Decoding Multimodal Causality: End-to-End Multimodal Mediation Pathways Inference

Yulong Li[1,2], Xiwei Liu[1], Niranjana Arun Menon[3], Yuxuan Zhang[1,2], Jianxu Chen[1], Rong Xia[1,2], Haolin Yang[1], Peixin Guo[2], Yutong Xie[1†], and Imran Razzak[1†]

[1] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[2] Xi'an Jiaotong-Liverpool University, Suzhou, China
[3] University of New South Wales, Sydney, Australia
{yulong.li,Yutong.Xie,Imran.Razzak}@mbzuai.ac.ae
[†]Corresponding authors.

Abstract. The nonparametric identification of causal mediation effects from heterogeneous multimodal observational data remains an open theoretical problem. Existing single-modal methods cannot handle cross-modal causal relationships, while multimodal learning methods lack causal inference capabilities, preventing reliable mediation pathway identification and effect quantification under structural uncertainty. We propose the Multimodal Structure-Informed Guided Mediation Analysis (MM-SIGMA) framework, which achieves automated cross-modal mediation pathway identification and end-to-end uncertainty propagation through probabilistic causal structure discovery. We establish nonparametric identification conditions for the Cross-Modal Natural Direct Effect and Cross-Modal Natural Indirect Effect from heterogeneous multimodal observational data, and build MM-SIGMA upon this theoretical foundation. Specifically, it employs multimodal variational autoencoders for causality-preserving latent representation, differentiable Flow-Structural Equation Models for asymptotically consistent latent structure learning, Cross-Modal Path Stability Scoring for high-confidence pathway identification, and Efficient Influence Functions with Bayesian Model Averaging for end-to-end uncertainty propagation. Experiments on synthetic data demonstrate state-of-the-art performance under structural uncertainty, nonlinearity, and cross-modal heterogeneity. On the HPP dataset, MM-SIGMA identifies cross-modal mediation pathways connecting sleep, fundus imaging, and cardiovascular health, revealing undiscovered cross-system mechanisms.

Keywords: Multimodal causal inference · Mediation analysis

1 Introduction

Causal mediation analysis has emerged as a critical framework bridging causal inference and machine learning [12, 24, 39]. Recent studies emphasize the evolution from estimating total effects to elucidating underlying mechanisms [26, 41],

a transition essential for developing clinical decision systems capable of intervention reasoning and interpretable explanations [25, 28]. Modern disease processes span multiple biological layers: molecular and cellular abnormalities can trigger changes in organ morphology and function, ultimately manifesting as clinical outcome differences [16,40]. Single modalities can only capture fragmentary information from one layer [9], while disease mechanisms exhibit cross-modal cascade effects–for example, metabolic disorders first manifest in molecular biomarkers [1], subsequently reflect in retinal vascular morphological changes in fundus images [43], and further impact cardiovascular functional indicators [11]. Existing single-modal causal analysis methods cannot adequately reveal such cross-level, cross-modal mediation chains [37, 44]. Therefore, achieving biologically plausible cross-modal mechanism modeling with robust pathway effect quantification constitutes the current core challenge in multimodal causal inference.

Disease modalities differ substantially in dimensional scale, distribution, and temporal granularity [17], with cross-modal associations further complicated by nonlinearity and potential time delays [8]. Empirical evidence confirms such cross-modal causal chains: sleep apnea has been shown to affect cardiovascular outcomes via inflammatory mediators [36], while retinal vascular morphology independently predicts cardiovascular risk [18]. Yet the complete causal mechanisms underlying complex diseases cannot be captured by any single modality [31]—a heterogeneity that fundamentally violates the homogeneity assumptions of existing mediation analysis methods [20].

Extending causal mediation analysis to multimodal environments faces four interconnected technical challenges: *i)* Multimodal Unified Representation Learning (MURL): existing methods optimize reconstruction rather than causal objectives [2, 29], causing latent representations to destroy causal dependencies in original data [4]. *ii)* Latent Space Causal Structure Discovery (LSCSD): traditional causal discovery algorithms cannot guarantee that latent causal relationships correspond to those among original variables [15, 46, 47]. *iii)* Cross-Modal Mediation Pathway Identification (CMMPI): existing methods rely on predefined pathways and cannot operate cross-modally [7, 42], leaving cross-modal mediation identifiability theoretically unestablished. *iv)* Uncertainty Propagation and Effect Estimation (UPEE): existing estimators assume known structures [30, 34], preventing valid confidence interval construction [6] under multimodal structural uncertainty. We propose unifying these four challenges within MM-SIGMA to quantify cross-modal mediation pathways under structural uncertainty.

To bridge this theoretical gap, we first establish nonparametric identification conditions for CM-NDE and CM-NIE from heterogeneous multimodal observational data under standard causal assumptions, providing the theoretical foundation upon which the entire framework rests. Building on this, we propose the Multimodal Structure-Informed Guided Mediation Analysis (MM-SIGMA) framework, which achieves automatic identification of cross-modal mediation pathways and end-to-end uncertainty propagation through probabilistic causal structure discovery and uncertainty quantification. Specifically, MM-SIGMA constructs a

multimodal variational autoencoder architecture that maps heterogeneous data to a shared latent space through structured and image data encoders along with cross-modal alignment loss. Based on unified representations, MM-SIGMA employs differentiable Flow-Structural Equation Models for structure learning in latent space, learning edge parameters encoding structural uncertainty and sampling to generate directed acyclic graph ensembles. On this ensemble, we apply Path Stability Score to identify high-confidence latent mediation pathways and map them to interpretable cross-modal mediation pathways based on correspondence between latent and original variables, thereby avoiding manual pathway pre-specification limitations. For identified mediation pathways, MM-SIGMA employs efficient influence functions for pathway-specific effect estimation and uses Bayesian model averaging to fuse within-structure estimation uncertainty and between-structure effect variation, propagating uncertainty to final effect estimates. We validate MM-SIGMA's effectiveness on high-dimensional, heterogeneous HPP data, identifying and quantifying cross-modal mediation pathways connecting sleep, retinal image features, and cardiovascular health.

The contributions are summarized as follows:

1. We establish nonparametric identification conditions for CM-NDE and CM-NIE from heterogeneous multimodal observational data, filling a fundamental theoretical gap in cross-modal causal mediation analysis.
2. We propose MM-SIGMA, an end-to-end framework built upon the above identification theory, unifying multimodal representation learning with probabilistic causal structure discovery via differentiable Flow-SEM, and achieving uncertainty propagation from representation learning to effect estimation via extended Efficient Influence Functions with Bayesian Model Averaging.
3. We introduce CM-PSS, identifying cross-modal mediation pathways by quantifying marginal probabilities of latent paths in posterior DAG samples, with interpretable mappings to original variables.
4. We validate MM-SIGMA on high-dimensional heterogeneous HPP data, identifying and quantifying cross-modal mediation pathways connecting sleep, retinal images, and cardiovascular health outcomes.

2 Real-world Multimodal Dataset

We employ the Human Phenotype Project (HPP)[4] dataset for real-world validation. As illustrated in Figure 1(a), HPP is a phenotyping cohort study with home sleep apnea test data from 6,366 adults(3,043 male and 3,323 female participants) across 16,812 nights of continuous monitoring. The overall project includes approximately 28,000 participants, with over 13,000 completing their initial visit. The coordinated collection process, as shown in Figure 1(b), provides temporally aligned longitudinal measurements across multiple physiological systems. The project conducts measurements across 16 body systems within a ±6-month time window, ensuring temporally aligned longitudinal data quality. The study collected 448 sleep characteristics, aggregating clinical and physiological attributes into domain knowledge features through cross-system grouping.

[4] `https://knowledgebase.pheno.ai/`

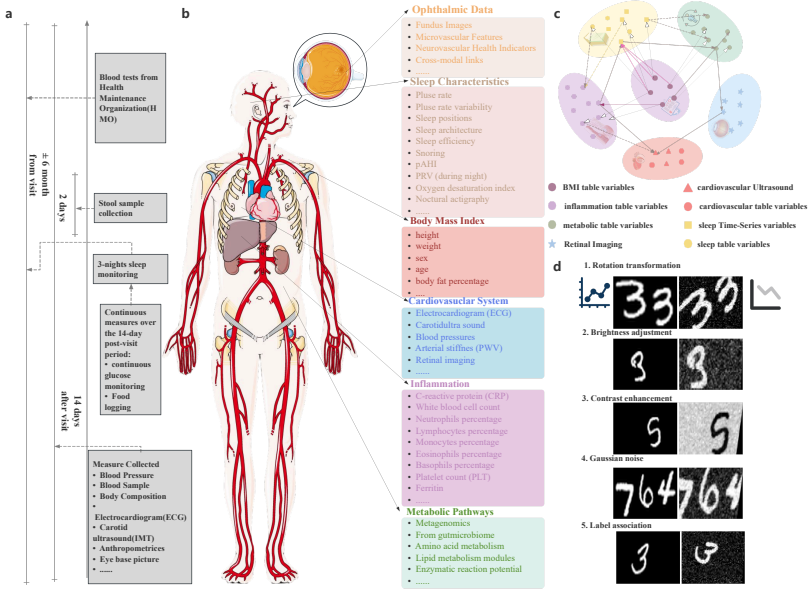

Fig. 1: MM-SIGMA: (a) HPP modality categories, (b) 14-day post-visit acquisition timeline, (c) schematic of cross-modal mediation discovery under structural uncertainty, and (d) synthetic image generation via five deterministic causal embeddings.

We employ five representative data categories to construct a cross-modal analysis framework, as depicted in Figure 1(a). Sleep features include sleep duration, sleep efficiency, apnea-hypopnea index (AHI), oxygen desaturation index, and arousal index as sleep quality parameters. Body composition indicators encompass body mass index, muscle mass, visceral fat ratio, waist-hip ratio, and other anthropometric parameters. The cardiovascular system includes systolic blood pressure, diastolic blood pressure, heart rate variability, carotid intima-media thickness, vascular stiffness, and other cardiovascular functional parameters. Inflammatory and metabolic pathways encompass inflammatory markers such as interleukin-6 and C-reactive protein, as well as metabolic indicators including fasting glucose, insulin, total cholesterol, and triglycerides. Fundus imaging features are acquired from 8,467 individuals via non-mydriatic fundus photography, extracting 12 quantitative vessel indicators (density, width, fractal dimension, tortuosity) and structural parameters (optic disc area, cup-to-disc ratio, macular thickness) through standardized deep learning preprocessing.

3 Synthetic Multimodal Dataset Construction

To evaluate MM-SIGMA under controlled conditions, we construct synthetic multimodal datasets with known cross-modal causal relationships. Let $X^{(s)} = \{X_1^{(s)}, ..., X_{d_s}^{(s)}\} \in \mathbb{R}^{d_s}$ denote structured variables and $I_0 \in \mathbb{R}^{H\times W\times C}$ the base image. We construct $\mathcal{T} : X^{(s)} \times I_0 \to I$ generating causally informative images via deterministic transformations, where $P(I|\text{do}(X_k^{(s)} = x)) = \delta(I - \mathcal{T}_k(I_0, x))$.

Five transformation types embed structured variables into image space: **rotation** $T_{\text{rot}} = \text{Rotate}(I_0, \theta_k \cdot X_k^{(s)})$; **brightness** $T_{\text{bright}} = \text{Clamp}(I_0 + \beta_k \cdot \frac{X_k^{(s)} - \mu_k}{\sigma_k}, 0, 1)$; **contrast** $T_{\text{contrast}} = \text{Clamp}(\mu + (1 + \alpha_k X_k^{(s)})(I_0 - \mu), 0, 1)$; **Gaussian noise** $T_{\text{noise}} = I_0 + \mathcal{N}(0, \sigma_{0,k}^2(1 + \lambda_k |X_k^{(s)}|))$; and **label association** $Y = f_{\text{label}}(X_k^{(s)}, \epsilon)$, $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. The final image combines these as $I = T_{K_n} \circ \cdots \circ T_{K_1}(I_0, X_{K_1}^{(s)}, ..., X_{K_n}^{(s)})$. We design three pathway configurations—parallel ($I = \bigotimes_{k \in K} T_k(I_0, X_k^{(s)})$), serial ($X_j^{(s)} = g_j(X_i^{(s)}, \epsilon_j)$, $I = T_j(I_0, X_j^{(s)})$), and hybrid—simulating mediation structures of varying complexity. These transformations represent generic image variation mechanisms (geometric, photometric, stochastic, and semantic) independent of any architectural assumption in MM-SIGMA. The five transformation types were selected prior to any model design decisions, drawn exclusively from domain-agnostic image variation literature [33] covering geometric, photometric, stochastic, and semantic transformations. All baselines are evaluated on identical synthetic datasets generated without reference to model architecture, precluding circularity between data generation and model design.

4 Definitions, Notation, and Identification

We consider causal mediation analysis in multimodal observational data. Let A denote the treatment variable, Y the outcome variable, $X^{(s)} = \{X_1^{(s)}, ..., X_{d_s}^{(s)}\}$ a set of structured variables, $X^{(i)} = \{X_1^{(i)}, ..., X_{d_i}^{(i)}\}$ a set of image variables, and C a set of covariates. Variables may be continuous, discrete, or mixed-type. Let $V = \{A, Y\} \cup X^{(s)} \cup X^{(i)} \cup C$ be the full set of observed variables and assume the underlying causal structure is a DAG $G^* = (V, E)$, where $\text{Pa}_{G^*}(V_i)$ denotes the parent set of node V_i. We assume that the true graph G^* satisfies the Causal Markov and Faithfulness assumptions. Due to the high-dimensionality and heterogeneity of original multimodal data, we introduce a latent representation space $Z = \{Z_1, ..., Z_p\}$ where $p \ll d_s + d_i$. Through an encoding function $f_{\text{enc}} : (X^{(s)}, X^{(i)}) \to Z$, heterogeneous data are mapped to a unified latent space while preserving causal semantics. We adopt the Potential Outcomes Framework to define cross-modal causal mediation effects. Let $M^{(s)}(a)$ and $M^{(i)}(a)$ be the potential structured and image mediator values under treatment $A = a$, and $Y(a, m^{(s)}, m^{(i)})$ the potential outcome if $A = a$, $M^{(s)} = m^{(s)}$, and $M^{(i)} = m^{(i)}$. We focus on two cross-modal mediation effects comparing treatment a to reference a^*: (1) The Cross-Modal Natural Direct Effect (CM-NDE) is defined as $\text{CM-NDE}(a, a^*) = \mathbb{E}[Y(a, M^{(s)}(a^*), M^{(i)}(a^*)) - Y(a^*, M^{(s)}(a^*), M^{(i)}(a^*))]$. (2) The Cross-Modal Natural Indirect Effect (CM-NIE) is: $\text{CM-NIE}(a, a^*) = \mathbb{E}[Y(a, M^{(s)}(a), M^{(i)}(a)) - Y(a, M^{(s)}(a^*), M^{(i)}(a^*))]$. Identifying cross-modal mediation effects from observational data requires several key assumptions. We use $\perp\!\!\!\perp$ to denote conditional independence.

Consistency: if $A = a'$, then $M^{(s)} = M^{(s)}(a')$ and $M^{(i)} = M^{(i)}(a')$; if $A = a'$, $M^{(s)} = m^{(s)}$, and $M^{(i)} = m^{(i)}$, then $Y = Y(a', m^{(s)}, m^{(i)})$, for all relevant treatment levels $a' \in \{a, a^*\}$.

Ignorability. For all relevant values $(a, a^*, m^{(s)}, m^{(i)})$, the following hold surely:

(i) $A \perp\!\!\!\perp \{Y(a^*, m^{(s)}, m^{(i)}),\ M^{(s)}(a),\ M^{(i)}(a)\} \mid C$,
(ii) $\{M^{(s)}, M^{(i)}\} \perp\!\!\!\perp Y(a, m^{(s)\prime}, m^{(i)\prime}) \mid A = a,\ C$,
(iii) $Y(a^*, m^{(s)}, m^{(i)}) \perp\!\!\!\perp \{M^{(s)}(a),\ M^{(i)}(a)\} \mid C$.

Positivity: For all a and c in the support of A and C, the treatment propensity score $\pi(a|C) \equiv \Pr(A = a \mid C = c) \in (c_1, c_2)$ for some constants $0 < c_1 \leq c_2 < 1$. For structured mediators $M^{(s)}$, the conditional density given A and $C = c$ is bounded between $[\rho_1, \rho_2]$ for some constants $0 < \rho_1 \leq \rho_2 < \infty$. For image mediators $M^{(i)}$, Positivity is imposed on the latent representation $Z^{(i)}$ rather than the raw image space, which is feasible given the compact support of the variational posterior $q_\phi(Z^{(i)} \mid X^{(i)})$.

Theorem 1 (Nonparametric Identification of CM-NDE and CM-NIE). *Under Consistency, Ignorability, Positivity, and:*

$$Y(a, m^{(s)}, m^{(i)}) \perp\!\!\!\perp m^{(i)} \mid Z^{(i)}, A = a, M^{(s)}, C \qquad \text{(Latent Sufficiency)}$$

CM-NDE and CM-NIE are nonparametrically identified:

$$\text{CM-NDE}(a, a^*) = \int \Delta\mu_Y \, dF_{M^{(s)}|a^*,c} \, dF_{Z^{(i)}|a^*,c} \, dF_C \tag{1}$$

$$\text{CM-NIE}(a, a^*) = \int \mu_Y(a, m^{(s)}, z^{(i)}, c) \, \Delta F_{M^{(s)},Z^{(i)}|c} \, dF_C \tag{2}$$

where $\Delta\mu_Y = \mu_Y(a, m^{(s)}, z^{(i)}, c) - \mu_Y(a^*, m^{(s)}, z^{(i)}, c)$, $\Delta F_{M^{(s)},Z^{(i)}|c} = dF_{M^{(s)},Z^{(i)}|a,c} - dF_{M^{(s)},Z^{(i)}|a^*,c}$.

Remark 1 (Distinction from Classical Mediation Formula). Eqs. (1)–(2) extend Pearl [26] by replacing $M^{(i)}$ with $Z^{(i)}$ under Latent Sufficiency, and accommodating cross-modal interactions via the joint mediator distribution $F_{M^{(s)},Z^{(i)}}$.

5 Methodology

We propose Multimodal Structure-Informed Guided Mediation Analysis (MM-SIGMA), a unified framework for cross-modal causal mediation analysis under structural uncertainty. Given data over treatment, outcome, mediators, and confounders, MM-SIGMA employs a multimodal variational autoencoder to map heterogeneous data to a causality-preserving latent space (Section 5.1), performs probabilistic structure discovery via Flow-SEM to generate a posterior DAG ensemble with interpretable latent-to-original variable mappings (Section 5.2), introduces CM-PSS to identify high-confidence cross-modal mediation pathways (Section 5.3), and estimates pathway-specific effects via efficient influence functions with Bayesian model averaging for uncertainty propagation (Section 5.4).

5.1 Multimodal Variational Representation Learning

To model structural uncertainty in multimodal causal mediation analysis, MM-SIGMA employs a multimodal variational autoencoder to learn unified latent representations of heterogeneous data. The goal is to infer a posterior distribution $P(Z|X^{(s)}, X^{(i)})$ such that conditional dependencies among latent variables accurately reflect causal mechanisms of original cross-modal variables.

Multimodal Encoder Architecture MM-SIGMA builds a variable-level encoder architecture with independent encoders per structured variable to maintain interpretability. For each $X_k^{(s)} \in X^{(s)} = \{X_1^{(s)}, ..., X_{d_s}^{(s)}\}$, an independent encoder maps to its latent counterpart:

$$Z_k^{(s)} = f_{\text{enc},k}^{(s)}(X_k^{(s)}; \theta_{s,k})$$

For image data $X^{(i)}$, a CNN encoder maps to latent representation:

$$Z^{(i)} = f_{\text{enc}}^{(i)}(X^{(i)}; \theta_i)$$

The final latent space $Z = \{Z_1^{(s)}, ..., Z_{d_s}^{(s)}, Z^{(i)}\}$ directly maps structured and image variables with interpretable latent-to-original correspondence.

Causality-Preserving Latent Space Construction MM-SIGMA learns latent representations via variational inference. The variational posterior for latent variables is $q_\phi(Z|X^{(s)}, X^{(i)}) = \mathcal{N}(Z; \mu_\phi, \sigma_\phi^2 \mathbf{I})$, where μ_ϕ, σ_ϕ are neural network-parameterized mean and variance functions. The variational objective function combines reconstruction loss, KL divergence, and causal constraints:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} - \beta \mathcal{L}_{\text{KL}} + \lambda_{\text{causal}} \mathcal{L}_{\text{causal}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}}$$

where $\mathcal{L}_{\text{recon}} = \mathbb{E}_{q_\phi}[\log p_\theta(X|Z)]$, $\mathcal{L}_{\text{KL}} = \text{KL}[q_\phi(Z|X) \| p(Z)]$, the causal constraint loss $\mathcal{L}_{\text{causal}}$ penalizes violations of domain-grounded conditional independencies in the latent space, and the cross-modal alignment loss $\mathcal{L}_{\text{align}}$ achieves alignment by maximizing mutual information between structured and image representations. Model parameters are jointly trained through stochastic gradient optimization. Upon completion of training, latent representations Z preserve causal dependencies of original multimodal data. The causal constraint loss is defined over a set $\mathcal{K} \subseteq \mathcal{V}^3$ of conditional independence triplets (i, j, k), each encoding $Z_i \perp\!\!\!\perp Z_j \mid Z_k$, constructed from two sources *prior to* model training without requiring knowledge of G^*: **(i)** domain knowledge constraints from established biomedical findings, and **(ii)** kernel-based conditional independence tests on a held-out split, retaining triplets with $p > 0.1$ under Bonferroni correction:

$$\mathcal{L}_{\text{causal}} = \sum_{(i,j,k) \in \mathcal{K}} \hat{I}(Z_i; Z_j \mid Z_k)$$

where $\hat{I}(\cdot\,; \cdot \mid \cdot)$ is the empirical conditional mutual information estimator. Residual structural uncertainty beyond $\mathcal{K}$ is propagated through the Flow-SEM posterior $P(G_Z \mid \mathcal{D}_Z)$ (Section 5.2). We establish the following guarantee:

Proposition 1 (Causal Semantic Preservation). *Let $\mathcal{K}$ be the set of conditional independence triplets constructed as defined above. If $\mathcal{L}_{causal} = 0$, then the learned latent representations Z satisfy $Z_i \perp\!\!\!\perp Z_j \mid Z_k$ for all $(i, j, k) \in \mathcal{K}$. If additionally $\mathcal{K}$ forms a d-separating set of G^*, the complete Markov structure of G^* is recovered in Z.*

5.2 Latent Space Probabilistic Structure Discovery

MM-SIGMA employs Flow-SEM to infer $P(G_Z|D_Z)$ encoding dependencies among latent variables, where $G_Z = (Z, E_Z)$ is the latent space DAG.

Flow-SEM in Latent Space Each latent variable Z_i is modeled as a stochastic function of its parents in a candidate DAG G_Z:

$$Z_i = f_i(\mathrm{Pa}_{G_Z}(Z_i); \phi_i) + \sigma_i(\mathrm{Pa}_{G_Z}(Z_i); \phi_i) \cdot U_i, \quad U_i \sim \mathcal{N}(0, 1)$$

To enable differentiable structure learning, a continuous weight matrix $W_\theta \in \mathbb{R}^{p\times p}$ is introduced, where $[W_\theta]_{ji}$ reflects the directed edge strength from Z_j to Z_i. The log-likelihood of an observed latent sample $Z^{(j)}$ is:

$$\log P(Z^{(j)}|W_\theta, \Phi) = \sum_{i=1}^{p} \log p_U(u_i^{(j)}) - \sum_{i=1}^{p} \log \sigma_i^{(j)}$$

where $u_i^{(j)} = (Z_i^{(j)} - \mu_i^{(j)})/\sigma_i^{(j)}$, with $\mu_i^{(j)} := f_i(\mathrm{Pa}_{G_Z}(Z_i^{(j)}); \phi_i)$ and $\sigma_i^{(j)} := \sigma_i(\mathrm{Pa}_{G_Z}(Z_i^{(j)}); \phi_i)$.

Latent Structural Posterior and DAG Sampling Model parameters (W_θ, Φ) are optimized by minimizing a regularized loss:

$$\mathcal{L}(W_\theta, \Phi; D_Z) = -\frac{1}{n} \sum_{j=1}^{n} \ell^{(j)} + \lambda_{\mathrm{dag}} h(W_\theta) + \alpha ||W_\theta||_1$$

where $\ell^{(j)} = \log P(Z^{(j)}|W_\theta, \Phi)$, $h(W_\theta) = \mathrm{tr}(\exp(W_\theta \odot W_\theta)) - p$ enforces acyclicity, and $||W_\theta||_1$ encourages sparsity. To approximate the structural uncertainty over DAGs, each element of W_θ parameterizes an edge inclusion score $p_{ij} = \sigma([W_\theta]_{ji})$, serving as a variational structural uncertainty score quantifying relative edge confidence under the Flow-SEM likelihood. Candidate DAGs are sampled via Bernoulli(p_{ij}) with greedy cycle removal; the resulting sampling bias is bounded by $\sum_k p_{ik}p_{kj}$ and vanishes as $\lambda_{\mathrm{dag}} \to \infty$. Residual uncertainty is absorbed into $\hat{V}_{\mathrm{between}}$ of the BMA estimator (Section 5.4). Proposition 2 provides justification.

Proposition 2 (Consistency of Structural Uncertainty Scores). *Assume **(C1)** Flow-SEM identifiability up to Markov equivalence; **(C2)** $\lambda_{\mathrm{dag}} = O(\sqrt{\log n/n})$; **(C3)** bounded in-degree $\max_i |\mathrm{Pa}_{G^*}(Z_i)| \le d_{\max} < \infty$. Then the optimal scores p_{ij} satisfy:*

$$\left|p_{ij} - \mathbb{P}(G_{ij}{=}1 \mid \mathcal{D}_Z)\right| \xrightarrow{p} 0 \quad \textit{and} \quad \mathrm{KL}[q(G; W_{\theta^*}) \| P(G \mid \mathcal{D}_Z)] = O_p\Big(\tfrac{p^2 \log n}{n}\Big).$$

Latent-to-Original Variable Mapping MM-SIGMA establishes direct mapping mechanisms. For structured data latent variables, one-to-one mapping is adopted: $\phi_s(Z_k^{(s)}) = X_k^{(s)}$; for image data latent variables: $\phi_i(Z^{(i)}) = X^{(i)}$. For a latent space path $\pi_Z : Z_a \to Z_b \to Z_c$, the corresponding original space cross-modal path is: $\pi_X : \phi(Z_a) \to \phi(Z_b) \to \phi(Z_c)$. This mapping mechanism ensures that causal relationships identified in latent space accurately reflect causal mechanisms among original cross-modal variables.

Proposition 3 (Latent-to-Original Identifiability for Structured Variables). *Under the encoder in Section 5.1, where each structured variable $X_k^{(s)}$ has a dedicated encoder $f_{enc,k}^{(s)}$ with output $Z_k^{(s)}$, the mapping $\phi_s(Z_k^{(s)}) = X_k^{(s)}$ is identifiable: causal relationships among $\{Z_k^{(s)}\}$ correspond uniquely to causal relationships among $\{X_k^{(s)}\}$. For image variables, the encoder $f_{enc}^{(i)}$ maps $X^{(i)}$ to a summarizing representation $Z^{(i)}$; causal relationships involving $Z^{(i)}$ are interpreted as relationships involving the* causally relevant features *of $X^{(i)}$ captured by encoder, rather than a unique inversion of $X^{(i)}$ itself.*

5.3 Cross-Modal Mediation Pathway Identification

Given a posterior DAG ensemble $\{G_{Z,s}\}_{s=1}^{N_{\text{DAG}}}$, MM-SIGMA identifies cross-modal mediation pathways via Cross-Modal Path Stability Score (CM-PSS). Let $\pi_Z = (Z_0 \to Z_1 \to \cdots \to Z_k)$ denote a directed path from treatment to outcome node. CM-PSS is defined as:

$$\text{CM-PSS}(\pi_Z) = \frac{1}{N_{\text{valid}}} \sum_{s=1}^{N_{\text{DAG}}} \mathbf{I}(\pi_Z \subseteq G_{Z,s})$$

MM-SIGMA enumerates all simple directed paths from A to Y in each valid DAG, retaining those above threshold τ as high-confidence candidates: $\mathit{\Pi}_{\text{stable}} = \{\pi_Z | \text{CM-PSS}(\pi_Z) \geq \tau\}$. Each stable latent path is mapped to an original-space cross-modal path via $\pi_X = \phi(\pi_Z)$, yielding $\mathit{\Pi}_{\text{cross-modal}} = \{\phi(\pi_Z) | \pi_Z \in \mathit{\Pi}_{\text{stable}}\}$ without requiring predefined pathway structures.

Remark 2 (Distinction from Edge Inclusion Probability). Unlike Edge Inclusion Probability (EIP), which operates on individual edges, CM-PSS estimates the marginal posterior probability of entire directed paths, capturing path-level structural uncertainty across heterogeneous latent subspaces. It further integrates the interpretable mapping ϕ for direct causal interpretation in the original variable space, and is an unbiased consistent estimator of path posterior inclusion probability .

Proposition 4 (Identification Error under Approximate Latent Sufficiency). *Let $\varepsilon = \hat{I}(Y; X^{(i)} \mid Z^{(i)}, A, M^{(s)}, C)$ and let $B = \sup |Y| < \infty$. Then:*

$$|\widehat{\text{CM-NDE}} - \text{CM-NDE}| \leq B\sqrt{2\varepsilon}, \quad |\widehat{\text{CM-NIE}} - \text{CM-NIE}| \leq B\sqrt{2\varepsilon} \tag{3}$$

by Pinsker's inequality applied to $\mathbb{E}[\mathrm{TV}(P(Y \mid X^{(i)}, \ldots), P(Y \mid Z^{(i)}, \ldots))]$. *As encoder capacity increases,* $\varepsilon \to 0$ *and both bounds vanish. In the HPP setting,* Y *is a standardized cardiovascular risk score bounded in* $[0, 1]$*, and the 12 pre-extracted vessel indicators constitute a domain-validated sufficient statistic for cardiovascular outcomes [38], providing empirical support that* ε *is small in practice.*

5.4 Multimodal Effect Estimation and Uncertainty Propagation

Upon identifying the stable cross-modal mediation pathway set, MM-SIGMA estimates pathway-specific CM-NDE and CM-NIE, accounting for multi-level uncertainties: representation learning, structure discovery, pathway identification, and effect estimation. This achieves end-to-end uncertainty propagation through structure-specific nuisance estimation, extended efficient influence functions, and Bayesian model averaging.

Structure-Guided Nuisance Estimation For each stable pathway $\pi_X \in \mathit{\Pi}_{\text{cross-modal}}$, MM-SIGMA estimates nuisance functions $\hat{\eta}_s$ leveraging structure-specific conditioning sets $\mathrm{Pa}_{G_{Z,s}}(\cdot)$ from supporting DAG $G_{Z,s}$:

$$\hat{\pi}_s(A = a|C) = \hat{\mathbb{P}}(A = a|\mathrm{Pa}_{G_{Z,s}}(A))$$

$$\hat{m}_s(M^{(s)}, M^{(i)}|A, C) = \hat{\mathbb{P}}(M^{(s)}, M^{(i)}|A, \mathrm{Pa}_{G_{Z,s}}(M^{(s)}, M^{(i)}))$$

$$\hat{\mu}_{Y,s}(A, M^{(s)}, M^{(i)}, C) = \hat{\mathbb{E}}[Y|A, M^{(s)}, M^{(i)}, \mathrm{Pa}_{G_{Z,s}}(Y)]$$

where $\hat{\pi}_s$ is the structure-specific propensity score, $\hat{m}_s$ is the conditional mediator distribution, and $\hat{\mu}_{Y,s}$ is the outcome regression. Neural networks fit these via K-fold cross-fitting for bias reduction, yielding counterfactual predictions $\hat{\mu}_Y^{(i)}(a, M(a^*))$ for downstream EIF and plugin estimators.

Extended Efficient Influence Functions Given $\hat{\eta}_s$, MM-SIGMA computes single-DAG effects $\hat{\theta}_s(\pi_X) = [\widehat{\text{CM-NDE}}_s(\pi_X), \widehat{\text{CM-NIE}}_s(\pi_X)]^T$ via a hybrid strategy: extended influence functions for $L(\pi_X) = 3$ (i.e., $A \to M \to Y$), and plugin estimators for $L(\pi_X) > 3$. All effects compare pre-specified exposure a to reference a^*. For binary A, levels are 1 and 0; for continuous A, $a^* = \tilde{\mu}_A$ and $a = \tilde{\mu}_A + k \cdot \tilde{\sigma}_A$ $(k = 1)$, with $\tilde{\mu}_A$, $\tilde{\sigma}_A$ from the first K-fold training set.

Extended Influence Function Estimation $(L(\pi_X) = 3)$: For $L(\pi_X) = 3$, the estimand is the natural indirect effect under nested counterfactuals (Section 4, Theorem 1), where cross-world independence holds under the single-mediator Ignorability condition (iii). The point estimate $\hat{\theta}_s(\pi_X)$ is the sample mean of influence function scores $\psi_i(Z_i; \hat{\eta}_s, G_{Z,s})$, i.e., $\hat{\theta}_s(\pi_X) = \frac{1}{n}\sum_{i=1}^n \psi_i(Z_i; \hat{\eta}_s, G_{Z,s})$. MM-SIGMA extends influence functions in multimodal settings by constructing bias-corrected counterfactual expectation estimates through integrating cross-modal nuisance function components. These include cross-modal propensity scores, mediator models, and outcome prediction functions. The single-DAG variance $\hat{V}(\hat{\theta}_s(\pi_X))$ is estimated by $\frac{1}{n^2}\sum_{i=1}^n(\psi_i - \bar{\psi})^2$, where $\bar{\psi} = \hat{\theta}_s(\pi_X)$. **Plugin Estimation** $(L(\pi_X) > 3)$: For longer paths, cross-world independence assumptions

accumulate with each additional node, rendering nested counterfactual identification fragile. MM-SIGMA therefore adopts the *interventional indirect effect* formulation [42], which requires only sequential ignorability and avoids cross-world independence. The estimand is:

$$\text{CM-NIE}_s(\pi_X) = \int \mu_Y(a, m, c)\big[dF_{M|a,c} - dF_{M|a^*,c}\big]\, dF_C$$

estimated via plug-in:

$$\widehat{\text{CM-NIE}}_s(\pi_X) = \frac{1}{n}\sum_{i=1}^{n} \Delta_{\text{NIE}}^{(i)}$$

where $\Delta_{\text{NIE}}^{(i)} = \hat{\mu}_Y^{(i)}(a, M(a)) - \hat{\mu}_Y^{(i)}(a, M(a^*))$. Its variance is calculated as:

$$\hat{V}(\hat{\theta}_s(\pi_X)) = \frac{1}{n}\hat{\sigma}^2_\Delta, \quad \hat{\sigma}^2_\Delta = \frac{1}{n-1}\sum_{i=1}^{n}(\Delta_{CF}^{(i)} - \bar{\Delta}_{CF})^2$$

where $\Delta_{CF}^{(i)}$ denotes individual-level counterfactual differences. **End-to-End Uncertainty Propagation** To generate final mediation effect estimates that account for structural uncertainty, MM-SIGMA aggregates single-DAG estimates $\hat{\theta}_s(\pi_X)$ and their variances $\hat{V}(\hat{\theta}_s(\pi_X))$ across supporting DAGs via Bayesian model averaging. Let N_{path} be the number of DAGs supporting pathway π_X. The Bayesian model averaged point estimate is $\hat{\theta}_{\text{BMA}}(\pi_X) = \frac{1}{N_{\text{path}}}\sum_{s=1}^{N_{\text{path}}} \hat{\theta}_s(\pi_X)$, using uniform weights for each supporting DAG $G_{Z,s}$. The total BMA variance is computed via the law of total variance:

$$\hat{V}(\hat{\theta}_{\text{BMA}}(\pi_X)) = \hat{V}_{\text{within}} + \hat{V}_{\text{between}}$$

where $\hat{V}_{\text{within}} = \frac{1}{N_{\text{path}}}\sum_{s=1}^{N_{\text{path}}} \hat{V}(\hat{\theta}_s(\pi_X))$ reflects estimation uncertainty for each structure, and $\hat{V}_{\text{between}} = \frac{1}{N_{\text{path}}}\sum_{s=1}^{N_{\text{path}}}(\hat{\theta}_s(\pi_X) - \hat{\theta}_{\text{BMA}}(\pi_X))^2$ captures structural uncertainty across structures. Confidence intervals are constructed using standard normal approximation: $\hat{\theta}_{\text{BMA}}(\pi_X) \pm z_{1-\alpha/2}\sqrt{\hat{V}(\hat{\theta}_{\text{BMA}}(\pi_X))}$. This framework propagates estimation (within-DAG) and structural (between-DAG) uncertainty into final cross-modal mediation effect estimates, providing robust uncertainty quantification for multimodal causal inference. Uniform weights are adopted following the Bayesian model averaging literature under prior DAG uncertainty [10]: when the structural posterior $P(G \mid D_Z)$ is approximated by the variational distribution $q(G; W_\theta^*)$, likelihood-based weights would require evaluating the marginal likelihood per DAG, which is intractable in the Flow-SEM setting. The resulting additional variance is absorbed into $\hat{V}_{\text{between}}$ via the law of total variance, preserving valid uncertainty propagation.

6 Experiment

We generate multiple synthetic datasets to comprehensively evaluate MM-SIGMA performance. Table 1 summarizes the configuration parameters. Each configuration generates 10 independent datasets with algorithms running 50 Monte Carlo repetitions. Hyperparameters are optimized through grid search and 5-fold cross-validation. All methods share the same multimodal encoder input layer, ensuring performance differences reflect algorithmic design rather than input representation; no existing method handles cross-modal structure discovery and mediation estimation jointly, necessitating the multimodal-adapted baselines described above.

Table 1: Dataset Configuration.

Parameter	Values
d_s	10, 20, 50
$H \times W$	64^2, 128^2
C	1, 3
n	1k, 5k, 10k
Relationship	Lin., Nonlin., Mixed
σ_{obs}	0.1, 0.2, 0.5
Cross-modal edges	25%–35%†

†Set to reflect cross-system association proportions in multimodal biological networks [19].

Experimental Setup. Treatment A is the primary upstream sleep variable (AHI for known pathways; sleep efficiency and arousal index for novel pathways) and outcome Y is the cardiovascular endpoint (carotid IMT or composite CVD risk score). Covariates C include age, sex, BMI, and smoking status [13, 27]. Missing values are handled via MICE [5], with missingness rates below 8%.

6.1 Synthetic Multimodal Dataset Construction

Structure discovery is evaluated by adjacency $F1_{\text{adj}}$, orientation $F1_{\text{ori}}$, cross-modal edge $F1_{\text{cm}}$, and causal accuracy ($\text{Acc}_{\text{causal}} = |G_{\text{pred}} \cap G_{\text{true}}|/|G_{\text{true}}|$). Mediation analysis is evaluated by pathway recovery rate ($\text{PR} = |\mathcal{P}_{\text{recovered}} \cap \mathcal{P}_{\text{true}}|/|\mathcal{P}_{\text{true}}|$), CM-NDE/NIE bias, and 95% CI coverage rate.

6.2 Results

Table 2 shows that cross-modal edge identification is harder than intra-modal, with all methods scoring lower on CM-Edge F1 than Adjacency F1. Multimodal adaptation outperforms concatenation based baselines (MM-NOTEARS: +36.9% over Concat-NOTEARS), confirming the importance of preserving modal structure. Orientation identification remains difficult across all methods; even MM-DECI reaches only 0.348 in Ori F1. MM-SIGMA achieves 0.421 (+21.0% over best baseline) through end-to-end learning, with gains across all metrics despite absence of ground-truth graph supervision.

Table 2: Cross-modal Structure Discovery. Results over 50 Monte Carlo runs.

Method	Adj F1	Ori F1	CM-Edge F1	Acc
MM-PC	0.342±0.045	0.287±0.052	0.234±0.067	0.156±0.043
MM-NOTEARS	0.389±0.038	0.312±0.041	0.278±0.053	0.201±0.037
MM-ICALiNGAM	0.356±0.042	0.298±0.048	0.251±0.059	0.178±0.041
MM-DECI	0.412±0.035	0.348±0.039	0.315±0.048	0.267±0.033
Concat-NOTEARS	0.321±0.047	0.256±0.054	0.203±0.068	0.142±0.045
Concat-DECI	0.345±0.043	0.278±0.049	0.225±0.061	0.168±0.042
CausalMixNet	0.367±0.041	0.289±0.045	0.263±0.055	0.189±0.039
MM-SIGMA	**0.487±0.028**	**0.421±0.032**	**0.398±0.035**	**0.345±0.029**

Table 3 reveals differential impacts of data conditions and pathway complexity. High-resolution images pose greater challenges than high-dimensional structured data, and performance gains diminish from medium to large samples, suggesting an information saturation point. MM-SIGMA exhibits only 15.4%

Table 3: Pathway Recovery Rate under Different Conditions. Top: robustness across data conditions. Bottom: identification performance by pathway type. S=Structured variables, I=Image variables.

Condition / Pathway Type	MM-SIGMA	MM-DECI+EIF	MM-NOTEARS+DeepMed	CausalMixNet+EIF
Data Conditions				
Small Sample	0.421±0.045	0.312±0.052	0.298±0.048	0.267±0.051
Medium Sample	0.487±0.028	0.389±0.038	0.356±0.041	0.321±0.042
Large Sample	0.512±0.022	0.431±0.029	0.398±0.035	0.362±0.038
High-dim Structured	0.468±0.031	0.367±0.041	0.342±0.044	0.298±0.047
High-res Images	0.453±0.033	0.348±0.043	0.329±0.046	0.285±0.049
Strong Noise	0.412±0.038	0.298±0.045	0.276±0.048	0.245±0.052
Pathway Types				
S→I→S	0.72±0.05	0.54±0.07	0.49±0.08	0.46±0.09
I→S→S	0.68±0.06	0.51±0.08	0.46±0.09	0.43±0.10
S→S→I	0.71±0.05	0.53±0.07	0.48±0.08	0.45±0.09
S→I→S→I	0.63±0.07	0.42±0.09	0.38±0.10	0.35±0.11
Multiple Parallel	0.69±0.06	0.48±0.08	0.43±0.09	0.40±0.10

Table 4: Validation Results on HPP Dataset. HRV=Heart Rate Variability, PWV=Pulse Wave Velocity, IMT=Intima-Media Thickness, OSA=Obstructive Sleep Apnea, CVD=Cardiovascular Disease, BP=Blood Pressure. †Methods require pre-specified pathway structures and cannot discover cross-modal pathways.

		MM-SIGMA		**MM-DECI+EIF**		**MM-NOTEARS+DeepMed**		**CausalMixNet+EIF**		**Logistic Regression**	
Pathway Type	Pathway Description	NDE	NIE	NDE	NIE	NDE	NIE	NDE	NIE	NDE	NIE
Known Pathways	Sleep HRV→PWV→Carotid IMT	−0.089±0.028	−0.156±0.034	−0.071±0.039	−0.118±0.048	−0.063±0.043	−0.107±0.055	−0.068±0.041	−0.112±0.051	−0.067±0.045	−0.098±0.067
	OSA Severity→Inflammation→CVD Risk	−0.054±0.021	−0.089±0.027	−0.041±0.033	−0.068±0.042	−0.037±0.037	−0.061±0.048	−0.039±0.035	−0.064±0.045	−0.043±0.038	−0.071±0.051
Novel Pathways	Sleep→Gut Microbiome→Metabolism→CVD	−0.045±0.019	−0.078±0.025	N/A†	N/A†	N/A†	N/A†	N/A†	N/A†	N/A†	N/A†
	Retinal Features→Body Fat→BP→IMT	−0.032±0.015	−0.067±0.022	N/A†	N/A†	N/A†	N/A†	N/A†	N/A†	N/A†	N/A†

degradation under strong noise versus >25% for all baselines. Among pathway types, single cross-modal transition pathways (S→I→S, I→S→S, S→S→I) perform similarly while multi-transition paths show significant degradation, indicating cumulative identification difficulty per modal transition. MM-SIGMA consistently outperforms all baselines across every condition and pathway type.

Table 5 highlights key differences between end-to-end and two-stage frameworks. Two-stage methods exhibit "structure-effect inconsistency": MM-DECI discovers structure well, but MM-DECI+EIF reaches only 0.64 pathway recovery, suggesting loss in uncertainty propagation. MM-SIGMA achieves 0.94±0.02 coverage, slightly below nominal 0.95 yet outperforming two-stage (0.89–0.90) and predefined-pathway methods (0.85–0.88). This undercoverage is consistent with finite-sample bias in plug-in variance estimation under structural uncertainty, where $\hat{V}_{\text{between}}$ underestimates between-DAG dispersion when N_{path} is small, and aligns with the low NDE/NIE bias. Ablations show $\mathcal{L}_{\text{causal}}$ drives pathway recovery, while BMA is crucial for coverage; τ=0.5 yields the best overall balance.

6.3 HPP Real-world Data Validation

Validation on HPP dataset covers pathway replication and novel pathway discovery (Table 4).

Known Pathways. The sleep HRV→PWV→carotid IMT pathway (CM-NDE: −0.089±0.028, CM-NIE: −0.156±0.034, p<0.001) is consistent with [14]. The OSA→inflammation→CVD pathway (CM-NDE: −0.054±0.021, CM-NIE: −0.089±0.027, p<0.005) aligns with established mechanisms [13, 27]. Effect match with reference values was confirmed via one-sample t-tests (p<0.05).

Table 5: End-to-end mediation analysis. $p <$ 0.01 vs. best baseline (paired t-test, 50 runs).

Method	Recov.	NDE Bias	NIE Bias	Cov.
Full Model				
MM-SIGMA	0.74±0.04	0.012±0.008	0.015±0.009	0.94±0.02
Ablations				
w/o $\mathcal{L}_{\text{causal}}$	0.61±0.06	0.031±0.014	0.034±0.015	0.90±0.03
w/o $\mathcal{L}_{\text{align}}$	0.65±0.05	0.026±0.012	0.029±0.013	0.91±0.03
w/o BMA	0.71±0.05	0.018±0.010	0.021±0.011	0.87±0.04
τ=0.3	0.79±0.05	0.024±0.011	0.027±0.012	0.91±0.03
τ=0.5	0.74±0.04	0.012±0.008	0.015±0.009	0.94±0.02
τ=0.7	0.68±0.05	0.015±0.009	0.018±0.010	0.93±0.02
Two-stage Baselines				
MM-NOTEARS+ DeepMed	0.61±0.06	0.045±0.021	0.041±0.023	0.89±0.04
MM-DECI+EIF	0.64±0.05	0.039±0.019	0.037±0.020	0.90±0.03
Single-stage Baselines				
MM-PC+ Baron-Kenny	0.42±0.08	0.087±0.032	0.094±0.038	0.82±0.06
MM-DeepMed	0.58±0.06	0.052±0.024	0.048±0.026	0.88±0.04
MM-Lasso	0.51±0.07	0.061±0.027	0.057±0.028	0.86±0.05
MM-RF	0.54±0.08	0.068±0.030	0.063±0.031	0.85±0.06

Methodological Comparison. MM-SIGMA yields significantly narrower confidence intervals than logistic regression (-0.156 ± 0.034 vs. -0.098 ± 0.067) and outperforms the best baseline MM-DECI+EIF (-0.118 ± 0.048). Baselines requiring pre-specified structures and logistic regression on image modalities cannot handle novel pathways, confirming MM-SIGMA's ability.

Novel Pathways. CM-PSS filtering (Section 5.3) restricts inference to pathways with marginal posterior inclusion probability $\geq \tau$, providing multiplicity control analogous to Bayesian variable selection [32]. Among retained pathways, the retinal features→body fat→BP→carotid IMT pathway (CM-NIE: -0.067 ± 0.022, $p<0.01$) is consistent with established cross-modal links: retinal vessel morphology is associated with adiposity measures [38], and body fat accumulation is a known mediator of hypertension and arterial stiffness [21]. The sleep→gut microbiome→metabolism→CVD pathway (CM-NIE: -0.078 ± 0.025, $p<0.01$) aligns with converging evidence that sleep disruption alters gut microbial function through circadian dysregulation, producing metabolites such as TMAO and SCFAs that elevate cardiometabolic risk [22, 23]. These pathways represent cross-modal mechanisms not discoverable by single-modality methods, with stability confirmed via 5-fold cross-validation (CV < 0.12).

7 Related Work and Conclusion

Traditional mediation analysis methods [3, 35] rely on linear assumptions and single-modal data, while semiparametric extensions [30, 39] improve robustness but remain confined to predefined pathway structures. Multimodal representation learning approaches [2,29] optimize reconstruction objectives without causal constraints, disrupting causal dependencies [4]. Recent causal discovery methods [46, 47] and multimodal causal models such as CausalMixNet [45] address specific aspects of this problem but rely on pre-specified structures and do not establish nonparametric identification theory for cross-modal mediation effects.

The proposed MM-SIGMA framework enables cross-modal causal mediation analysis through end-to-end learning. Results show superior performance in cross-modal edge identification (F1: 0.398) and pathway recovery (0.74) compared to existing methods. HPP dataset validation confirmed effectiveness in replicating known pathways and discovering novel cross-modal mechanisms, providing valuable tools for precision medicine applications.

References

1. Baharum, S.N., Azizan, K.A.: Metabolomics in systems biology. In: Advances in Experimental Medicine and Biology, vol. 1102, pp. 51–68. Springer (2018). https://doi.org/10.1007/978-3-319-98758-3_4
2. Baltrusaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. **41**(2), 423–443 (Feb 2019). https://doi.org/10.1109/TPAMI.2018.2798607, https://doi.org/10.1109/TPAMI.2018.2798607
3. Baron, R.M., Kenny, D.A.: The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology **51**(6), 1173–1182 (1986). https://doi.org/10.1037/0022-3514.51.6.1173
4. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (Aug 2013). https://doi.org/10.1109/TPAMI.2013.50, https://doi.org/10.1109/TPAMI.2013.50
5. van Buuren, S., Groothuis-Oudshoorn, K.: mice: Multivariate imputation by chained equations in r. Journal of Statistical Software **45**(3), 1–67 (2011). https://doi.org/10.18637/jss.v045.i03, https://www.jstatsoft.org/index.php/jss/article/view/v045i03
6. Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J.: Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal **21**(1), C1–C68 (01 2018). https://doi.org/10.1111/ectj.12097, https://doi.org/10.1111/ectj.12097
7. Daniel, R.M., De Stavola, B.L., Cousens, S.N., Vansteelandt, S.: Causal mediation analysis with multiple mediators. Biometrics **71**(1), 1–14 (Mar 2015). https://doi.org/10.1111/biom.12248
8. Glymour, C., Zhang, K., Spirtes, P.: Review of causal discovery methods based on graphical models. Frontiers in Genetics **10**, 524 (2019). https://doi.org/10.3389/fgene.2019.00524
9. Hasin, Y., Seldin, M., Lusis, A.J.: Multi-omics approaches to disease. Genome Biology **18**(1), 83 (2017). https://doi.org/10.1186/s13059-017-1215-1
10. Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T.: Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and ei george, and a rejoinder by the authors. Statistical science **14**(4), 382–417 (1999)
11. Ikram, M.K., de Jong, F.J., Bos, M.J., Vingerling, J.R., Hofman, A., Koudstaal, P.J., de Jong, P.T., Breteler, M.M.B.: Retinal vessel diameters and risk of stroke: the rotterdam study. Neurology **66**(9), 1339–1343 (May 2006). https://doi.org/10.1212/01.wnl.0000210533.24338.ea
12. Imai, K., Keele, L., Tingley, D.: A general approach to causal mediation analysis. Psychological Methods **15**(4), 309–334 (2010). https://doi.org/10.1037/a0020761, https://pubmed.ncbi.nlm.nih.gov/20954780/
13. Javaheri, S., Barbe, F., Campos-Rodriguez, F., Dempsey, J.A., Khayat, R., Javaheri, S., et al.: Sleep apnea: types, mechanisms, and clinical cardiovascular consequences. Journal of the American College of Cardiology **69**(7), 841–858 (2017)
14. Kadoya, M., Koyama, H., Kurajoh, M., Kanzaki, A., Kakutani-Hatayama, M., Okazaki, H., Shoji, T., Moriwaki, Y., Yamamoto, T., Emoto, M., Inaba, M., Namba, M.: Sleep, cardiac autonomic function, and carotid atherosclerosis in patients with cardiovascular risks: Hscaa study. Atherosclerosis **238**(2), 409–414 (Feb

2015). https://doi.org/10.1016/j.atherosclerosis.2014.12.032, epub 2014 Dec 20
15. Kocaoglu, M., Snyder, C., Dimakis, A.G., Vishwanath, S.: Causalgan: Learning causal implicit generative models with adversarial training (2017), https://arxiv.org/abs/1709.02023
16. Kumar, V., Abbas, A.K., Aster, J.C.: Robbins & Cotran Pathologic Basis of Disease. Elsevier Health Sciences, 9 edn. (2014)
17. Lahat, D., Adali, T., Jutten, C.: Multimodal data fusion: An overview of methods, challenges, and prospects. Proceedings of the IEEE **103**(9), 1449–1477 (2015). https://doi.org/10.1109/JPROC.2015.2460697
18. McGeechan, K., Liew, G., Macaskill, P., Irwig, L., Klein, R., Klein, B.E., Wang, J.J., Mitchell, P., Vingerling, J.R., de Jong, P.T., Witteman, J.C., Breteler, M.M., Shaw, J., Zimmet, P., Wong, T.Y.: Prediction of incident stroke events based on retinal vessel caliber: a systematic review and individual-participant meta-analysis. American Journal of Epidemiology **170**(11), 1323–1332 (Dec 2009). https://doi.org/10.1093/aje/kwp306
19. Menche, J., Sharma, A., Kitsak, M., Ghiassian, S.D., Vidal, M., Loscalzo, J., Barabási, A.L.: Uncovering disease-disease relationships through the incomplete interactome. Science **347**(6224), 1257601 (2015). https://doi.org/10.1126/science.1257601, https://www.science.org/doi/abs/10.1126/science.1257601
20. Morgan, S.L., Winship, C.: Counterfactuals and Causal Inference: Methods and Principles for Social Research. Analytical Methods for Social Research, Cambridge University Press, 2 edn. (2014)
21. Natale, F., Tedesco, M.A., Mocerino, R., de Simone, V., Di Marco, G.M., Aronne, L., Credendino, M., Siniscalchi, C., Calabrò, P., Cotrufo, M., Calabrò, R.: Visceral adiposity and arterial stiffness: echocardiographic epicardial fat thickness reflects, better than waist circumference, carotid arterial stiffness in a large population of hypertensives. European Journal of Echocardiography **10**(4), 549–555 (02 2009). https://doi.org/10.1093/ejechocard/jep002, https://doi.org/10.1093/ejechocard/jep002
22. Pala, B., Pennazzi, L., Nardoianni, G., Fogacci, F., Cicero, A.F.G., Di Renzo, L., Barbato, E., Tocci, G.: Gut microbiota dysbiosis and sleep disorders: Culprit in cardiovascular diseases. Journal of Clinical Medicine **13**(11), 3254 (2024). https://doi.org/10.3390/jcm13113254
23. Parkar, S.G., Kalsbeek, A., Cheeseman, J.F.: Potential role for the gut microbiota in modulating host circadian rhythms and metabolic health. Microorganisms **7**(2), 41 (2019). https://doi.org/10.3390/microorganisms7020041
24. Pearl, J.: Direct and indirect effects. In: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. p. 411–420. UAI'01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)
25. Pearl, J.: Causality: Models, Reasoning and Inference. Cambridge University Press, 2 edn. (2009). https://doi.org/10.5555/1642718
26. Pearl, J.: The causal mediation formula—a guide to the assessment of pathways and mechanisms. Prevention science **13**(4), 426–436 (2012)
27. Peppard, P.E., Young, T., Palta, M., Skatrud, J.: Prospective study of the association between sleep-disordered breathing and hypertension. New England Journal of Medicine **342**(19), 1378–1384 (2000)
28. Peters, J., Janzing, D., Schölkopf, B.: Elements of Causal Inference: Foundations and Learning Algorithms. The MIT Press (2017). https://doi.org/10.5555/3202377

29. Ramachandram, D., Taylor, G.W.: Deep multimodal learning: A survey on recent advances and trends. IEEE Signal Processing Magazine **34**(6), 96–108 (2017). https://doi.org/10.1109/MSP.2017.2738401
30. Robins, J.M., Greenland, S.: Identifiability and exchangeability for direct and indirect effects. Epidemiology **3**(2), 143–155 (1992). https://doi.org/10.1097/00001648-199203000-00013
31. Schadt, E.E.: Molecular networks as sensors and drivers of common human diseases. Nature **461**(7261), 218–223 (2009). https://doi.org/10.1038/nature08454, https://doi.org/10.1038/nature08454
32. Scott, J.G., Berger, J.O.: Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. The Annals of Statistics **38**(5) (Oct 2010). https://doi.org/10.1214/10-aos792, http://dx.doi.org/10.1214/10-AOS792
33. Shorten, C., Khoshgoftaar, T.M.: A survey on Image Data Augmentation for Deep Learning. Journal of Big Data **6**(1), 60 (2019). https://doi.org/10.1186/s40537-019-0197-0
34. Shpitser, I., Pearl, J.: Complete identification methods for the causal hierarchy. J. Mach. Learn. Res. **9**, 1941–1979 (Jun 2008)
35. Sobel, M.E.: Asymptotic confidence intervals for indirect effects in structural equation models. Sociological Methodology **13**, 290–312 (1982). https://doi.org/10.2307/270723
36. Song, S.Y., He, K., Narla, R.R., Kang, H.G., Ryu, H.U., Boyko, E.J.: Metabolic consequences of obstructive sleep apnea especially pertaining to diabetes mellitus and insulin sensitivity. Diabetes & Metabolism Journal **43**(2), 144–155 (Apr 2019). https://doi.org/10.4093/dmj.2018.0256
37. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search. The MIT Press, 2 edn. (2001). https://doi.org/10.7551/mitpress/1754.001.0001
38. Tapp, R.J., Owen, C.G., Barman, S.A., Welikala, R.A., Foster, P.J., Whincup, P.H., Strachan, D.P., Rudnicka, A.R.: Retinal vascular tortuosity and diameter associations with adiposity and components of body composition. Obesity (Silver Spring) **28**(9), 1750–1760 (2020). https://doi.org/10.1002/oby.22885
39. Tchetgen Tchetgen, E.J., Shpitser, I.: Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. Annals of Statistics **40**(3), 1816–1845 (Jun 2012). https://doi.org/10.1214/12-AOS990
40. Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine **25**(1), 44–56 (2019). https://doi.org/10.1038/s41591-018-0300-7, https://doi.org/10.1038/s41591-018-0300-7
41. VanderWeele, T.J.: Explanation in causal inference: methods for mediation and interaction. Oxford University Press (2015)
42. VanderWeele, T.J.: Explanation in causal inference: developments in mediation and interaction. International Journal of Epidemiology **45**(6), 1904–1908 (Dec 2016). https://doi.org/10.1093/ije/dyw277
43. Wang, J.J., Liew, G., Wong, T.Y., Smith, W., Klein, R., Leeder, S.R., Mitchell, P.: Retinal vascular calibre and the risk of coronary heart disease-related death. Heart **92**(11), 1583–1587 (Nov 2006). https://doi.org/10.1136/hrt.2006.090522
44. Zhang, C., Bütepage, J., Kjellström, H., Mandt, S.: Advances in variational inference. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(8), 2008–2026 (2019). https://doi.org/10.1109/TPAMI.2018.2889774
45. Zhang, H., Puyol-Antón, E., Jacob, J., Marsden, P., Glocker, B., Bai, W.: Causalmixnet: Weakly supervised mediators for debiased causal learning in multimodal medical ai. Medical Image Analysis **98**, 103371 (2025). https://doi.org/10.1016/j.media.2024.103371

46. Zheng, X., Aragam, B., Ravikumar, P., Xing, E.P.: Dags with no tears: continuous optimization for structure learning. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 9492–9503. NIPS'18, Curran Associates Inc., Red Hook, NY, USA (2018)
47. Zhu, S., Ng, I., Chen, Z.: Causal discovery with reinforcement learning (2020), https://arxiv.org/abs/1906.04477