict subset of modalities that share $\mathbf{z}^{(b_2)}$, i.e., $\mathcal{M}(\mathbf{z}^{(b_1)}) \subsetneq \mathcal{M}(\mathbf{z}^{(b_2)})$. Therefore, we have either $\mathbf{z}^{(H(b))} = \mathbf{z}^{(b)}$ (it is identifiable itself) or $\mathbf{z}^{(b)} \prec \mathbf{z}^{(H(b))} \setminus \mathbf{z}^{(b)}$ (it forms an identifiable block with a more deeply shared block). We denote former blocks as $b^+ \in B^+ \subset B$, i.e., $\mathbf{z}^{(H(b^+))} = \mathbf{z}^{(b^+)}$, and the latter blocks as $b^- \in B^- = B \setminus B^+$. In the two-modal case, we have $\mathbf{z}^{(m \setminus n)} \prec \mathbf{z}^{(m \cap n)}$ and $\mathbf{z}^{(n \setminus m)} \prec \mathbf{z}^{(m \cap n)}$, and $B^+ = \{(m \cap n)\}$ and $B^- = \{(m \setminus n), (n \setminus m)\}$.

We note that all shared blocks $\mathbf{z}^{(b^+)}$ are identified, and thus their bijective indeterminacies are w.r.t, themselves, i.e., $\mathbf{z}^{(b^+)} \mapsto \hat{\mathbf{z}}^{(b^+)}$, which implies the square shape of their indeterminacy matrices $[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}]_{(b^+),(b^+)}$ (as both downstream and upstream variables on both sides of the true graph matrix, e.g., Eq. (40)). In contrast, the unidentifiable blocks $\mathbf{z}^{(b^-)}$ can potentially receive the influence from all components in its modality $\mathbf{z}^{(H(b^-))}$. However, they do not influence the complement block $\mathbf{z}^{(H(b^-))} \setminus \mathbf{z}^{(b^{(-1)})}$ in its modality, since $\mathbf{z}^{(b^-)} \prec \mathbf{z}^{(H(b^-))} \setminus \mathbf{z}^{(b^-)}$. Consequently, their indeterminacy matrices are $[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}]_{(b^-),(H(b^-))}$ (as downstream variables at the right-side of the true graph matrix, e.g., Eq. (43)) and $[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}]_{(b^-),(b^-)}$ (as upstream variables at the right-side of the true graph matrix, e.g., Eq. (43)).

The indeterminacy matrix $\boldsymbol{T} := \boldsymbol{T}_{\text{on}} + \boldsymbol{T}_{\text{off}}$ is not strictly block-diagonal anymore. The matrix $\boldsymbol{T}_{\text{on}}$ contains all the on-diagonal square matrices $\boldsymbol{T}_{\text{on}} := \begin{bmatrix} \boldsymbol{T}_{b_1} & & & \\ & \boldsymbol{T}_{b_2} & & \\ & & \dots & \\ & & & \boldsymbol{T}_{b_{|B|}} \end{bmatrix}$, where each $\boldsymbol{T}_b$ is an invertible matrix. The matrix $\boldsymbol{T}_{\text{off}}$ contains all the off-diagonal nonzero elements such that $\text{Supp}(\boldsymbol{T}_{\text{on}}) \cap \text{Supp}(\boldsymbol{T}_{\text{off}}) = \emptyset$ and $\boldsymbol{T} := \begin{bmatrix} \tilde{\boldsymbol{T}}_{b_1} & & & \\ & \tilde{\boldsymbol{T}}_{b_2} & & \\ & & \dots & \\ & & & \tilde{\boldsymbol{T}}_{b_{|B|}} \end{bmatrix}$, where each $\tilde{\boldsymbol{T}}_b$ is of shape $d(\mathbf{z}^{(b)}) \times d(\mathbf{z}^{H(b)} \setminus \mathbf{z}^{(b)})$. Rows of distinct blocks b_1 and b_2 are disjoint, but their columns might due to the shared variables between their corresponding $H(b_1)$ and $H(b_2)$. We denote a set of blocks $E(b)$ whose memberships are not a subset of those of block b as: $E(b) := B \setminus \{\tilde{b} \in B | \mathbf{z}^{(\tilde{b})} \prec \mathbf{z}^{(b)}\}$.

With these notations, we state the generalized component identification result in Theorem C.9.

Condition C.10 (Generalized Component Identifiability Conditions)**.** Over the domain of $(\mathbf{z}, \boldsymbol{\epsilon})$, for any nonempty $C^{(m)} \subset b$, nonempty $R^{(m)} \subset H(b)$ with $b \in B$, and $\boldsymbol{T}$, sub-matrices $[\boldsymbol{T}\boldsymbol{G}]_{E(b),C^{(b)}}$ and $\left([\boldsymbol{G}\boldsymbol{T}_{\text{on}}^{-1}]_{R^{(b)},E(b)}\right)^\top$ satisfy:

$$\begin{aligned} &\left| \bigcup_{j \in C^{(b)}} \text{Supp}([\boldsymbol{T}\boldsymbol{G}]_{E(b),j}) \right| - d^*\left([\boldsymbol{T}\boldsymbol{G}]_{E(b),C^{(b)}}\right) > \max_{j \in C^{(b)}} \left\| [\boldsymbol{G}]_{E(b),j} \right\|_0; \\ &\left| \bigcup_{j \in R^{(b)}} \text{Supp}(\left([\boldsymbol{G}\boldsymbol{T}_{\text{on}}^{-1}]_{j,E(b)}\right)^\top) \right| - d^*\left(\left([\boldsymbol{G}\boldsymbol{T}_{\text{on}}^{-1}]_{R^{(b)},E(b)}\right)^\top\right) > \max_{j_1 R^{(b)}} \left\| [\boldsymbol{G}]_{j,E(b)} \right\|_0. \end{aligned} \tag{35}$$

Proof. This proof closely follows that of Theorem 4.4. We illustrate the key discrepancies as follows.

We start with only two modalities $\mathbf{z}^{(m)}$ and $\mathbf{z}^{(m)}$ for simplicity and then move on to general cases.

The structure of the indeterminacy matrix $\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}$. Identical to Equation 20, we have the relationship between Jacobian matrices:

$$\begin{bmatrix} \boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \tilde{\mathbf{z}}}} & \boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \tilde{\boldsymbol{\epsilon}}}} \end{bmatrix} \begin{bmatrix} \boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}} \\ \boldsymbol{T}_{\frac{\partial \tilde{\boldsymbol{\epsilon}}}{\partial \mathbf{z}}} \end{bmatrix} = \boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \tilde{\mathbf{z}}}} \boldsymbol{G}_{\frac{\partial \tilde{\mathbf{z}}}{\partial \mathbf{z}}}. \tag{36}$$

The presence of the shared block $\mathbf{z}^{(m \cap n)}$ alters the indeterminacy matrix $\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}$ – instead of the disjoint diagonal-block shape, $\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}$, the columns belonging to the shared variables $\mathbf{z}^{(m,n)}$ (shared between two modalities) are possibly nonzero over rows belonging to $\mathbf{z}^{(m \cap n)}$. That is, the shared variables $\mathbf{z}^{(m \cap n)}$ can still mix in the estimates of the two individual parts $\hat{\mathbf{z}}^{(m \setminus n)}$ and $\hat{\mathbf{z}}^{(n \setminus m)}$. However, since we have identified the subspace of $\mathbf{z}^{(m \cap n)}$, its estimates would not contain information of the individual blocks $\mathbf{z}^{(m \setminus n)}$ and $\mathbf{z}^{(n \setminus m)}$, rendering the blocks $\frac{\partial \hat{\mathbf{z}}^{(m \cap n)}}{\partial \mathbf{z}^{(m \setminus n)}} = 0$ and $\frac{\partial \hat{\mathbf{z}}^{(m \cap n)}}{\partial \mathbf{z}^{(n \setminus m)}} = 0$.

The sparse connection among modalities. The reasoning in **Step 2** in the proof of Theorem 4.4 implies that the structure of the matrix $\boldsymbol{T}_{\frac{\partial \hat{\boldsymbol{\epsilon}}}{\partial \mathbf{z}}}$ is consistent with that of the matrix $\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}$. That is, they have zero block matrices at the same positions. In particular, since the subspace identifiability in Theorem C.9 implies that the estimated shared variable $\hat{\mathbf{z}}^{(m\cap n)}$ and the modality-specific variable $\hat{\mathbf{z}}^{(n\setminus m)}$ are not influenced by that the other modality-specific variable $\mathbf{z}^{(m\setminus n)}$, the same applies to the estimated exogenous variable $\hat{\boldsymbol{\epsilon}}^{(m\cap n)}$ and $\hat{\boldsymbol{\epsilon}}^{(m\setminus n)}$. This structure permits us to disregard and $\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \boldsymbol{\epsilon}}}$ (an identity matrix) and $\boldsymbol{T}_{\frac{\partial \hat{\boldsymbol{\epsilon}}}{\partial \mathbf{z}}}$ on the left-hand side of Eq. (36) when computing a sub-matrix of the right-hand side product:

$$\left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(n),(m\setminus n)} = \left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(n),(m\setminus n)}. \tag{37}$$

We further divide the block $[(n),(m\setminus n)]$ into two blocks along their rows: $[(m\cap n),(m\setminus n)]$ and $[(n\setminus m),(m\setminus n)]$ that represent the influence from $\mathbf{z}^{(m\setminus n)}$ to $\hat{\mathbf{z}}^{(m\cap n)}$ and $\hat{\mathbf{z}}^{(n\setminus m)}$, due to the matrix structural disparity.

Expressing the block $[(m\cap n),(m\setminus n)]$ on the left-hand side of Eq. (37) gives:

$$\left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(m\cap n),(m\setminus n)} = \left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}\right]_{(m\cap n),:}\left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{:,(m\setminus n)} = \left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}\right]_{(m\cap n),(m\setminus n)}\left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(m\setminus n),(m\setminus n)}. \tag{38}$$

Analogously, this block on the right-hand side of Eq. (37) can be expressed as:

$$\left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(m\cap n),(m\setminus n)} = \left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(m\cap n),:}\left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{:,(m\setminus n)} = \left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(m\cap n),(m\cap n)}\left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(m\cap n),(m\setminus n)}. \tag{39}$$

Thus, we have the equality for the block $[(m\cap n),(m\setminus n)]$:

$$\begin{aligned}
&\left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}\right]_{(m\cap n),(m\setminus n)}\left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(m\setminus n),(m\setminus n)} = \left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(m\cap n),(m\cap n)}\left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(m\cap n),(m\setminus n)} \\
&\quad\Longrightarrow \\
&\left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}\right]_{(m\cap n),(m\setminus n)} = \left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(m\cap n),(m\cap n)}\left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(m\cap n),(m\setminus n)}\left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(m\setminus n),(m\setminus n)}.
\end{aligned} \tag{40}$$

This graphical relation is identical to that in Eq. (25). However, the relation for the block $[(n\setminus m),(m\setminus n)]$ between two modality-specific parts varies, due to the potential mixing of the shared part into these blocks, which may increase the inbound edges (not outbound edges), as we show below.

For the block $[(n\setminus m),(m\setminus n)]$ on the left-hand side of Eq. (37) gives:

$$\left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(n\setminus m),(m\setminus n)} = \left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}\right]_{(n\setminus m),:}\left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{:,(m\setminus n)} = \left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}\right]_{(n\setminus m),(m\setminus n)}\left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(m\setminus n),(m\setminus n)}. \tag{41}$$

Unlike previous cases, the right-hand side of Eq. (37) for the block involves more than atomic blocks (i.e., it involves the entire modality (n)):

$$\left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(n\setminus m),(m\setminus n)} = \left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(n\setminus m),:}\left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{:,(m\setminus n)} = \left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(n\setminus m),(n)}\left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(n),(m\setminus n)}. \tag{42}$$

Then, it follows from Eq. (41) and Eq. (42) that

$$\begin{aligned}
&\left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}\right]_{(n\setminus m),(m\setminus n)}\left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(m\setminus n),(m\setminus n)} = \left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(n\setminus m),(n)}\left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(n),(m\setminus n)} \\
&\quad\Longrightarrow \\
&\left[\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}\right]_{(n\setminus m),(m\setminus n)} = \left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(n\setminus m),(n)}\left[\boldsymbol{G}_{\frac{\partial \mathbf{z}}{\partial \mathbf{z}}}\right]_{(n),(m\setminus n)}\left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(m\setminus n),(m\setminus n)}.
\end{aligned} \tag{43}$$

We can observe that the existence of the shared variables $\mathbf{z}^{(m,n)}$ divides the latent space into finer blocks $\mathbf{z}^{(m\setminus n)}$, $\mathbf{z}^{(n\setminus m)}$, and $\mathbf{z}^{(m\cap n)}$. Eq. (40) and Eq. (43) reveal that the bijective indeterminacy relation hold over these finer blocks, exception for the non-square transition matrix $\left[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}\right]_{(n\setminus m),(n)}$ on the right-hand side of Eq. (43). This is because that the shared part $\mathbf{z}^{(m\cap n)}$ can potentially mix in $\hat{\mathbf{z}}^{(n\setminus m)}$, so $\hat{\mathbf{z}}^{(n\setminus m)}$ may receive edges inbound to $\mathbf{z}^{(m\cap n)}$.

Interplay among multiple modalities. In light of the graphical condition for the two-modality case (Eq. (40) and Eq. (43)), we can derive the conditions for the multi-modality case.

The participation of multiple modalities requires a new definition of the shared blocks in $\mathbf{z}$ since the sharing structure could be nested and various numbers of modalities could share one partition. We partition the entire latent space $\mathbf{z}$ into disjoint blocks $\{\mathbf{z}^{(b)}\}_{b\in B}$, whose components z have exactly the same modality membership $\mathcal{M}(z) := \{m \in [M] : z \in \mathbf{z}^{(m)}\}$. We define the $\mathbf{z}^{(H(b))}$ as the smallest (the least components) identified partition in $\mathbf{z}$ that contains $\mathbf{z}^{(b)}$. In the two-modal case, we have $B = \{(m\cap n), (m\setminus n), (n\setminus m)\}$ and $\mathbf{z}^{(H(m\cap n))} = \mathbf{1}_{|\mathbf{z}^{(m)}|\leq|\mathbf{z}^{(n)}|}\mathbf{z}^{(m)} + \mathbf{1}_{|\mathbf{z}^{(m)}|>|\mathbf{z}^{(n)}|}\mathbf{z}^{(n)}$.

We denote $\mathbf{z}^{(b_1)} \prec \mathbf{z}^{(b_2)}$ if block $\mathbf{z}^{(b_1)}$ is shared by a strict subset of modalities that share $\mathbf{z}^{(b_2)}$, i.e., $\mathcal{M}(\mathbf{z}^{(b_1)}) \subsetneq \mathcal{M}(\mathbf{z}^{(b_2)})$. Therefore, we have either $\mathbf{z}^{(H(b))} = \mathbf{z}^{(b)}$ (it is identifiable itself) or $\mathbf{z}^{(b)} \prec \mathbf{z}^{(H(b))} \setminus \mathbf{z}^{(b)}$ (it forms an identifiable block with a more deeply shared block). We denote former blocks as $b^+ \in B^+ \subset B$, i.e., $\mathbf{z}^{(H(b^+))} = \mathbf{z}^{(b^+)}$, and the latter blocks as $b^- \in B^- = B\setminus B^+$. In the two-modal case, we have $\mathbf{z}^{(m\setminus n)} \prec \mathbf{z}^{(m\cap n)}$ and $\mathbf{z}^{(n\setminus m)} \prec \mathbf{z}^{(m\cap n)}$, and $B^+ = \{(m\cap n)\}$ and $B^- = \{(m\setminus n), (n\setminus m)\}$.

We note that all shared blocks $\mathbf{z}^{(b^+)}$ are identified, and thus their bijective indeterminacies are w.r.t, themselves, i.e., $\mathbf{z}^{(b^+)} \mapsto \hat{\mathbf{z}}^{(b^+)}$, which implies the square shape of their indeterminacy matrices $[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}]_{(b^+),(b^+)}$ (as both downstream and upstream variables on both sides of the true graph matrix, e.g., Eq. (40)). In contrast, the unidentifiable blocks $\mathbf{z}^{(b^-)}$ can potentially receive the influence from all components in its modality $\mathbf{z}^{(H(b^-))}$. However, they do not influence the complement block $\mathbf{z}^{(H(b^-))}\setminus\mathbf{z}^{(b^{(-1)})}$ in its modality, since $\mathbf{z}^{(b^-)} \prec \mathbf{z}^{(H(b^-))}\setminus\mathbf{z}^{(b^-)}$. Consequently, their indeterminacy matrices are $[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}]_{(b^-),(H(b^-))}$ (as downstream variables at the right-side of the true graph matrix, e.g., Eq. (43)) and $[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}]_{(b^-),(b^-)}$ (as upstream variables at the right-side of the true graph matrix, e.g., Eq. (43)). Thus, we have the following categories:

Block 1 : Identifiable blocks $\mathbf{z}^{(b^+)}$: $[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}]_{(b^+),(b^+)}$ as both downstream and upstream variables on both sides of the true graph matrix $\boldsymbol{G}$.

Block 2 : Unidentifiable blocks $\hat{\mathbf{z}}^{(b^-)}$ as downstream variables at the right-side of the graph $\boldsymbol{G}$: their indeterminacy matrices are $[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}]_{(b^-),(H(b^-))}$.

Block 3 : Unidentifiable blocks $\hat{\mathbf{z}}^{(m\setminus\cdot)}$ as upstream variables at the right-side of the graph $\boldsymbol{G}$: their indeterminacy matrices are $[\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}]_{(b^-),(b^-)}$.

Subsequently, we classify the blocks in the estimation graph $\hat{\boldsymbol{G}}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\mathbf{z}}}}$ into the following categories according to the row (as downstream) and column (as upstream) indices: for an upstream block U and a downstream block D such that they belong to distinct blocks b_1 and b_2.

Region 1 : Blocks b_1 and b_2 do not have nested memberships, i.e., $\mathcal{M}(\mathbf{z}^{(b_1)}) \not\subset \mathcal{M}(\mathbf{z}^{(b_2)})$ and $\mathcal{M}(\mathbf{z}^{(b_2)}) \not\subset \mathcal{M}(\mathbf{z}^{(b_1)})$;

Region 2 : Block b_1 has fewer memberships than block b_2: $\mathbf{z}^{(b_1)} \prec \mathbf{z}^{(b_2)}$;

Region 3 : Block b_1 has more memberships than block b_2: $\mathbf{z}^{(b_2)} \prec \mathbf{z}^{(b_1)}$.

The sparsity for Region 1 and Region 2 is informative, whereas Region 3 is not. This is because in these the inherent indeterminacy from the subspace identifiability within each modality (Theorem 4.2) will engage the product $\boldsymbol{T}_{\frac{\partial \hat{\mathbf{z}}}{\partial \hat{\epsilon}}} \boldsymbol{T}_{\frac{\partial \hat{\epsilon}}{\partial \mathbf{z}}}$ in Eq. (36) in addition to the sparsity in the estimated graph $\boldsymbol{G}_{\frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}}}$. In comparison with Theorem 4.4, Region 2 becomes informative thanks to the identification of such shared blocks.

Overall conditions. Consolidating all the considerations above, we re-define objects in Condition 4.3 as follows.

1. The upstream variables $U^{(m)}$ in modality $\mathbf{z}^{(m)}$ is confined to blocks in B.
2. The downstream variables $D^{(m)}$ in modality $\mathbf{z}^{(m)}$ are over minimal identifiable blocks $\{H(b) : b \in B\}$.
3. The indeterminacy matrix $\boldsymbol{T} := \boldsymbol{T}_{\text{on}} + \boldsymbol{T}_{\text{off}}$ is not strictly block-diagonal. The matrix $\boldsymbol{T}_{\text{on}}$ contains all the on-diagonal square matrices $\boldsymbol{T}_{\text{on}} := \begin{bmatrix} \boldsymbol{T}_{b_1} & & & \\ & \boldsymbol{T}_{b_2} & & \\ & & \dots & \\ & & & \boldsymbol{T}_{b_{|B|}} \end{bmatrix}$, where each $\boldsymbol{T}_b$ is an invertible matrix. The matrix $\boldsymbol{T}_{\text{off}}$ contains all the off-diagonal nonzero elements such that $\text{Supp}(\boldsymbol{T}_{\text{on}}) \cap \text{Supp}(\boldsymbol{T}_{\text{off}}) = \emptyset$ and $\boldsymbol{T} := \begin{bmatrix} \tilde{\boldsymbol{T}}_{b_1} & & & \\ & \tilde{\boldsymbol{T}}_{b_2} & & \\ & & \dots & \\ & & & \tilde{\boldsymbol{T}}_{b_{|B|}} \end{bmatrix}$, where each $\tilde{\boldsymbol{T}}_b$ is of shape $d(\mathbf{z}^{(b)}) \times d(\mathbf{z}^{H(b)} \setminus \mathbf{z}^{(b)})$. Rows of distinct blocks b_1 and b_2 are disjoint, but their columns might due to the shared variables between their corresponding $H(b_1)$ and $H(b_2)$.
4. The sub-matrices on which we impose the sparsity controls are exactly the union of Region 2 and Region 1, i.e., the complement of Region 3. We denote such a region as the function of the block index $E(b) := B \setminus \{\tilde{b} \in B | \mathbf{z}^{(\tilde{b})} \prec \mathbf{z}^{(b)}\}$. Therefore, the condition becomes $[\boldsymbol{T}\boldsymbol{G}]_{E(b),C^{(b)}}$ and $\left([\boldsymbol{G}\boldsymbol{T}_{\text{on}}^{-1}]_{R^{(b)},E(b)}\right)^{\top}$. Note that $\boldsymbol{T}_{\text{on}}$ is invertible, although $\boldsymbol{T}$ may not.

With these modifications, the rest of the proof follows exactly from that of Theorem 4.4.

□

D Experimental Details

D.1 Numerical Dataset

Six numerical datasets are used in this paper, including three multi-group settings that satisfy our assumptions and three datasets that slightly violate the sparsity assumptions in the proposed theorems. The observations are generated using a multi-layer perceptron (MLP), following previous work Von Kügelgen et al. (2021); Yao et al. (2021); Zimmermann et al. (2021). Specifically, the mixing function g is modeled as a three-layer MLP with randomly initialized weights, Leaky ReLU activations, and hidden layers of sizes 8.

Multi-modality Cases In the multi-modality case, we generate $n = 10000$ samples according to Eq. (3), and the dimensionality of the observations in each modality ranged from $d_x = 15$ to 20. The causal noise terms ϵ are i.i.d. sampled from a Gaussian distribution, and the exogenous variables are also assumed to follow a Gaussian distribution. Sparse causal relations between inter-group variables are randomly generated, with a sparsity ratio controlled between 50% and 75%, ensuring that the latent variables in each group keep at least one causal connection with another group.

Ablation Cases We create two 15-dimensional modality observations under different sparsity ratios. Each observations are generated from three latent variables, two of which are causally latent and one exogenous. The sample size for each dataset was set to $n = 10000$, and the dimensionality of the observations in each modality is $d_x = 15$. Suppose the four latent variables are denoted as $z_1^{(1)}, z_2^{(1)}, z_1^{(2)}$, and $z_2^{(2)}$. For sparsity ratio = 0%, all inter-modality latent variables are fully connected. For sparsity ratio = 25%, three pairs of inter-modality connections among the latent variables are present. For sparsity ratio = 50%, two pairs of the inter-modality latent variables are connected. For sparsity ratio = 75%, only one pair of inter-modality latent variables remains connected.

D.2 Synthetic Dataset

Augmented MNIST The MNIST dataset (LeCun, 1998) is a widely-used benchmark for image classification, consisting of handwritten digits from 0 to 9. Building on this foundation, we designed our own variant of MNIST to investigate causal relationships across modalities. Our Variant MNIST consists of two modalities: colored MNIST and fashion MNIST. In the colored MNIST, each digit is assigned a specific color, where the digit class serves as the causal factor and the image color as the effect. In fashion MNIST, we use the class of clothing items as the cause and introduce the rotation angle of the image as the effect. Additionally, we establish a causal relation between the class labels of colored MNIST and fashion MNIST, allowing us to explore cross-modal dependencies in a controlled setting. These relationships are visually shown in the Figure 7 for clarity.

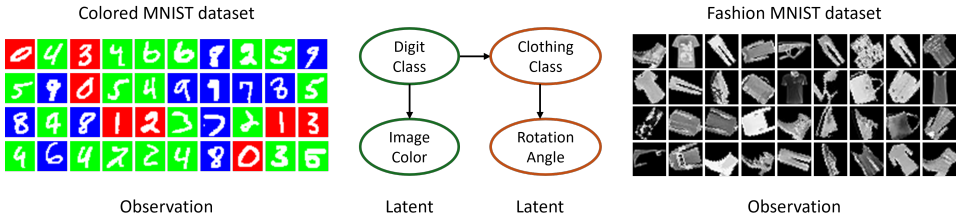


Figure 7: The ground-truth data generation process of MNIST images dataset.

D.3 Real-world Dataset

In this paper, we consider three types of datasets covering image, time series, and tabular data. Visualizations of the image and time-series datasets are shown in Figure 8.

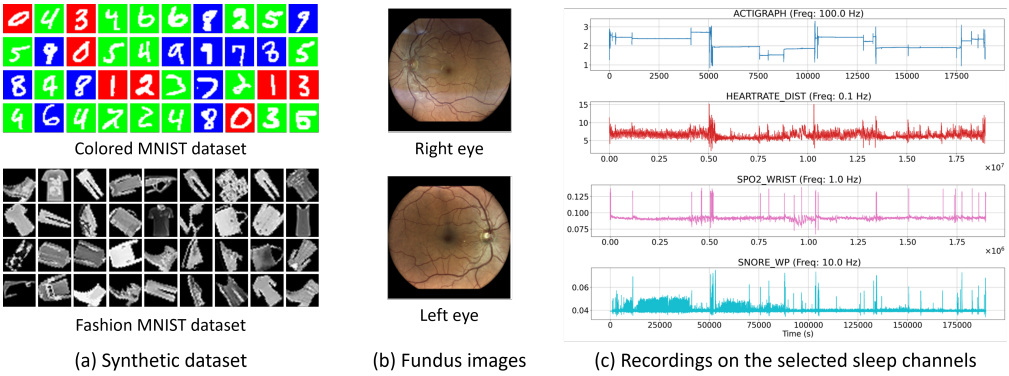


Figure 8: Visualization on the datasets: (a) Synthetic dataset: Augmented MNIST. (b) Real-world dataset: Fundus imaging shows the interior surface of the eyes. (c) Real-world dataset: Sleep monitoring shows the time-series recording of sleep-related metrics overnights.

Fundus imaging is the visualization of the interior surface of the fundus, which includes structures such as the optic disc, retina, and retinal microvasculature. High-resolution images of the back of the eye are essential for diagnosing and monitoring a variety of eye diseases and conditions.

For example, the retinal microvasculature, which consists of small blood vessels that supply blood to the retina, provides valuable information about eye health. Moreover, fundus imaging can enhance the understanding of the underlying mechanisms of various eye diseases. It serves as a non-invasive tool for assessing the overall health of the microvascular circulation health and provides a direct view of part of the central nervous system.

Sleep monitoring is a time-series dataset collected over three consecutive nights that records various metrics including sleep stage, body position, respiratory events, heart rate, oxygen saturation, and snoring. This dataset focuses on obstructive sleep apnea (OSA), a sleep disorder where a person's breathing is interrupted during sleep due to the relaxation of throat muscles, causing upper airway obstruction. These interruptions often lead to loud snoring, reduced blood oxygen levels, stress responses, awakenings, and fragmented sleep.

This dataset is collected from a Home Sleep Apnea Test (HSAT), a non-invasive diagnostic method for sleep apnea. Patients wear a portable device overnight to monitor their breathing patterns, heart rate, oxygen levels, snoring, and other sleep patterns. The dataset includes multiple channels, such as ACTIGRAPH for movement, HEARTRATE_DIST for heart rate, SPO2_WRIST for blood oxygen saturation, and SBORE_WP for snoring, capturing key aspects of physical activity and sleep patterns during the HSAT. The device calculates apnea-related indices, including the Apnea/Hypopnea Index (AHI), Respiratory Disturbance Index (RDI), and Oxygen Desaturation Index (ODI), as well as indices for diagnosing conditions such as atrial fibrillation.

D.4 Evaluation Metrics

MCC: Mean Correlation Coefficient MCC is a standard metric used to evaluate the recovery of latent factors in causal representation learning. It measures the alignment between ground-truth factors and estimated latent variables. Specifically, MCC first computes the absolute values of the correlation coefficients between each ground-truth factor and every estimated latent variable. To account for possible permutations of the latent variables, the metric solves a linear sum assignment problem on the computed correlation matrix in polynomial time, ensuring optimal matching between the factors and their corresponding latent representations.

SHD: Structural Hamming Distance SHD is a widely used metric for evaluating the accuracy of graph structure recovery in causal discovery. It quantifies the difference between the true causal graph and the estimated graph. Specifically, SHD counts the number of edge modifications—additions, deletions, or reversals—required to transform the estimated graph into the ground-truth graph. This metric provides a simple yet effective measure of structural similarity, with a lower SHD indicating closer alignment between the estimated and true causal structures.

R2: Coefficient of Determination R2 is a standard metric used to assess the goodness of fit in regression models. It measures the proportion of variance in the dependent variable that is explained by the independent variables in the model. Specifically, R2 compares the residual sum of squares of the model with the total sum of squares, providing a value between 0 and 1. A higher R2 indicates that the model explains a larger portion of the variance in the data, with 1 representing a perfect fit and 0 indicating that the model explains none of the variability.

D.5 Detailed Discussion on Human Phenotype

Without learning such latent variables, we cannot provide a causal explanation between different modalities. The estimated model shows all causal influences involved, suggests the existence of hidden causal variables, and illustrates their relationships with each other and with observable data. Asymptotically, the learned adjacency matrix A corresponds to a graph within the Markov equivalence class given by the PC algorithm.

To interpret the learned hidden variables, we primarily refer to existing medical literature, which supports their alignment with background knowledge, thereby adding validity to our results. For example, the latent variable FRight3 relates handgrip strength to fundus imaging, consistent with findings showing that handgrip strength correlates with intraocular pressure (IOP) (Pérez-Castilla et al., 2021). Additionally, the association between the cataract and changes in IOP (Slabaugh et al., 2013) aligns with the model's discovery. These connections underline the physiological relevance

of the learned hidden variable. Similarly, FRight1 and FLeft1, associated with fundus imaging and age estimation, are consistent with studies demonstrating age-related changes in fundus image color content (Ege et al., 2002). Another latent variable Sleep1 associated with oxygen saturation and sleep metrics aligns with findings that oxygen saturation is a strong predictor of obstructive sleep apnea (OSA) severity (Wali et al., 2020). This indicates that the model's latent variable effectively captures critical factors related to sleep disorders.

E EXTENDED EXPERIMENT

E.1 EXTENDED RESULTS ON COMPLEX SCENARIOS

To evaluate the scalability and generalizability of our method to complex causal structures, we conducted additional experiments on higher-dimensional simulated tasks with diverse configurations of latent variables and modalities. These setups introduce significantly more intricate causal relationships among variables compared to our original experiments.

The extended scenarios include: (1) Two-mods: 30-dimensional observations from two modalities with four latent variables and one exogenous variable per modality. (2) Five-mods: 30-dimensional observations from five modalities with two latent variables and one exogenous variable per modality. (3) Six-mods: 30-dimensional observations from six modalities with two latent variables and one exogenous variable per modality. The results, summarized in Table 4, show that our method maintains robust performance under these challenging conditions. Metrics including MCC and R2 demonstrate that our method continues to perform well under these more challenging conditions.

Metric	**Two mods**	**Five mods**	**Six mods**
R2	$0.89\pm1e-4$	$0.89\pm1e-4$	$0.97\pm8e-7$
MCC	$0.83\pm4e-6$	$0.84\pm3e-4$	$0.82\pm4e-4$

Table 4: Extended results on complex scenarios.

E.2 DISCUSSION ON THE NUMBER OF LATENT VARIABLES

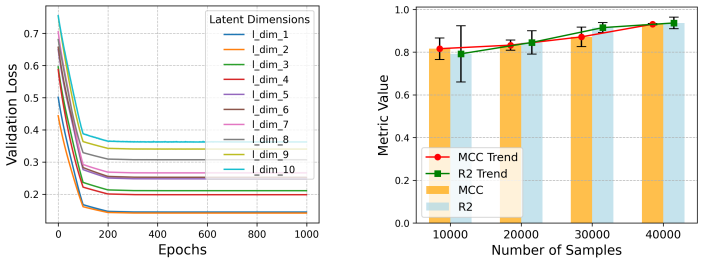
In real-world applications, the true number of latent variables is typically unknown, and arbitrarily predefining this number can introduce bias and degrade model performance. In this section, we discuss how our method can eliminate the redundant effect of the latent variables, and introduce a cross-validation-based method for determining the appropriate number of latent nodes.

At the same time, we can manually set the range of latent variable numbers and use cross-validation to select the one with the lowest validation loss. This approach is conceptually simple and widely applicable Khemakhem et al. (2020b), making it a reliable choice in scenarios where computational resources are not constrained. By directly linking the number of latent variables to model performance, it provides interpretable results that are easy to understand and justify. Here we conduct synthetic experiments to validate its effectiveness. We followed the data generation process in Section D.1, where the ground-truth number of latent variables is two for each modality. The results, as shown in Figure 9(a), demonstrate that our approach accurately recovers the correct number of latent variables.

E.3 DISCUSSION ON THE EFFECT OF SAMPLE NUMBERS

To investigate the impact of sample size on model performance, we conducted an additional experiment evaluating the MCC as the number of data samples increased. For this study, we followed the data generation process in Section D.1, where the ground-truth number of latent variables is two for two modalities. The experiment systematically increased the sample size from 10,000 to 40,000, while measuring MCC and R2 accordingly.

As shown in Figure 9(b), the results indicate that the MCC improves consistently with larger sample sizes, thereby confirming the hypothesis that a greater amount of data enhances the model's ability to recover the underlying causal structure. This experiment highlights the effectiveness of our approach in leveraging increased data availability to improve causal representation learning.



(a) (b)

Figure 9: (a) Comparison of loss across different latent dimensions. (b) The effect of sample size.

E.4 Discussion on the non-DAG models

The assumptions regarding Directed Acyclic Graphs (DAGs) for latent variable structures within or across modalities are not strictly necessary for the theoretical results in this paper. In this section, we conduct synthetic experiments to evaluate the performance of the model under non-DAG settings, both for cycles within modalities and across modalities. Specifically, we followed the data generation process in Section D.1, and considered additional: (1) cyclic influence within modality; and (2) cyclic influence across modalities, to generate the data. Empirical results in Table 5 demonstrate that the presence of cycles does not hinder the identification of latent variables.

Metric	Cyclic within mod	Cyclic across mods
R2	$0.95 \pm 1e-5$	$0.94 \pm 2e-4$
MCC	$0.89 \pm 2e-4$	$0.92 \pm 1e-5$

Table 5: Results under non-DAG settings.

E.5 Discussion on the Shared Latent Variables

In this section, we present how to extend the current network architecture to allow for the shared variables across modalities and provide corresponding experimental results.

The extended theorem in Section C.4 provides the theoretical guarantee on the generality of our framework to the shared latent variable scenario. The extended framework incorporates an additional mechanism to estimate the shared latent variable. Following the similar idea in Yao et al. (2023); Daunhawer et al. (2023); Von Kügelgen et al. (2021), an additional contrastive loss is introduced to enforce similarity in the shared latent representations across modalities, ensuring that shared variables capture the common causal structure.

To demonstrate the effectiveness of the extended framework, we extend the data generation process in Section D.1 and allow the existence of a shared variable across modalities. We conducted experiments on various synthetic settings and reported the results in Table 6. The high MCC shows the accurate recovery of both shared and modality-specific latent variables across different scenarios, confirming the theoretical guarantees of the extended framework.

Metric	Two mods	Three mods
R2	$0.86 \pm 5e-4$	$0.90 \pm 1e-4$
MCC	$0.83 \pm 7e-4$	$0.83 \pm 4e-6$

Table 6: Results for shared latent variables.

F Implementation Details

In this section, we provide details of the network architecture. The optimization scheme and hyperparameter settings are summarized. The computation efficiency is analyzed.

F.1 Network Architecture

We summarize our network architecture below and describe it in detail in Table 7.

- **(1,2) Encoder and Decoder**: The encoder transforms raw observations into latent representations, while the decoder reconstructs the inputs from the latent variables. The encoder-decoder design varies depending on the downstream task. For synthetic data, MLPs with LeakyReLU activation were used. For image data (e.g., MNIST), ResNet18 with LeakyReLU activation served as the encoder, while ConvTranspose2D was used as the decoder. LSTMs were used to extract latent features from time series data. By leveraging the universal approximation theorem, the model is theoretically capable of approximating the underlying mixing function.
- **(3) Learnable Adjacency Matrix**: The causal relationships are embedded in the learned adjacency matrix, where the binary elements indicate whether specific pairs of vertices contribute to the generation of components. It initializes a learnable matrix that captures these dependencies. During the forward pass, the matrix is processed to ensure a directional structure where only certain connections are allowed based on a threshold. This allows the model to learn sparse, meaningful relationships between the latent variables.
- **(4) Flow-based Transformation**: The flow-based transformation is implemented using a three-layer MLP to process the latent variable and a DDSF model for the transformation. The MLP first extracts features from the latent variable, which are then used to compute the parameters for the flow model. The DDSF applies an invertible transformation to the latent space, allowing the model to estimate the noise distribution.

F.2 Training Details

Optimization Scheme. The models were implemented in PyTorch and trained on a GPU. The estimation framework was trained using the Adam optimizer with an initial learning rate of 0.002, and the StepLR scheduler was used to reduce the learning rate periodically. The training process ran for a maximum of 10000 epochs, with early stopping applied if the validation loss does not improve for 20 consecutive epochs. Random seeds were used to ensure reproducibility, and results were averaged across experiments.

The training loss combines multiple components.

- Reconstruction Loss: Mean squared error between reconstructed inputs and original data.
- KL Divergence Loss: Encourages estimated variables to follow a standard normal prior.
- Sparsity Loss: An L1-norm penalty is applied to the adjacency matrix to enforce sparsity.

Hyperparameter Details. The hyperparameters $\alpha = [\alpha_{\text{Ind}}, \alpha_{\text{Sp}}, \alpha_{\text{Recon}}]$ represent the weights assigned to each term in the composite objective function. For the experiments, the following settings were applied: $\alpha = [1e{-}2, 1, 10]$ for the synthetic dataset, $\alpha = [1e{-}5, 2, 10]$ for the MNIST dataset, and $\alpha = [1e-1, 1e-2, 1]$ for the phenotype dataset.

G Algorithm Pseudocode

The pseudocode for the proposed algorithm is presented in Algorithm 1.

Configuration	Description	Output
1.1 MLP-Encoder	Encoder for synthetic data	
Input	Multi-modality observations	BS × d_x
Dense	32 neurons, LeakyReLU	BS × 32
Dense	32 neurons, LeakyReLU	BS × 32
Dense	Latent embeddings	BS × l_dim
2.1 MLP-Decoder	Decoder for synthetic data	
Input	Latent embeddings	BS × l_dim
Dense	32 neurons, LeakyReLU	BS × 32
Dense	32 neurons, LeakyReLU	BS × 32
Dense	Reconstructed observations	BS × d_x
1.2 Image-Encoder	Encoder for image data	
Input	Image input	BS × 3 × H × W
ResNet18	ResNet backbone, LeakyReLU	BS × h_dim
Dense	Latent embeddings	BS × l_dim
2.2 Image-Decoder	Decoder for image data	
Input	Latent embeddings	BS × l_dim
Dense	h_dim neurons	BS × h_dim × 7 × 7
ConvTranspose2D	BatchNorm2D, LeakyReLU	BS × h_dim × 14 × 14
ConvTranspose2D	Sigmoid, Reconstructed observations	BS × 3 × H × W
1.3 Time-series Encoder	Encoder for time-series data	
Input	Multi-channel time-series data	BS × seq_len × n_channel
LSTM	Sequences into hidden representations	BS × h_dim
Output	Latent representation	BS × l_dim
2.3 Time-series Decoder	Decoder for time-series data	
Input	Latent representation	BS × l_dim
LSTM	Sequence into output features	BS × seq_len × h_dim
Output	Reconstructed time-series data	BS × seq_len × n_channel
3. Adjacency Matrix	Sparsity regularization	
Input	Latent variables from encoders	BS × z_all
Masking	Lower triangular mask	z_all × z_all
Thresholding	Retain entries exceeding threshold	z_all × z_all
Output	Learned causal adjacency matrix	z_all × z_all
4. Flow Transformation	Conditional independence constraints	
Input	Latent variables across modalities	BS × z_all
Condition Input	Apply adjacency matrix to latent	BS × z_all × z_all
MLP-transformation	A lower-dimensional feature space	BS × z_all × 32
Flow Parameter Net	Flow parameters for the transformation	BS × z_all × n_para
Flow Transformation	DDSF to the reshaped latent variables	BS × z_all
Output	Estimated noise variables	BS × z_all

Table 7: Architecture details. BS: batch size, d_x: input dimension, l_dim: latent dimension in each modality, z_all: latent dimensions across all modalities, h_dim: hidden dimension, H/W: height/width of the input image, seq_len: sequence length, n_channel: number of channels, LeakyReLU: Leaky Rectified Linear Unit.

Algorithm 1 Pseudocode for the proposed algorithm.

Input: Grouped observations $\{\mathbf{x}^{(m)}\}_{m=1}^{M}$
Output: Estimated latent variables $\{\hat{\mathbf{z}}^{(m)}\}_{m=1}^{M}$; Inferred causal graph $\hat{\mathcal{G}}$

 # Random Initialization
Initialize adjacency matrix $\hat{\mathbf{A}}$
Initialize encoders $\{\text{En}^{(m)}\}_{m=1}^{M}$ and decoders $\{\text{De}^{(m)}\}_{m=1}^{M}$ for each group

 # Conditional Independence Constraint
Input: Grouped observations $\{\mathbf{x}^{(m)}\}_{m=1}^{M}$
Output: Estimated latent variables $\hat{\mathbf{z}}^{(m)}$ for each group m
for each group $m = 1$ **to** M **do**
    Encode the current group latent and exogenous variables: $\hat{\mathbf{z}}^{(m)}, \hat{\eta}^{(m)} = \text{En}^{(m)}(\mathbf{x}^{(m)})$
end for
Concatenate latent representations: $\{\hat{\mathbf{z}}^{(m)}\}_{m=1}^{M} = \hat{\mathbf{z}}^{(1)} \oplus \hat{\mathbf{z}}^{(2)} \oplus \ldots \oplus \hat{\mathbf{z}}^{(M)}$
return Estimated latent variables and exogenous variables $\{\hat{\mathbf{z}}^{(m)}, \hat{\eta}^{(m)}\}_{m=1}^{M}$

 # Flow-based Noise Estimation
Input: Estimated latent variables for each group $\{\hat{\mathbf{z}}^{(m)}\}_{m=1}^{M}$
Output: Inferred causal graph $\mathcal{G}$
Initialize an empty causal graph $\mathcal{G} = \emptyset$
Choose the parents of each latent variable $\hat{z}_i$ based on the adjacency matrix
for each flow block **do**
    Pass $\text{Pa}(\hat{z}_i)$ through flow to obtain estimated residuals $\hat{\epsilon}_i$ and log determinant Jacobian
    Compute sparsity loss based on L_1 norm of the adjacency matrix
    Update the estimated causal graph based on the variable influence with threshold
end for
Optimize the KL divergence between $[\{\hat{\eta}^{(m)}\}_{m=1}^{M}, \hat{\epsilon}_{i=1}^{d_s}]$ and Gaussian prior
return Inferred causal graph $\hat{\mathcal{G}}$

 # Decoder
Input: Estimated latent and exogenous variables in each group $\{\hat{\mathbf{z}}^{(m)}, \hat{\eta}^{(m)}\}_{m=1}^{M}$
Output: Reconstructed grouped features $\{\hat{\mathbf{x}}^{(m)}\}_{m=1}^{M}$
for each group $m = 1$ **to** M **do**
    Decode $(\hat{\mathbf{z}}^{(m)}, \hat{\eta}^{(m)})$ to reconstruct features $\hat{\mathbf{x}}^{(m)}$: $\hat{\mathbf{x}}^{(m)} = \text{De}^{(m)}(\hat{\mathbf{z}}^{(m)}, \hat{\eta}^{(m)})$
    Compute reconstruction loss using MSE: $\mathcal{L}_{\text{Recon}}^{(m)} = \text{MSE}(\hat{\mathbf{x}}^{(m)}, \mathbf{x}^{(m)})$
end for