FROM REPRESENTATION IDENTIFIABILITY TO COUNTERFACTUAL IDENTIFIABILITY IN LATENT STRUCTURAL CAUSAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper investigates whether latent structural causal models (SCMs) learned from factual data can be trusted on counterfactuals, that is, whether latent SCMs that fit the same factual data agree on every counterfactual. To answer this question, we define counterfactual identifiability on latent SCMs, introduce the latent SCM isomorphism that implies it, and establish sufficient assumptions for reaching the isomorphism: an auxiliary variable that modulates the exogenous variables and one perfect intervention on each non-root endogenous variable. We further show that the isomorphism supports causal effect analysis at the counterfactual level, identifying natural direct and indirect effects up to a common scaling. Our theoretical results extend causal representation learning by turning representation identification, in the form of the isomorphism, into a guarantee for counterfactual identification. Finally, we develop FlowSCM, an estimator built on our assumptions and theory, and show on synthetic and image data that it recovers the latent SCM and returns its counterfactuals and causal effects.

1 INTRODUCTION

Would this patient avoid heart failure had they started treatment a year earlier? This is a *counterfactual* question, asking what would have happened to one individual under an action that did not occur. Counterfactuals sit at the top of Pearl's causal hierarchy (Pearl, 2009). Answering one takes three steps on a structural causal model (SCM): *abducting* the *exogenous variable* specific to an individual, *acting* on the SCM to alter its mechanism, and *predicting* the *endogenous variables* via the altered SCM on the abducted exogenous variable. *Counterfactual identification* asks if a learned SCM can be trusted on counterfactuals: whether, under reasonable assumptions, two SCMs agreeing on factual data distributions would agree on counterfactual distributions (Pearl, 2009, Def. 3.2.3).

While studied extensively (Correa et al., 2021; Khemakhem et al., 2021; Javaloy et al., 2023; Chen & Du, 2025), most progress in counterfactual identification is established when causal endogenous variables are directly observed. In practice, however, causal variables can be hidden under high-dimensional observations, such as medical images and records. Recovering latent causal variables and their SCMs from such observations is the focus of a vibrant research area known as causal representation learning (CRL) (Schölkopf et al., 2021). Most existing CRL works focus on identifying the latent SCM or parts of it, leveraging auxiliary labels of the causal variables (Yang et al., 2021; Komanduri et al., 2024), interventions (Ahuja et al., 2023; von Kügelgen et al., 2023), multiple views (Yao et al., 2024), or temporal structure (Yao et al., 2022; Lippe et al., 2022b). Measured against the three steps of a counterfactual, these results serve action and prediction but leave aside the abduction of exogenous variables, leaving counterfactuals under-explored in latent SCMs.

In two existing CRL works, counterfactuals enter as data: ILCM (Brehmer et al., 2022) identifies the latent SCM, including exogenous variables, from observed pairs before and after a perfect intervention. Because such pairs are often lacking in nature, Liu et al. (2023) explore the concept of "actionable counterfactuals" where an image pair provides counterfactual observations only for some object-level variables. SuaVE (Liu et al., 2026; 2024) recovers the exogenous variables with an auxiliary variable under linear or polynomial mechanisms, although without discussing counterfactuals. Only one recent work treats counterfactuals as the objective in latent SCMs: Zhou et al.

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

(2024) define counterfactual equivalence in the observation space and show that it does not require recovery of the latent SCM. However, this work does not address counterfactual identifiability, and the equivalence is restricted to counterfactuals from interventions present in the factual data. A prominent theoretical gap thus remains: counterfactual identifiability of latent SCMs under unseen interventions.

In this paper, we bridge this gap by examining the counterfactual identifiability of latent SCMs. Motivated by the work of Chen & Du (2025) that examines the counterfactual identifiability of SCMs with *observed* causal variables, we introduce the definition of counterfactual identifiability for latent SCMs, and develop an equivalence relation on the latent SCMs — *latent SCM isomorphism* — that implies counterfactual identifiability (Theorem 1). The isomorphism requires the recovery of the exogenous and endogenous causal variables up to a permutation and element-wise affine maps, a class that we show is tight for counterfactuals (Lemma 2). We further show that two isomorphic latent SCMs return the same natural direct and indirect causal effects (Robins & Greenland, 1992; Pearl, 2001) up to a common scaling factor (Proposition 3). We then establish the assumptions that induce latent SCM isomorphism, for an additive-noise SCM under an injective mixing. We show that an auxiliary variable (*exogenous index*) that modulates the distribution of the exogenous variables identifies them up to a permutation and element-wise affine maps. Building on this, we show that one perfect intervention per non-root node, with its target unknown, identifies the endogenous causal variables up to the same maps and recovers the graph. Together they establish latent SCM isomorphism (Theorem 4) and with it counterfactual identifiability. Finally, we introduce FlowSCM, a flow-based estimator built on Theorem 4, such that maximum likelihood returns the exogenous variables, the causal variables, and the graph (§6). Through experiments on synthetic latent SCMs and two image datasets, we test FlowSCM for identification of the latent SCM and estimation of counterfactuals and causal effects. We further ablate each assumption that we establish for inducing latent SCM isomorphism, and compare FlowSCM with ILCM and SuaVE on data that meet each model's assumptions (§7). The main contributions of this work include:

1. **A framework for counterfactual identification with latent causal variables.** We define counterfactual identifiability for latent SCMs and prove that *latent SCM isomorphism* implies it (Theorem 1), that its affine class is tight (Lemma 2), and that two isomorphic latent SCMs return the same natural direct and indirect effects up to a common scaling factor (Proposition 3).
2. **Reaching the isomorphism.** We show that the exogenous index and one perfect intervention per non-root node with its target unknown identify the exogenous and endogenous causal variables and recover the graph, inducing latent SCM isomorphism (Theorem 4).
3. **Estimator and experiments.** We build FlowSCM, a normalizing flow structured as a latent SCM and built on Assumptions 1 and 2 of Theorem 4 (§6), and show on synthetic and image data that it recovers the latent SCM, the counterfactuals, and the causal effects (§7).

2 Related Works

Identifiability in CRL: CRL recovers latent causal variables and graphs from their mixed observations. Because observational data alone is not sufficient for identifiability (Hyvärinen & Pajunen, 1999; Locatello et al., 2019), additional information must be leveraged. Interventions are most commonly used: one or two interventions per node identify the causal variables up to a permutation and element-wise diffeomorphisms, under a linear (Squires et al., 2023; Varıcı et al., 2025), polynomial (Ahuja et al., 2023) or nonlinear mixing (Buchholz et al., 2023; von Kügelgen et al., 2023; Varıcı et al., 2024); soft interventions, or several mechanisms changing at once, may leave a variable mixed with some of its ancestors or neighbours (Zhang et al., 2023; 2024; Ng et al., 2025). Additional sources of information include labels of causal variables (Yang et al., 2021), multiple views (Yao et al., 2024), and temporal structure (Yao et al., 2022), possibly with known intervention targets (Lippe et al., 2022b). Identification of exogenous variables is less examined, using paired data before and after an intervention (Brehmer et al., 2022), or an auxiliary variable that modulates the SCM along with linear or polynomial mechanisms (Liu et al., 2026; 2024). All of these results stop at recovering the latent SCM up to an equivalence class; the next step from this representation identifiability to counterfactual identifiability is yet to be addressed in the context of latent SCMs.

Counterfactual identification: With the causal variables observed and the graph known, Correa et al. (2021) give a graphical condition and an algorithm for identifying nested counterfactuals,

such as the natural direct effect, from observational and interventional distributions. For a learned SCM, Nasr-Esfahany & Kiciman (2023) show counterfactual identifiability impossible for general mechanisms with multi-dimensional exogenous variables; for bijective mechanisms, which include additive-noise mechanisms, Nasr-Esfahany et al. (2023) establish it in three specific causal structures, while Chen & Du (2025) extend it through an exogenous isomorphism, an equivalence on the SCMs that preserves every counterfactual. With latent SCMs, counterfactuals have entered CRL mainly as data: paired observations before and after a perfect intervention (Brehmer et al., 2022) and actionable counterfactuals available for some variables only (Liu et al., 2023). There is only one recent work that addresses counterfactuals as an objective in the context of latent SCMs: Zhou et al. (2024) define counterfactual equivalence among observed domains of interventions and show that it needs no recovery of the latent SCM. This work, however, does not address counterfactual identifiability or counterfactuals beyond domains present in the data.

3 Counterfactual Identifiability in Latent SCMs

Latent structural causal models: To begin with, we formulate the data generating process as a latent SCM: *exogenous variables* $\boldsymbol{\varepsilon}$ drive the *endogenous variables* $\mathbf{z}$, the causal variables, through an SCM, and a decoder f renders $\mathbf{z}$ as the observation $\mathbf{x}$.

Definition 1 (Latent SCM)**.** *A **latent SCM** is a tuple $\mathcal{M} = \big(A,\ \{m_i^{\mathbf{v}}\}_{i\in[n],\,\mathbf{v}\in\mathcal{V}},\ \{p(\cdot \mid \mathbf{u})\}_{\mathbf{u}\in\mathcal{U}},\ f\big)$ generating the observation from exogenous variables through an additive-noise SCM and a decoder,*

$$z_i \;=\; m_i^{\mathbf{v}}\big(z_{\mathrm{pa}(i)}\big) + \varepsilon_i, \qquad \mathbf{x} = f(\mathbf{z}), \qquad \boldsymbol{\varepsilon} \mid \mathbf{u} \;\sim\; p(\boldsymbol{\varepsilon} \mid \mathbf{u}) = \prod\nolimits_{i=1}^{n} p(\varepsilon_i \mid \mathbf{u}), \tag{1}$$

*where $\boldsymbol{\varepsilon}, \mathbf{z} \in \mathbb{R}^n$ and $\mathbf{x} \in \mathbb{R}^d$ with $d \geq n$; A is a directed acyclic graph (DAG) on the index set $[n]$; the causal mechanism of each node, $m_i^{\mathbf{v}} : \mathbb{R}^{|\mathrm{pa}(i)|} \to \mathbb{R}$, is smooth, modulated by the **mechanism regime** $\mathbf{v}$, and depends on each parent in some regime (causal minimality; Peters et al., 2014); the exogenous variables are mutually independent given the **exogenous index** $\mathbf{u}$, so that within each value of $\mathbf{u}$ the SCM is Markovian (Pearl, 2009); and the decoder $f : \mathbb{R}^n \to \mathbb{R}^d$ is smooth, diffeomorphic onto its image, and shared by every mechanism regime. The spaces $\mathbb{R}^n$, $\mathbb{R}^d$ and the label sets $\mathcal{U}, \mathcal{V}$ are shared by all latent SCMs under comparison.*

The construction of $p(\boldsymbol{\varepsilon} \mid \mathbf{u})$ in Equation (1) is motivated by nonlinear ICA (Hyvärinen et al., 2019; Khemakhem et al., 2020). We write $g_A^{\mathbf{v}} : \boldsymbol{\varepsilon} \mapsto \mathbf{z}$ for the solution of Equation (1) in regime $\mathbf{v}$, a bijection of $\mathbb{R}^n$, and $F^{\mathbf{v}} := f \circ g_A^{\mathbf{v}}$ for the composite carrying $\boldsymbol{\varepsilon}$ to $\mathbf{x}$.

Definition 2 (Interventions)**.** *An intervention ι acts on a target set $I \subseteq [n]$: at each target $i \in I$, it replaces the original mechanism, $z_i = m_i^{\mathbf{v}}(\mathbf{z}_{\mathrm{pa}(i)}) + \varepsilon_i$ in Equation* (1)*, and leaves every other mechanism and the decoder untouched. Three forms are considered in this paper, with $c_i^\iota \in \mathbb{R}$ an **intervention value** and $\tilde{m}_i$ a continuous **intervention function** of the parents: a **perfect** intervention sets $z_i = c_i^\iota + \varepsilon_i$, a **do**-intervention sets $z_i = c_i^\iota$, and a **soft** intervention sets $z_i = \tilde{m}_i(\mathbf{z}_{\mathrm{pa}(i)}) + \varepsilon_i$.*

Two observationally equivalent latent SCMs may label and transform their causal variables differently, so an intervention on one may act on the other at another node and value; the correspondence below matches the two, as in causal abstraction (Rubenstein et al., 2017; Beckers & Halpern, 2019).

Definition 3 (Intervention correspondence)**.** *An **intervention correspondence** between two latent SCMs $\mathcal{M}$ and $\widehat{\mathcal{M}}$ is a bijection $\iota \mapsto \hat{\iota}$ between their interventions, specified by a permutation σ of $[n]$ and, at each node i, two continuous bijections $\rho_i, \tau_i : \mathbb{R} \to \mathbb{R}$. The corresponding intervention $\hat{\iota}$ acts at $\sigma(i)$ for each intervention target i of ι, as follows:*

$$\begin{aligned}
&\textbf{\textit{perfect} intervention:} && z_i = c + \varepsilon_i && \mapsto && \hat{z}_{\sigma(i)} = \rho_i(c) + \hat{\varepsilon}_{\sigma(i)},\\
&\textbf{\textit{do}-intervention:} && z_i = c && \mapsto && \hat{z}_{\sigma(i)} = \tau_i(c),\\
&\textbf{\textit{soft} intervention:} && z_i = \tilde{m}_i(\mathbf{z}_{\mathrm{pa}(i)}) + \varepsilon_i && \mapsto && \hat{z}_{\sigma(i)} = \rho_i\big(\tilde{m}_i(\mathbf{z}_{\mathrm{pa}(i)})\big) + \hat{\varepsilon}_{\sigma(i)},
\end{aligned} \tag{2}$$

where for the soft intervention, the parents' values are read from $\widehat{\mathcal{M}}$, $z_j = \tau_j^{-1}(\hat{z}_{\sigma(j)})$ for $j \in \mathrm{pa}(i)$; for this to be a soft intervention of $\widehat{\mathcal{M}}$, σ must carry the parents of i to the parents of $\sigma(i)$.

Counterfactual: For a latent SCM $\mathcal{M}$, the three steps of a counterfactual read as follows. *Abduction* recovers the unit's exogenous variables from a factual record $\mathbf{x}$ in regime $\mathbf{v}_0$, $\boldsymbol{\varepsilon} = (F^{\mathbf{v}_0})^{-1}(\mathbf{x})$; the

action is an intervention ι_k that turns $\mathcal{M}$ into the derived model $\mathcal{M}_{[\iota_k]}$ and $F^{\mathbf{v}_0}$ into $F^{\mathbf{v}_k}_{[\iota_k]}$; *prediction* runs the altered map on the recovered $\boldsymbol{\varepsilon}$, $\mathbf{x}^{\mathrm{CF}} = F^{\mathbf{v}_k}_{[\iota_k]}\big((F^{\mathbf{v}_0})^{-1}(\mathbf{x})\big)$.

Definition 4 (Counterfactual distribution and consistency)**.** *Let $\theta = \big(\mathbf{v}_0; (\mathbf{v}_k, \iota_k)_{k\le K}\big)$ collect a factual regime $\mathbf{v}_0 \in \mathcal{V}$ and finitely many regimes $\mathbf{v}_k \in \mathcal{V}$, each carrying an intervention ι_k. For a factual observation $\mathbf{x}$, write $\mathbf{x}^{\mathrm{CF}}_k(\mathbf{x}) := F^{\mathbf{v}_k}_{[\iota_k]}\big((F^{\mathbf{v}_0})^{-1}(\mathbf{x})\big)$ for its k-th counterfactual observation. The **counterfactual distribution** of $\mathcal{M}$ under θ is the joint distribution of $\mathbf{x}$ and its counterfactual observations, all generated from one shared draw of the exogenous variables:*

$$P^{\mathcal{M}}_{\theta}(\,\cdot \mid \mathbf{u}) \;:=\; \big(F^{\mathbf{v}_0},\, F^{\mathbf{v}_1}_{[\iota_1]},\, \ldots,\, F^{\mathbf{v}_K}_{[\iota_K]}\big)_{\#}\, p(\boldsymbol{\varepsilon} \mid \mathbf{u}). \tag{3}$$

*Latent SCMs $\mathcal{M}, \widehat{\mathcal{M}}$ are **counterfactually consistent**, denoted $\mathcal{M} \sim_{\mathrm{CF}} \widehat{\mathcal{M}}$, if there is an intervention correspondence $\iota \mapsto \hat{\iota}$ (Definition 3) such that*

$$P^{\mathcal{M}}_{\theta}(\,\cdot \mid \mathbf{u}) \;=\; P^{\widehat{\mathcal{M}}}_{\hat{\theta}}(\,\cdot \mid \mathbf{u}) \qquad \text{for every family } \theta \text{ and every } \mathbf{u}, \tag{4}$$

where $\hat{\theta} = \theta^{\sigma,\rho,\tau} := \big(\mathbf{v}_0; (\mathbf{v}_k, \hat{\iota}_k)_{k\le K}\big)$ replaces each ι_k by $\hat{\iota}_k$ and keeps the regimes.

Definition 5 (Counterfactual identifiability, after Chen & Du, 2025, Def. 2.3)**.** *Let $\mathcal{A}$ denote a set of assumptions, $[\mathcal{A}] = \{\mathcal{M} : \mathcal{M} \models \mathcal{A}\}$ the latent SCMs satisfying them, and call two latent SCMs **observationally equivalent** if they induce the same $p(\mathbf{x} \mid \mathbf{u}, \mathbf{v})$ for every $\mathbf{u}$ and $\mathbf{v}$. A latent SCM is **counterfactually identifiable** from $\mathcal{A}$ if $\mathcal{M} \sim_{\mathrm{CF}} \widehat{\mathcal{M}}$ for all observationally equivalent $\mathcal{M}, \widehat{\mathcal{M}} \in [\mathcal{A}]$.*

Causal effect: Following Pearl (2001), we define the natural direct and indirect effects for three causal variables: a treatment z_X, a mediator z_M, and an outcome z_Y, with $X, M, Y \in [n]$ (M may be a set of nodes) and $do(z_X{=}x)$ the do-intervention of Definition 2 that sets z_X to the value x.

Definition 6 (Natural direct and indirect effects)**.** *Let x, x' be two intervention values for the treatment z_X. On one shared draw $\boldsymbol{\varepsilon} \sim p(\,\cdot \mid \mathbf{u})$ in a fixed regime $\mathbf{v}$, write $Y_{x,m}$ for z_Y under $do(z_X{=}x,\, z_M{=}m)$ and M_x for z_M under $do(z_X{=}x)$; Y_{x',M_x} is then a **nested** counterfactual. The **natural direct** and **natural indirect effects** of the change $x \to x'$ on z_Y are, for finite expectations,*

$$\mathrm{NDE}_{x\to x'} := \mathbb{E}\big[Y_{x',M_x}\big] - \mathbb{E}\big[Y_{x,M_x}\big], \qquad \mathrm{NIE}_{x\to x'} := \mathbb{E}\big[Y_{x',M_{x'}}\big] - \mathbb{E}\big[Y_{x',M_x}\big]. \tag{5}$$

With the total effect $\mathrm{TE}_{x\to x'} := \mathbb{E}[Y_{x',M_{x'}}] - \mathbb{E}[Y_{x,M_x}]$, the two effects sum to the total, $\mathrm{TE}_{x\to x'} = \mathrm{NDE}_{x\to x'} + \mathrm{NIE}_{x\to x'}$. Here the indirect effect is the total indirect effect of Robins & Greenland (1992); the indirect effect of Pearl (2001) holds the treatment at x instead.

4 Latent SCM Isomorphism

Counterfactual identifiability (Definition 5) is not straightforward to establish as it is defined through the counterfactual distributions a latent SCM generates. Following Chen & Du (2025), we seek an alternative equivalence relation, stated on the latent SCMs, that implies counterfactual consistency.

The isomorphism: The equivalence relation, latent SCM isomorphism, is defined as follows.

Definition 7 (Latent SCM isomorphism)**.** *Latent SCMs $\mathcal{M}, \widehat{\mathcal{M}}$ are **isomorphic**, denoted $\mathcal{M} \sim_{\mathrm{LSI}} \widehat{\mathcal{M}}$, if there are a permutation σ of $[n]$, scales $s_i \neq 0$ and offsets b_i, c_i such that*

*(i) **(exogenous variables)** the exogenous variables correspond by $\hat{\boldsymbol{\varepsilon}} = \Phi_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon})$ with $\hat{\varepsilon}_{\sigma(i)} = s_i\,\varepsilon_i + b_i$, and $\Phi_{\boldsymbol{\varepsilon}}$ carries $p(\,\cdot \mid \mathbf{u})$ to $\hat{p}(\,\cdot \mid \mathbf{u})$ for every $\mathbf{u}$;*

*(ii) **(causal variables)** the causal variables correspond by $\hat{\mathbf{z}} = \Phi_{\mathbf{z}}(\mathbf{z})$ with $\hat{z}_{\sigma(i)} = s_i\, z_i + c_i$, and $\hat{g}^{\mathbf{v}}_{\widehat{A}} \circ \Phi_{\boldsymbol{\varepsilon}} = \Phi_{\mathbf{z}} \circ g^{\mathbf{v}}_{A}$ for every regime $\mathbf{v}$;*

*(iii) **(graph and decoder)** $j \in \mathrm{pa}(i)$ in A if and only if $\sigma(j) \in \widehat{\mathrm{pa}}(\sigma(i))$ in $\widehat{A}$, and $\hat{f} = f \circ \Phi_{\mathbf{z}}^{-1}$.*

From isomorphism to counterfactual identifiability: The isomorphism induces an intervention correspondence (Definition 3): its permutation σ carries parents to parents, and its value maps are affine with the shared scales. Under it, isomorphism implies counterfactual consistency.

Theorem 1 ($\sim_{\mathrm{LSI}}$ implies $\sim_{\mathrm{CF}}$; proof in Appendix B)**.** *Let $\mathcal{M} \sim_{\mathrm{LSI}} \widehat{\mathcal{M}}$ with the permutation σ, scales s_i and offsets b_i, c_i of Definition 7. Then $\mathcal{M} \sim_{\mathrm{CF}} \widehat{\mathcal{M}}$ under the intervention correspondence (Definition 3) specified by σ and the affine maps $\rho_i(t) = s_i\, t + c_i - b_i$ and $\tau_i(t) = s_i\, t + c_i$, $i \in [n]$.*

Definition 7 asks for element-wise affine maps while prior equivalences allow more: node-wise bijections of the exogenous variables in the exogenous isomorphism of Chen & Du (2025) on observed causal variables, and element-wise diffeomorphisms of the causal variables in causal representation learning (Brehmer et al., 2022; von Kügelgen et al., 2023). With *latent* causal variables, such maps no longer preserve every counterfactual. The affine class is tight, as the following lemma states.

Lemma 2 (Tightness of the affine class; proof in Appendix C)**.** *Let $\mathcal{M}$ be a latent SCM whose exogenous variables have a positive density on $\mathbb{R}^n$ for every $\mathbf{u}$, and let $\widehat{\mathcal{M}}$ be related to $\mathcal{M}$ as in Definition 7 with the affine maps replaced by element-wise diffeomorphisms $h_{\varepsilon,i}, h_{z,i} : \mathbb{R} \to \mathbb{R}$, $i \in [n]$. If for some node i the pair $(h_{\varepsilon,i}, h_{z,i})$ is not affine with a common slope, then $\mathcal{M} \not\sim_{\mathrm{CF}} \widehat{\mathcal{M}}$.*

From isomorphism to causal effects: Comparing causal effects on the *latent* causal variables across two latent SCMs requires consideration of the transformation established between the two SCMs. Under the permutation σ and the affine maps of Definition 7, two isomorphic latent SCMs agree on the effects of Definition 6 up to the scale of the outcome variable, as stated below.

Proposition 3 (Isomorphism preserves causal effects up to scale; proof in Appendix D)**.** *Let $\mathcal{M} \sim_{\mathrm{LSI}} \widehat{\mathcal{M}}$ with the permutation σ, scales s_i, and offsets c_i of Definition 7; let z_X, z_M, z_Y be causal variables of $\mathcal{M}$ and x, x' two intervention values for z_X, with $\hat{x} := \tau_X(x) = s_X x + c_X$ and $\hat{x}' := \tau_X(x') = s_X x' + c_X$ the corresponding values for $\hat{z}_{\sigma(X)}$. Then, for every $\mathbf{u}$ and $\mathbf{v}$, the effects of Definition 6 for $\hat{z}_{\sigma(X)}, \hat{z}_{\sigma(M)}, \hat{z}_{\sigma(Y)}$ in $\widehat{\mathcal{M}}$ equal s_Y times those for z_X, z_M, z_Y in $\mathcal{M}$, with s_Y the scale of Definition 7 at Y:*

$$\widehat{\mathrm{TE}}_{\hat{x}\to\hat{x}'} = s_Y\, \mathrm{TE}_{x\to x'}, \qquad \widehat{\mathrm{NDE}}_{\hat{x}\to\hat{x}'} = s_Y\, \mathrm{NDE}_{x\to x'}, \qquad \widehat{\mathrm{NIE}}_{\hat{x}\to\hat{x}'} = s_Y\, \mathrm{NIE}_{x\to x'}. \tag{6}$$

The three effects are thus identified up to the common factor s_Y, and their proportions exactly.

5 Reaching Latent SCM Isomorphism

This section establishes the assumptions under which any two observationally equivalent latent SCMs are isomorphic, hence counterfactually consistent by Theorem 1.

Assumption 1 (For identifying the exogenous variables)**.** *The exogenous index satisfies:*

(i) ***(Gaussian prior)*** *The prior is Gaussian, $p(\boldsymbol{\varepsilon} \mid \mathbf{u}) = \prod_{i=1}^{n} \mathcal{N}\big(\varepsilon_i; \mu_i(\mathbf{u}), \omega_i^2(\mathbf{u})\big)$, with natural parameters $\boldsymbol{\lambda}(\mathbf{u}) = \big(\mu_i(\mathbf{u})/\omega_i^2(\mathbf{u}),\ -1/(2\omega_i^2(\mathbf{u}))\big)_{i\le n} \in \mathbb{R}^{2n}$.*

(ii) ***(sufficient variability)*** *There exist $2n+1$ values $\mathbf{u}^0, \mathbf{u}^1, \ldots, \mathbf{u}^{2n}$ of the exogenous index, carried by every regime, such that the contrast matrix $\mathbf{L} = \big(\boldsymbol{\lambda}(\mathbf{u}^1) - \boldsymbol{\lambda}(\mathbf{u}^0), \ldots, \boldsymbol{\lambda}(\mathbf{u}^{2n}) - \boldsymbol{\lambda}(\mathbf{u}^0)\big) \in \mathbb{R}^{2n\times 2n}$ is invertible.*

Assumption 2 (For identifying the causal variables and the graph)**.** *A* ***reference regime*** *$\mathbf{v}_0$ and a set $\mathcal{D}$ of* ***designated regimes*** *are known and shared. $\mathcal{D}$ consists of one regime $\mathbf{v}_i$ per non-root node i ($\mathrm{pa}(i) \neq \emptyset$), obtained from $\mathbf{v}_0$ by a perfect intervention at i (Definition 2): the dependence of node i on its parents is removed and replaced by a constant, and every other mechanism is kept,*

$$m_i^{\mathbf{v}_i}\big(\mathbf{z}_{\mathrm{pa}(i)}\big) \;=\; c_i^{\mathbf{v}_i} \quad \textit{for all } \mathbf{z}_{\mathrm{pa}(i)}, \qquad m_j^{\mathbf{v}_i} \;=\; m_j^{\mathbf{v}_0} \quad \textit{for all } j \neq i, \tag{7}$$

with $m_i^{\mathbf{v}_i} \neq m_i^{\mathbf{v}_0}$. The designated regime $\mathbf{v}_i$ in the data is a label only: the target i is unknown.

Assumption 1 identifies the exogenous variables from the exogenous index; with the exogenous variables identified, Assumption 2 identifies the causal variables and the graph from the interventional data. Together the two force the isomorphism, formally stated in the following theorem.

Theorem 4 (Reaching latent SCM isomorphism; proof in Appendix E)**.** *Let $\mathcal{M}, \widehat{\mathcal{M}}$ be latent SCMs satisfying Assumptions 1 and 2. If $\mathcal{M}$ and $\widehat{\mathcal{M}}$ are observationally equivalent, then $\mathcal{M} \sim_{\mathrm{LSI}} \widehat{\mathcal{M}}$.*

Together with Theorem 1, Theorem 4 establishes that a latent SCM is counterfactually identifiable from Assumptions 1 and 2 (Definition 5). On the data side, the two assumptions ask for variation of the exogenous index, a reference regime, and one designated regime per non-root node, each realizing a perfect intervention at an unknown target. Under the intervention correspondence of Theorem 1, the identified latent SCM answers every counterfactual family of Definition 4 at any set of intervention targets and at intervention values and functions the data never realized, and the natural effects of Definition 6 up to the scale of the outcome variable (Proposition 3).

6 Learning the Latent SCM

We now translate Theorem 4 into a method, **FlowSCM**, for learning the latent SCM. Theorem 4 places two requirements on the learned latent SCM: Assumptions 1 and 2, and observational equivalence to the true latent SCM. FlowSCM meets both through its architecture and its objective.

Architecture: Neural networks with parameters ψ realize the latent SCM of Definition 1:

$$z_i \;=\; \gamma_i(\mathbf{v})\, m^{\mathbf{v}}_{i,\psi}\big(\mathbf{z}_{<i} \odot \mathbf{a}_i\big) + \varepsilon_i, \qquad \mathbf{x} = f_\psi(\mathbf{z}), \qquad \varepsilon_i \mid \mathbf{u} \;\sim\; \mathcal{N}\big(\mu_{i,\psi}(\mathbf{u}),\, \omega^2_{i,\psi}(\mathbf{u})\big), \tag{8}$$

where $m^{\mathbf{v}}_{i,\psi}$ is the mechanism of node i in regime $\mathbf{v}$, $\mathbf{a}_i \in [0,1]^{\,i-1}$ the i-th row of the learned graph weights A, strictly lower triangular in a fixed order (the flow may permute the coordinates, so the order carries no assumption), $\gamma_i(\mathbf{v}) \in [0,1]$ a per-regime gain, $\mu_{i,\psi}(\mathbf{u}), \omega_{i,\psi}(\mathbf{u})$ the prior's mean and scale at index value $\mathbf{u}$, and f_ψ an invertible flow. Because an invertible flow needs as many inputs as outputs, we design the decoder to take $d-n$ extra dimensions besides the causal variables as: $\mathbf{x} = f_\psi(\mathbf{z}, \boldsymbol{\eta})$ with $\boldsymbol{\eta} \in \mathbb{R}^{d-n}$, following GIN (Sorrenson et al., 2020). $\boldsymbol{\eta}$ carries no causal variable and follows a Gaussian prior, $\eta_k \sim \mathcal{N}(\nu_{k,\psi}, \kappa^2_{k,\psi})$, whose learned mean and scale do not depend on the exogenous index or the intervention regime. On images, FlowSCM is embedded in a convolutional autoencoder sharing the architecture with β-VAE and iVAE.

Assumption 1(i) is met by the model: the prior of the exogenous variables is conditionally independent and Gaussian, with a mean $\mu_{i,\psi}(\mathbf{u})$ and a scale $\omega_{i,\psi}(\mathbf{u})$ learned for every index value. Condition (ii) is an assumption on the data, the index varying sufficiently in every regime; the model keeps the separation of Definition 1 by letting the index enter only the prior and the regime only the mechanisms. Assumption 2 is met by the gain $\gamma_i(\mathbf{v})$: setting $\gamma_i(\mathbf{v}) = 0$ removes the parent term of node i, the perfect intervention of Equation (7), and every other gain is 1. Since the target of a designated regime is not known, the gains are estimated as part of the objective below.

Objective: Observational equivalence is approached by maximum likelihood. Inverting Equation (8) gives the exact per-sample log-likelihood:

$$\log p_\psi(\mathbf{x} \mid \mathbf{u}, \mathbf{v}) = \sum\nolimits_{j=1}^{n} \log \mathcal{N}\big(\varepsilon_j;\, \mu_{j,\psi}(\mathbf{u}), \omega^2_{j,\psi}(\mathbf{u})\big) + \log\big|\det J_{f_\psi^{-1}}(\mathbf{x})\big|, \tag{9}$$

where $\boldsymbol{\varepsilon}$ is obtained from $\mathbf{x}$ by inverting Equation (8) at the mechanisms and gains of regime $\mathbf{v}$, and the mechanism map is triangular with unit diagonal and contributes no Jacobian term. For $d > n$, the inversion also returns $\boldsymbol{\eta}$, and its log-prior, $\sum_{k=1}^{d-n} \log \mathcal{N}\big(\eta_k;\, \nu_{k,\psi}, \kappa^2_{k,\psi}\big)$, is added to Equation (9).

Denote by $\mathcal{D}$ the designated intervention regimes of Assumption 2. The target of a regime $\mathbf{v} \in \mathcal{D}$ is unknown. We thus treat it as a latent variable over the n nodes, shared by all samples of $\mathbf{v}$, and locate it by comparison with the reference regime $\mathbf{v}_0$. By Assumption 2, $\mathbf{v}$ differs from $\mathbf{v}_0$ at one mechanism only, so the model lets $\mathbf{v}$ share the mechanisms of $\mathbf{v}_0$ and sets the gain of a single node to zero. Denote by $D_{\mathbf{v}}$ the samples $(\mathbf{x}, \mathbf{u})$ of regime $\mathbf{v}$; the log-likelihood of $D_{\mathbf{v}}$ given that the target is node i is $\log p_\psi(D_{\mathbf{v}} \mid \mathbf{v}_0, i) = \sum_{(\mathbf{x},\mathbf{u}) \in D_{\mathbf{v}}} \log p_\psi(\mathbf{x} \mid \mathbf{u}, \mathbf{v}_0, i)$, where $\log p_\psi(\mathbf{x} \mid \mathbf{u}, \mathbf{v}_0, i)$ is Equation (9) evaluated with the mechanisms of $\mathbf{v}_0$, the gain of node i at zero and every other gain at one. Let $\alpha_{\mathbf{v}}$ be the variational distribution of the target: it weights the n candidate log-likelihoods, and the target is the node whose zero gain gives the highest likelihood on the data of $\mathbf{v}$. Training maximizes the following lower bound:

$$\mathcal{L}(\psi, \alpha) = \sum\nolimits_{\mathbf{v} \notin \mathcal{D}} \log p_\psi(D_{\mathbf{v}} \mid \mathbf{v}) + \sum\nolimits_{\mathbf{v} \in \mathcal{D}} \big[\sum\nolimits_{i=1}^{n} \alpha_{\mathbf{v},i} \log p_\psi(D_{\mathbf{v}} \mid \mathbf{v}_0, i) + H(\alpha_{\mathbf{v}})\big] + \lambda H(\bar{\alpha}), \tag{10}$$

where $\bar{\alpha}$ is the average of the $\alpha_{\mathbf{v}}$ over $\mathcal{D}$, H is the entropy, and $\lambda > 0$ is a weight; the last term keeps the targets of different regimes on different nodes. The derivation of the lower bound is given in Appendix F.

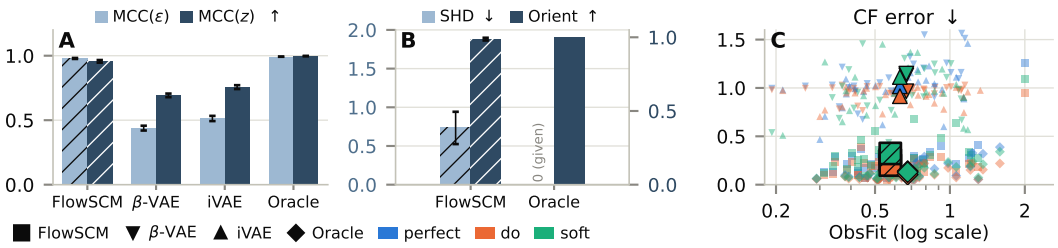


Figure 1: Results on synthetic data over 30 causal graphs. A–B: identification metrics, mean and standard error over the graphs. C: CF error of each model (shape) under the perfect, do, and soft interventions (colour) against its ObsFit; small points are single graphs, large markers their mean.

7 Experiments

In this section, we test the theory with FlowSCM on one synthetic and two rendered image datasets: whether the learned latent SCM reaches the isomorphism of Theorem 4, and whether, as Theorem 1 and Proposition 3 state, this returns the true counterfactuals and causal effects. We then ablate Assumptions 1 and 2 of Theorem 4, and compare FlowSCM with closely related CRL works.

Data: The synthetic data have $n = 5$ causal variables and $d = 10$-dimensional observations produced by a nonlinear mixing. The pendulum data (adapted from Yang et al., 2021) have $n = 4$ causal variables: the pendulum angle, the light angle, the shadow's position, and the shadow's length, rendered into 64×64 RGB images. Causal3DIdent (adapted from von Kügelgen et al., 2021; Zimmermann et al., 2021) has $n = 5$ causal variables: the object's two ground-plane coordinates, the spotlight's angle, and the hues of the object and of the spotlight, rendered into 64×64 RGB images. The generating processes, the latent SCMs, and the training are detailed in Appendices G.1–G.2.

Baselines: We consider two baselines without latent SCMs: β-VAE (Higgins et al., 2017) that receives neither the index nor the regimes, and iVAE (Khemakhem et al., 2020) that conditions its prior on the exogenous index. We further include an Oracle, FlowSCM with the true graph and mechanisms given, as a reference. Two additional CRL methods are considered: ILCM (Brehmer et al., 2022) that learns from counterfactual pairs, and SuaVE (Liu et al., 2026) that utilizes a *merged label* of the exogenous index and the intervention regime, under a linear latent SCM. Because each model has different identifiability assumptions, we compare FlowSCM with ILCM and SuaVE under separate data-generating conditions that meet each model's assumptions.

7.1 Identification, counterfactual identifiability and causal effects

We evaluate every model from three angles: identification up to the isomorphism of Definition 7 (Theorem 4), counterfactual identifiability (Theorem 1), and causal effects (Proposition 3), reported on the synthetic data in Figure 1 (with quantitative values in Appendix H.1) and on the two image datasets in Table 1. The metrics are defined below, with their computations in Appendix G.3.

Identification: Identification is scored by *MCC(ε)* and *MCC(z)* (Khemakhem et al., 2020), the mean correlation between the learned and true variables after a one-to-one matching of the coordinates. Because β-VAE and iVAE learn a single set of latent variables, they are matched once against the exogenous and once against the causal variables. The graph is scored by its structural Hamming distance (*SHD*) and orientation accuracy (*Orient*), read at the k strongest learned edges, k the number of true edges (Appendix G.3). Figure 1 and Table 1 show that FlowSCM recovers the exogenous variables, the causal variables, and the graph on all three datasets, close to the Oracle on the synthetic data and Causal3DIdent, while the two baselines recover neither the variables nor the graph.

Counterfactual identifiability: Counterfactual identifiability is scored by the *CF error* of the counterfactual predictions against the ground truth, and by *ObsFit*, the model's fit to factual observations:

$$\text{CF error} \;=\; \|\hat{\mathbf{x}}^{\text{CF}} - \mathbf{x}^{\text{CF}}\| \,/\, \|\mathbf{x}^{\text{CF}} - \mathbf{x}\|, \qquad \text{ObsFit} \;=\; D\big(\hat{p}(\mathbf{x} \mid \mathbf{u}, \mathbf{v}),\, p(\mathbf{x} \mid \mathbf{u}, \mathbf{v})\big), \tag{11}$$

where $\mathbf{x}$ is the factual observation, $\mathbf{x}^{\text{CF}}$ its true counterfactual, hats mark the learned model's outputs, and D is the Sinkhorn divergence on the synthetic data and the reconstruction error on the

Table 1: Results on Causal3DIdent and the pendulum: mean (standard error) over the four graphs of Causal3DIdent and three mechanism strengths of the pendulum. —: not returned by the method, or undefined on the dataset. Bold: the best value per column among FlowSCM and the two VAEs. †: the Oracle fails to identify the pendulum's coordinates (Table 4), possibly an optimization failure.

	Identification				Counterfactual				Effects			
Model	MCC($\boldsymbol{\varepsilon}$) ↑	MCC($\mathbf{z}$) ↑	SHD ↓	Orient ↑	ObsFit ↓	CF error ↓ perfect	CF error ↓ *do*	CF error ↓ soft	Δ_{TE} ↓	Δ_{NDE} ↓	Δ_{NIE} ↓	$\Delta_{\mathrm{NDE/TE}}$ ↓
Causal3DIdent												
FlowSCM	**0.951** (0.003)	**0.955** (0.012)	0.50 (0.50)	1.00 (0.00)	**0.0078** (0.0008)	**0.355** (0.033)	**0.323** (0.037)	**0.612** (0.132)	0.124 (0.031)	0.123 (0.078)	0.165 (0.047)	0.103 (0.060)
β-VAE	0.480 (0.005)	0.649 (0.017)	—	—	0.0131 (0.0007)	0.924 (0.052)	0.890 (0.040)	1.167 (0.161)	—	—	—	—
iVAE	0.578 (0.032)	0.754 (0.034)	—	—	0.0128 (0.0009)	0.778 (0.038)	0.751 (0.040)	1.043 (0.138)	—	—	—	—
Oracle	0.949 (0.007)	0.983 (0.001)	*given*	*given*	0.0078 (0.0006)	0.366 (0.049)	0.339 (0.049)	0.607 (0.141)	0.050 (0.008)	0.004 (0.003)	0.052 (0.008)	0.004 (0.003)
Pendulum												
FlowSCM	**0.954** (0.020)	**0.993** (0.001)	0.00 (0.00)	1.00 (0.00)	**0.0085** (0.0003)	**0.552** (0.017)	**0.521** (0.010)	—	0.071 (0.017)	0.130 (0.027)	0.085 (0.013)	0.074 (0.002)
β-VAE	0.659 (0.018)	0.827 (0.027)	—	—	0.0109 (0.0003)	0.751 (0.034)	0.702 (0.044)	—	—	—	—	—
iVAE	0.663 (0.030)	0.806 (0.010)	—	—	0.0109 (0.0002)	0.807 (0.084)	0.762 (0.106)	—	—	—	—	—
Oracle†	0.730 (0.092)	0.731 (0.108)	*given*	*given*	0.0087 (0.0001)	1.085 (0.040)	1.111 (0.080)	—	0.993 (0.007)	0.739 (0.102)	0.881 (0.055)	—

images (Appendix G.3). Note that the differences in ObsFit on the factual data are small across all models, on all three datasets. In comparison, FlowSCM delivers a much smaller CF error than the two baselines and approaches that of the Oracle, under every intervention form. This provides the empirical evidence for Theorem 1: recovery up to the isomorphism carries the counterfactuals, while the observational equivalence alone does not. Appendix H.6 shows visual examples of factual and counterfactual images generated by all models.

Causal effects: We report the error of predicted causal effects following Proposition 3 as:

$$\Delta_{\mathrm{TE}} = |s_Y\,\mathrm{TE} - \widehat{\mathrm{TE}}|\,/\,|s_Y\,\mathrm{TE}|, \qquad \Delta_{\mathrm{NDE/TE}} = \left|\mathrm{NDE/TE} - \widehat{\mathrm{NDE}}/\widehat{\mathrm{TE}}\right|, \tag{12}$$

and Δ_{NDE}, Δ_{NIE} likewise over the same denominator, where hats mark the learned latent SCM's effects and s_Y is the scale of Definition 7 at the outcome Y, estimated from the coordinate matching (Appendix G.3). Note that since NDE/TE and NIE/TE sum to one, $\Delta_{\mathrm{NIE/TE}} = \Delta_{\mathrm{NDE/TE}}$. Because the two VAEs have no latent SCM, they cannot be used for effect estimation. On the two image datasets, FlowSCM's effect errors are 7% to 17% of the true total effect. The Oracle's errors are smaller on Causal3DIdent, whereas on the pendulum it fails to identify the coordinates (†; Appendix G.4) and thus delivers poor estimates for counterfactuals and effects.

7.2 Ablations

Figure 2A–B show ablation results on Theorem 4 on the synthetic data, with FlowSCM as the full set; Appendix H.2 gives the complete tables. *Index only* drops the designated intervention regimes of Assumption 2: as a result, while the exogenous variables are still identified, neither the causal variables nor the graph is; the estimated counterfactuals and causal effects degrade accordingly. *Interventions only* drops the exogenous index of Assumption 1: as a result, the exogenous variables are not identified, nor the subsequent endogenous variables, counterfactuals, or causal effects.

Figure 2C shows the results as we vary the number of exogenous index values M_u, testing the sufficient variability of Assumption 1(ii): nearly every metric improves with M_u and keeps improving past the $2n+1$ values the assumption asks for, showing that more variability than the theory requires still helps in practice. Figure 2D shows the effect of the number k of single-node interventions starting from none: nearly every metric improves with k and saturates when every non-root node has an intervention ($k = 4$), the number Assumption 2 asks for. Results in Appendix H.4 illustrate Lemma 2: a non-affine map of a causal variable keeps the observational fit but degrades the counterfactuals. Results in Appendix H.5 show that while Assumption 1(i) stipulates Gaussian exogenous variables, FlowSCM still performs well on non-Gaussian exogenous variables in practice.

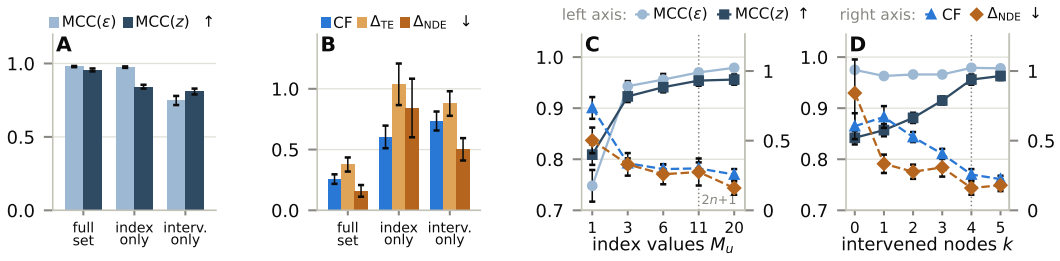


Figure 2: Ablations on the synthetic data; mean and standard error over the graphs. A–B: the full set against *index only* and *interventions only*. C: effect of the number of index values M_u; the dotted line marks $2n + 1 = 11$. D: effect of the number of nodes k with a designated regime, $k = 4$ every non-root, $k = 5$ the root added. CF error is the perfect form; Δ_{TE} and Δ_{NDE} as in Table 1.

Table 2: Comparison of FlowSCM with ILCM and SuaVE on data generators that meet their respective assumptions, over 30 datasets; DCI (Eastwood & Williams, 2018) and the ENCO SHD (Lippe et al., 2022a) are the scores ILCM reports. Bold: best per metric and M_u in (a), per column in (b).

(a) ILCM's generator

	$M_u = 1$		$M_u = 20$		$M_u = 100$	
	ILCM	FlowSCM	ILCM	FlowSCM	ILCM	FlowSCM
DCI ↑	**0.965** (0.014)	0.245 (0.012)	0.815 (0.019)	**0.893** (0.015)	0.916 (0.016)	**0.925** (0.014)
ENCO SHD ↓	**0.500** (0.171)	4.733 (0.225)	3.633 (0.227)	**1.900** (0.251)	1.667 (0.241)	**1.167** (0.259)
MCC($\boldsymbol{\varepsilon}$) ↑	**0.928** (0.011)	0.676 (0.010)	0.918 (0.013)	**0.993** (0.001)	0.920 (0.013)	**0.996** (0.001)
MCC($\mathbf{z}$) ↑	**0.973** (0.012)	0.657 (0.008)	0.920 (0.014)	**0.974** (0.005)	0.950 (0.015)	**0.981** (0.004)

(b) SuaVE's generator, $M_u = 20$

	variables		graph	
	MCC($\boldsymbol{\varepsilon}$) ↑	MCC($\mathbf{z}$) ↑	SHD ↓	Orient ↑
FlowSCM	**0.958** (0.003)	**0.932** (0.009)	**1.633** (0.290)	**0.992** (0.008)
SuaVE	0.510 (0.021)	0.902 (0.011)	5.133 (0.261)	0.694 (0.060)
β-VAE	0.380 (0.018)	0.629 (0.012)	—	—
iVAE	0.529 (0.021)	0.793 (0.016)	—	—

7.3 Comparisons with related CRL methods

Finally, Table 2 compares FlowSCM with ILCM and SuaVE on separate data generation processes that satisfy each model's assumptions. Block (a) reports results on data from ILCM's generator with M_u index values, with $M_u = 1$ representing ILCM's setting of a single Gaussian and $M_u > 2n+1$ representing FlowSCM's setting. At $M_u = 1$, ILCM returns its published performance, while FlowSCM does not identify due to the violation of sufficient variability (Assumption 1(ii)). At $M_u = 20$, ILCM's performance drops while FlowSCM leads on all metrics. At $M_u = 100$, ILCM regains its performance but FlowSCM still outperforms on all metrics. Note that this performance gain of FlowSCM is obtained without requiring the pairing of samples before and after an intervention, while ILCM does. Block (b) reports results on data from SuaVE's generator with linear causal mechanisms, which can be considered a special case of FlowSCM's setting. SuaVE recovers the causal variables but neither the exogenous variables nor the orientation of the graph, whereas FlowSCM recovers all three. The reproductions of ILCM and SuaVE in their original settings are in Appendix G.4, and their CF errors (higher than FlowSCM's) and causal effects in Appendix H.3.

8 Conclusion

In this work, we proved that latent SCM isomorphism (representation identification) implies consistency on every counterfactual (counterfactual identification) and supports the analysis of causal effects. We then developed FlowSCM, a flow-based estimator built on the identifiability assumptions, and tested the theory on synthetic and image data. The key **limitations** of our work lie in the assumptions (additive-noise mechanisms, a Gaussian prior on the exogenous variables modulated by an index, and one perfect intervention per non-root node) and in the asymptotic identifiability guarantees that do not address whether or how finite-sample optimization can reach the theoretical solution. Weakening the assumptions (for instance to soft interventions or non-Gaussian exogenous variables), establishing finite-sample and convergence guarantees, and applying the framework to real-world data are left for future work.

AI USE STATEMENT

The authors conceived the research questions and the framework, and decided every definition, result and experiment. Large language model assistants (Anthropic's Claude) supported the work under the authors' direction by retrieving and organizing literature; drafting and editing text and parts of the appendix proofs from the authors' outlines; checking the proofs; giving feedback on drafts, on the design of the experiments, and on the reading of their results; writing and debugging parts of the experimental code; and compiling logged results into tables and figures. The authors checked every proof, tested the code, and verified every reported number against the logged outputs. All data come from the simulators and renderers described in Appendix G; none were generated by a language model. We have not used generative AI tools for other tasks with required disclosures, and the rest of the required disclosure tasks are not applicable to this work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The proofs of Theorems 1 and 4, Lemma 2 and Proposition 3 are given in full in Appendices B–E, and the objective FlowSCM optimizes is derived in Appendix F. The data generating processes of the three datasets, the training, seed-selection and baseline protocols, and the computation of every metric are documented in Appendix G. The code, including the generators of the three datasets, will be made publicly available upon acceptance.

REFERENCES

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning (ICML)*, pp. 372–407, 2023.

Sander Beckers and Joseph Y. Halpern. Abstracting causal models. In *AAAI Conference on Artificial Intelligence*, volume 33, pp. 2678–2685, 2019.

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yikang Chen and Dehui Du. Exogenous isomorphism for counterfactual identifiability. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 8022–8064, 2025.

Juan Correa, Sanghack Lee, and Elias Bareinboim. Nested counterfactual identification from arbitrary surrogate experiments. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6856–6867, 2021.

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations (ICLR)*, 2018.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017.

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

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 859–868, 2019.

Adrián Javaloy, Pablo Sánchez-Martín, and Isabel Valera. Causal normalizing flows: from theory to practice. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Ilyes Khemakhem, Ricardo Pio Monti, Robert Leech, and Aapo Hyvärinen. Causal autoregressive flows. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 3520–3528, 2021.

Aneesh Komanduri, Yongkai Wu, Feng Chen, and Xintao Wu. Learning causally disentangled representations via the principle of independent causal mechanisms (ICM-VAE). In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

Phillip Lippe, Taco Cohen, and Efstratios Gavves. Efficient neural causal discovery without acyclicity constraints. In *International Conference on Learning Representations (ICLR)*, 2022a.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning (ICML)*, 2022b.

Yuejiang Liu, Alexandre Alahi, Chris Russell, Max Horn, Dominik Zietlow, Bernhard Schölkopf, and Francesco Locatello. Causal Triplet: An open challenge for intervention-centric causal representation learning. In *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213, pp. 553–573. PMLR, 2023.

Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifiable latent polynomial causal models through the lens of change. In *International Conference on Learning Representations (ICLR)*, 2024.

Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifying weight-variant latent causal models. *Journal of Machine Learning Research*, 27(23-1023), 2026. arXiv:2208.14153, 2022.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning (ICML)*, 2019.

Arash Nasr-Esfahany and Emre Kiciman. Counterfactual (non-)identifiability of learned structural causal models. *arXiv preprint arXiv:2301.09031*, 2023.

Arash Nasr-Esfahany, Mohammad Alizadeh, and Devavrat Shah. Counterfactual identifiability of bijective causal models. In *International Conference on Machine Learning (ICML)*, pp. 25733–25754, 2023.

Ignavier Ng, Shaoan Xie, Xinshuai Dong, Peter Spirtes, and Kun Zhang. Causal representation learning from general environments under nonparametric mixing. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 411–420, 2001.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

James M. Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

Paul K. Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M. Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

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.

Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ICA with general incompressible-flow networks (GIN). In *International Conference on Learning Representations (ICLR)*, 2020.

Chandler Squires, Anna Seigal, Salil S. Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *International Conference on Machine Learning (ICML)*, pp. 32540–32560, 2023.

Burak Varıcı, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

Burak Varıcı, Emre Acartürk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning: Linear and general transformations. *Journal of Machine Learning Research*, 2025.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. CausalVAE: Disentangled representation learning via neural structural causal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

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. In *International Conference on Learning Representations (ICLR)*, 2024.

Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. In *International Conference on Machine Learning (ICML)*, 2024.

Zeyu Zhou, Ruqi Bai, Sean Kulinski, Murat Kocaoglu, and David I. Inouye. Towards characterizing domain counterfactuals for invertible latent causal models. In *International Conference on Learning Representations (ICLR)*, 2024.

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12979–12990, 2021.

APPENDIX CONTENTS

A Notation **13**

B Proof of Theorem 1: counterfactual consistency under the isomorphism **14**

C Proof of Lemma 2: tightness of the affine class **15**

D Proof of Proposition 3: causal effects under the isomorphism **17**

E Proof of Theorem 4: reaching the isomorphism **17**

E.1 Identifying the exogenous variables 18

E.2 Identifying the causal variables and the graph 21

E.3 Where the argument needs the index to stay out of the mechanisms 22

F Derivation of the objective **23**

G Experimental setup **24**

G.1 Data generating processes 24

G.2 Training and cost 25

G.3 Metrics 26

G.4 Protocols of Tables 1 and 2 27

H Additional results **28**

H.1 Figure 1 in numbers 28

H.2 Ablation tables 28

H.3 Counterfactuals on ILCM's and SuaVE's generators 29

H.4 Counterfactuals under a non-affine reparametrization 30

H.5 Exogenous variables outside the Gaussian family 31

H.6 Factual and counterfactual image samples 32

A NOTATION

Throughout the appendix $\mathcal{M}, \widehat{\mathcal{M}}$ are latent SCMs (Definition 1), and hatted objects belong to $\widehat{\mathcal{M}}$. Under $\mathcal{M} \sim_{\mathrm{LSI}} \widehat{\mathcal{M}}$ (Definition 7) the element-wise affine maps are $\Phi_{\boldsymbol{\varepsilon}}, \Phi_{\mathbf{z}}$, with $\hat{\varepsilon}_{\sigma(i)} = s_i \varepsilon_i + b_i$ and $\hat{z}_{\sigma(i)} = s_i z_i + c_i$, and the induced value maps on the two scales of an intervention value are

$$\rho_i(t) \;=\; s_i t + c_i - b_i \quad \text{(mechanism-value scale)}, \qquad \tau_i(t) \;=\; s_i t + c_i \quad \text{(variable-value scale)}. \tag{13}$$

Solution maps: $g_A^{\mathbf{v}} : \boldsymbol{\varepsilon} \mapsto \mathbf{z}$ and $F^{\mathbf{v}} = f \circ g_A^{\mathbf{v}}$ (Definition 1); $g_{[\iota]}^{\mathbf{v}}$ and $F_{[\iota]}^{\mathbf{v}}$ are their analogues in the derived model $\mathcal{M}_{[\iota]}$ (Definition 2), well defined because an intervention only removes or keeps dependencies within the original parent sets, so the derived graph stays acyclic. Since $g_A^{\mathbf{v}}$ is a bijection of $\mathbb{R}^n$, the observed law of every regime is carried by one image, $F^{\mathbf{v}}(\mathbb{R}^n) = f(\mathbb{R}^n)$, whose closure is its support when the exogenous variables have a positive density. Two observationally equivalent models carry the same law, so their images $f(\mathbb{R}^n)$ and $\hat{f}(\mathbb{R}^n)$ agree up to a null set; the *observed support* below is their common part, which carries the whole law. The two images coincide when the models are related by maps of $\mathbb{R}^n$, as in Definition 7 and Lemma 2, and, under Assumption 1,

by Lemma A4. Where a result assumes it (Lemmas 2 and A2; automatic under Assumption 1), the exogenous variables have a positive density on $\mathbb{R}^n$, so every nonempty relatively open part of the observed support has positive probability and a full-measure subset of it is dense in it. All solution maps, the abductions $(F^{\mathbf{v}})^{-1}$ on $f(\mathbb{R}^n)$, and the answer maps $\mathbf{x}_k^{\mathrm{CF}}$ of Definition 4 are continuous, as compositions of the smooth mechanisms, the continuous intervention functions of Definition 2, the decoder, and their inverses; under a correspondence of Definition 3 the corresponding soft intervention $\rho_i \circ \tilde{m}_i \circ \tau_{\mathrm{pa}(i)}^{-1}$ is continuous as well, so the same holds for the answer maps of $\widehat{\mathcal{M}}$.

B Proof of Theorem 1: counterfactual consistency under the isomorphism

Theorem 1 ($\sim_{\mathrm{LSI}}$ implies $\sim_{\mathrm{CF}}$)**.** *Let $\mathcal{M} \sim_{\mathrm{LSI}} \widehat{\mathcal{M}}$ with the permutation σ, scales s_i and offsets b_i, c_i of Definition 7. Then $\mathcal{M} \sim_{\mathrm{CF}} \widehat{\mathcal{M}}$ under the intervention correspondence (Definition 3) specified by σ and the affine maps $\rho_i(t) = s_i\, t + c_i - b_i$ and $\tau_i(t) = s_i\, t + c_i$, $i \in [n]$.*

Lemma A1 (Transport of interventions)**.** *Let $\mathcal{M} \sim_{\mathrm{LSI}} \widehat{\mathcal{M}}$, and let ι be an intervention of Definition 2 with target set I. Let $\hat{\iota}$ act at $\sigma(I)$, translating a perfect value by ρ_i, a do-value by τ_i, and a soft mechanism to $\rho_i \circ \tilde{m}_i \circ \tau_{\mathrm{pa}(i)}^{-1}$, as in Equation* (13)*. Then $\widehat{F}^{\mathbf{v}}_{[\hat{\iota}]} \circ \Phi_{\boldsymbol{\varepsilon}} = F^{\mathbf{v}}_{[\iota]}$ for every regime $\mathbf{v}$.*

Proof. Write (σ, s_i, b_i, c_i) and $\Phi_{\boldsymbol{\varepsilon}}, \Phi_{\mathbf{z}}$ for the data of Definition 7, so that $\hat{\varepsilon}_{\sigma(i)} = s_i\varepsilon_i + b_i$, $\hat{z}_{\sigma(i)} = s_i z_i + c_i$, and $\widehat{\mathrm{pa}}(\sigma(i)) = \sigma(\mathrm{pa}(i))$ by the graph clause.

Step 0 (the mechanism identity). Clause (ii) of Definition 7 states $\hat{g}_A^{\mathbf{v}} \circ \Phi_{\boldsymbol{\varepsilon}} = \Phi_{\mathbf{z}} \circ g_A^{\mathbf{v}}$ for every regime. Fix $\boldsymbol{\varepsilon}$, put $\mathbf{z} := g_A^{\mathbf{v}}(\boldsymbol{\varepsilon})$ and $\hat{\mathbf{z}} := \Phi_{\mathbf{z}}(\mathbf{z})$; reading the hatted SCM at node $\sigma(i)$ and substituting the two element-wise maps,

$$s_i z_i + c_i \;=\; \hat{z}_{\sigma(i)} \;=\; \hat{m}^{\mathbf{v}}_{\sigma(i)}\big(\hat{\mathbf{z}}_{\sigma(\mathrm{pa}(i))}\big) + \hat{\varepsilon}_{\sigma(i)} \;=\; \hat{m}^{\mathbf{v}}_{\sigma(i)}\big(\hat{\mathbf{z}}_{\sigma(\mathrm{pa}(i))}\big) + s_i\varepsilon_i + b_i, \tag{14}$$

and cancelling $z_i = m_i^{\mathbf{v}}(\mathbf{z}_{\mathrm{pa}(i)}) + \varepsilon_i$ against the exogenous-variable term leaves

$$\hat{m}^{\mathbf{v}}_{\sigma(i)}\big(\Phi_{\mathbf{z}}(\mathbf{z})_{\sigma(\mathrm{pa}(i))}\big) \;=\; s_i\, m_i^{\mathbf{v}}(\mathbf{z}_{\mathrm{pa}(i)}) + c_i - b_i, \tag{15}$$

an identity for every $\mathbf{z} \in \mathbb{R}^n$, since $g_A^{\mathbf{v}}$ is onto.

Step 1 (the derived solutions correspond). Fix $\boldsymbol{\varepsilon}$ and let $\mathbf{z} := g^{\mathbf{v}}_{[\iota]}(\boldsymbol{\varepsilon})$ solve the derived model of $\mathcal{M}$. We claim $\Phi_{\mathbf{z}}(\mathbf{z})$ solves the derived model of $\widehat{\mathcal{M}}$ under $\hat{\iota}$ with exogenous variables $\Phi_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon})$; the derived solution is unique by acyclicity, so the claim proves $\hat{g}^{\mathbf{v}}_{[\hat{\iota}]} \circ \Phi_{\boldsymbol{\varepsilon}} = \Phi_{\mathbf{z}} \circ g^{\mathbf{v}}_{[\iota]}$, and it is checked node by node. At a node $\sigma(k)$ with $k \notin I$ the mechanism is the original $\hat{m}^{\mathbf{v}}_{\sigma(k)}$, and Equation (15) gives

$$\begin{aligned}\hat{m}^{\mathbf{v}}_{\sigma(k)}\big(\Phi_{\mathbf{z}}(\mathbf{z})_{\sigma(\mathrm{pa}(k))}\big) + \hat{\varepsilon}_{\sigma(k)} \;&=\; \big(s_k m_k^{\mathbf{v}}(\mathbf{z}_{\mathrm{pa}(k)}) + c_k - b_k\big) + \big(s_k\varepsilon_k + b_k\big) \\ &=\; s_k z_k + c_k \;=\; \Phi_{\mathbf{z}}(\mathbf{z})_{\sigma(k)}.\end{aligned} \tag{16}$$

At a target $i \in I$ the three forms of Definition 2 are checked separately, with z_i the value the derived model of $\mathcal{M}$ assigns. If ι is perfect with value c_i^ι, then $z_i = c_i^\iota + \varepsilon_i$ and $\hat{\iota}$ has intervention value $\rho_i(c_i^\iota)$, so

$$\rho_i(c_i^\iota) + \hat{\varepsilon}_{\sigma(i)} \;=\; \big(s_i c_i^\iota + c_i - b_i\big) + \big(s_i\varepsilon_i + b_i\big) \;=\; s_i\big(c_i^\iota + \varepsilon_i\big) + c_i \;=\; s_i z_i + c_i. \tag{17}$$

If ι is a *do*-intervention with value c_i^ι, then $z_i = c_i^\iota$, no exogenous variable is retained on either side, and $\hat{\iota}$ has intervention value

$$\tau_i(c_i^\iota) \;=\; s_i c_i^\iota + c_i \;=\; s_i z_i + c_i; \tag{18}$$

τ_i carries no $-b_i$ because there is no exogenous-variable term to compensate. If ι is soft with mechanism $\tilde{m}_i$, then $z_i = \tilde{m}_i(\mathbf{z}_{\mathrm{pa}(i)}) + \varepsilon_i$, the hatted mechanism is $\rho_i \circ \tilde{m}_i \circ \tau_{\mathrm{pa}(i)}^{-1}$, and $\Phi_{\mathbf{z}}(\mathbf{z})_{\sigma(j)} = s_j z_j + c_j = \tau_j(z_j)$, so $\tau_{\mathrm{pa}(i)}^{-1}$ returns the hatted parents to $\mathcal{M}$'s variable scale; hence

$$\rho_i\big(\tilde{m}_i(\mathbf{z}_{\mathrm{pa}(i)})\big) + \hat{\varepsilon}_{\sigma(i)} \;=\; \big(s_i\tilde{m}_i(\mathbf{z}_{\mathrm{pa}(i)}) + c_i - b_i\big) + \big(s_i\varepsilon_i + b_i\big) \;=\; s_i z_i + c_i. \tag{19}$$

In each case node $\sigma(i)$'s equation holds at $\Phi_{\mathbf{z}}(\mathbf{z})_{\sigma(i)} = s_i z_i + c_i$; the $-b_i$ of ρ_i appears exactly in the two forms that retain the exogenous variable.

Step 2 (compose with the decoder). By the decoder clause $\hat{f} = f \circ \Phi_{\mathbf{z}}^{-1}$, for every regime $\mathbf{v}$

$$\widehat{F}^{\mathbf{v}}_{[\hat{\iota}]} \circ \Phi_{\boldsymbol{\varepsilon}} \;=\; \hat{f} \circ \hat{g}^{\mathbf{v}}_{[\hat{\iota}]} \circ \Phi_{\boldsymbol{\varepsilon}} \;=\; f \circ \Phi_{\mathbf{z}}^{-1} \circ \Phi_{\mathbf{z}} \circ g^{\mathbf{v}}_{[\iota]} \;=\; F^{\mathbf{v}}_{[\iota]}. \tag{20}$$

The empty target set is the case $\iota = \hat{\iota} = \varnothing$ of Equation (20): $\widehat{F}^{\mathbf{v}} \circ \Phi_{\boldsymbol{\varepsilon}} = F^{\mathbf{v}}$. □

Proof of Theorem 1. Let (σ, s_i, b_i, c_i) realize $\mathcal{M} \sim_{\mathrm{LSI}} \widehat{\mathcal{M}}$, and take as the candidate correspondence the permutation σ together with the pairs of Equation (13), $\rho_i(t) = s_i t + c_i - b_i$ and $\tau_i(t) = s_i t + c_i$, which are continuous bijections of $\mathbb{R}$, since $s_i \neq 0$, as Definition 3 requires. Fix a family $\theta = (\mathbf{v}_0; (\mathbf{v}_k, \iota_k)_{k \leq K})$ and a value of $\mathbf{u}$, and write $\hat{\iota}_k$ for the image of ι_k under the correspondence, as in Lemma A1. Then

$$\begin{aligned} P^{\widehat{\mathcal{M}}}_{\theta^{\sigma,\rho,\tau}}(\,\cdot \mid \mathbf{u}) &= \big(\widehat{F}^{\mathbf{v}_0},\, \widehat{F}^{\mathbf{v}_1}_{[\hat{\iota}_1]},\, \ldots,\, \widehat{F}^{\mathbf{v}_K}_{[\hat{\iota}_K]}\big)_{\#}\, \hat{p}(\,\cdot \mid \mathbf{u}) \\ &= \big(\widehat{F}^{\mathbf{v}_0} \circ \Phi_{\boldsymbol{\varepsilon}},\, \widehat{F}^{\mathbf{v}_1}_{[\hat{\iota}_1]} \circ \Phi_{\boldsymbol{\varepsilon}},\, \ldots,\, \widehat{F}^{\mathbf{v}_K}_{[\hat{\iota}_K]} \circ \Phi_{\boldsymbol{\varepsilon}}\big)_{\#}\, p(\,\cdot \mid \mathbf{u}) \\ &= \big(F^{\mathbf{v}_0},\, F^{\mathbf{v}_1}_{[\iota_1]},\, \ldots,\, F^{\mathbf{v}_K}_{[\iota_K]}\big)_{\#}\, p(\,\cdot \mid \mathbf{u}) \;=\; P^{\mathcal{M}}_{\theta}(\,\cdot \mid \mathbf{u}). \end{aligned} \tag{21}$$

The first equality is Equation (3) in $\widehat{\mathcal{M}}$; the second is the exogenous-variable clause of Definition 7, $\hat{p}(\,\cdot \mid \mathbf{u}) = \Phi_{\boldsymbol{\varepsilon}\#}\, p(\,\cdot \mid \mathbf{u})$, pushed through the tuple; the third is Lemma A1 applied coordinate-wise (to the coordinate $k = 0$ as the empty intervention), and the last is Equation (3) in $\mathcal{M}$. The second equality is where the counterfactual structure enters: all $K{+}1$ coordinates of Equation (3) are functions of the *one* draw, so the single change of variables $\hat{\boldsymbol{\varepsilon}} = \Phi_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon})$ serves them simultaneously. As θ and $\mathbf{u}$ were arbitrary, Equation (4) holds and $\mathcal{M} \sim_{\mathrm{CF}} \widehat{\mathcal{M}}$. □

C Proof of Lemma 2: tightness of the affine class

Lemma 2 (Tightness of the affine class)**.** *Let $\mathcal{M}$ be a latent SCM whose exogenous variables have a positive density on $\mathbb{R}^n$ for every $\mathbf{u}$, and let $\widehat{\mathcal{M}}$ be related to $\mathcal{M}$ as in Definition 7 with the affine maps replaced by element-wise diffeomorphisms $h_{\varepsilon,i}, h_{z,i} : \mathbb{R} \to \mathbb{R}$, $i \in [n]$. If for some node i the pair $(h_{\varepsilon,i}, h_{z,i})$ is not affine with a common slope, then $\mathcal{M} \not\sim_{\mathrm{CF}} \widehat{\mathcal{M}}$.*

Two auxiliary lemmas carry the proof. The first reads the distribution-level consistency of Equation (4) one record at a time. The second is the functional-equation rigidity that forces the affine form.

Lemma A2 (Individual-level reading)**.** *Let the exogenous variables have a positive density on $\mathbb{R}^n$. Equality of the joints in Equation* (4) *for a family θ is equivalent to: the factual laws $p(\mathbf{x} \mid \mathbf{u}, \mathbf{v}_0)$ coincide, and the abduction–action–prediction outputs $\mathbf{x}_k^{\mathrm{CF}}$ of the two latent SCMs, the interventions matched by the correspondence, coincide at every factual record in the observed support.*

Proof. Write $G(\mathbf{x}) := \big(\mathbf{x}, \mathbf{x}_1^{\mathrm{CF}}(\mathbf{x}), \ldots, \mathbf{x}_K^{\mathrm{CF}}(\mathbf{x})\big)$ for $\mathcal{M}$'s answer map under θ and $\widehat{G}$ for $\widehat{\mathcal{M}}$'s under $\theta^{\sigma,\rho,\tau}$, and $\mu := p(\mathbf{x} \mid \mathbf{u}, \mathbf{v}_0)$, $\hat{\mu} := \hat{p}(\mathbf{x} \mid \mathbf{u}, \mathbf{v}_0)$ for the factual laws, so that, substituting $\mathbf{x} = F^{\mathbf{v}_0}(\boldsymbol{\varepsilon})$ in Equation (3), $P^{\mathcal{M}}_{\theta} = G_{\#}\mu$ and $P^{\widehat{\mathcal{M}}}_{\theta^{\sigma,\rho,\tau}} = \widehat{G}_{\#}\hat{\mu}$.

If $\mu = \hat{\mu}$ and $G = \widehat{G}$ on the observed support, the two pushforwards are equal; this is the easy direction. Conversely, suppose $G_{\#}\mu = \widehat{G}_{\#}\hat{\mu} =: P$. Projecting P to its first coordinate gives μ from the left representation and $\hat{\mu}$ from the right, so $\mu = \hat{\mu}$. Next, P charges the graphs of both answer maps: the sets

$$E := \big\{(\mathbf{x}, \mathbf{y}_1, \ldots, \mathbf{y}_K) : \mathbf{y}_k = \mathbf{x}_k^{\mathrm{CF}}(\mathbf{x}) \text{ for all } k\big\}, \qquad \widehat{E} \text{ its hatted analogue}, \tag{22}$$

are measurable, being relatively closed by the continuity of the answer maps (Appendix A), and satisfy $P(E) = 1$, immediately from $P = G_{\#}\mu$, and $P(\widehat{E}) = 1$ from $P = \widehat{G}_{\#}\hat{\mu}$; hence $P(E \cap \widehat{E}) = 1$. On $E \cap \widehat{E}$ the two answer maps agree at the first coordinate, so projecting once more,

$$\mu\big(\{\mathbf{x} : G(\mathbf{x}) = \widehat{G}(\mathbf{x})\}\big) \;=\; 1. \tag{23}$$

Finally the full-measure agreement upgrades to the whole observed support: both answer maps are continuous on it, so the agreement set is closed in it, and a relatively closed set of full measure is dense in it by the positive density (Appendix A), hence all of it. □

Lemma A3 (Pexider rigidity with a shared slope)**.** *Let $h_z, h_\varepsilon : \mathbb{R} \to \mathbb{R}$ be continuous and $\rho : \mathbb{R} \to \mathbb{R}$ be a map such that $h_z(c+e) = \rho(c) + h_\varepsilon(e)$ for every $c \in \mathbb{R}$ and almost every $e \in \mathbb{R}$. Then h_z, h_ε, and ρ are affine with one common slope.*

Proof. Step 0 (the identity holds everywhere). Fix c. The function $e \mapsto h_z(c+e) - h_\varepsilon(e)$ is continuous and equals the constant $\rho(c)$ for almost every e, hence for every e. The identity therefore holds for all $(c, e) \in \mathbb{R}^2$.

Step 1 (reduction to Cauchy's equation). Setting $e = 0$, $c = 0$, and $c = e = 0$ in turn,

$$\rho(c) = h_z(c) - h_\varepsilon(0), \qquad h_\varepsilon(e) = h_z(e) - \rho(0), \qquad h_z(0) = \rho(0) + h_\varepsilon(0). \tag{24}$$

Substituting the first two of Equation (24) back into the identity and writing $\varphi(t) := h_z(t) - h_z(0)$,

$$\varphi(c+e) \;=\; h_z(c) + h_z(e) - h_\varepsilon(0) - \rho(0) - h_z(0) \;=\; \varphi(c) + \varphi(e), \tag{25}$$

the second equality cancelling the constants by the third relation of Equation (24).

Step 2 (continuous additive maps are linear). Equation (25) is Cauchy's functional equation and φ is continuous, so $\varphi(t) = at$ for some $a \in \mathbb{R}$. By Equation (24),

$$h_z(t) = at + h_z(0), \qquad h_\varepsilon(t) = at + h_z(0) - \rho(0), \qquad \rho(t) = at + h_z(0) - h_\varepsilon(0), \tag{26}$$

all affine with the one slope a. No regularity of ρ was assumed; it inherits its form from h_z. □

Proof of Lemma 2. The hypothesis provides element-wise diffeomorphisms $\Phi_{\boldsymbol{\varepsilon}} : \hat{\varepsilon}_{\sigma(i)} = h_{\varepsilon,i}(\varepsilon_i)$ and $\Phi_{\mathbf{z}} : \hat{z}_{\sigma(i)} = h_{z,i}(z_i)$ realizing the clauses of Definition 7 with the affine maps relaxed: $\hat{p} = \Phi_{\boldsymbol{\varepsilon}\#}\, p$, the conjugation $\hat{g}^{\mathbf{v}}_{\hat{A}} \circ \Phi_{\boldsymbol{\varepsilon}} = \Phi_{\mathbf{z}} \circ g^{\mathbf{v}}_A$, and $\hat{f} = f \circ \Phi_{\mathbf{z}}^{-1}$. Composing the three exactly as in Equation (20) gives $\widehat{F}^{\mathbf{v}} \circ \Phi_{\boldsymbol{\varepsilon}} = F^{\mathbf{v}}$, so the two latent SCMs are observationally equivalent, and in particular $(\widehat{F}^{\mathbf{v}})^{-1} \circ F^{\mathbf{v}} = \Phi_{\boldsymbol{\varepsilon}}$ on $\mathbb{R}^n$. Suppose, for contradiction, that $\mathcal{M} \sim_{\mathrm{CF}} \widehat{\mathcal{M}}$ under some correspondence (σ', ρ, τ); we derive that for every node i the pair $(h_{\varepsilon,i}, h_{z,i})$ is affine with a common slope, against the hypothesis.

Step 0 (the candidate permutation is σ). Fix i, a regime $\mathbf{v}$, and $c \in \mathbb{R}$, and take the two-world family $\theta = (\mathbf{v}; (\mathbf{v}, \iota))$ with ι the *do*-intervention with intervention value c at node i. By Lemma A2, at every record $\mathbf{x}$ in the observed support the answer maps agree; substituting $\mathbf{x} = F^{\mathbf{v}}(\boldsymbol{\varepsilon})$, using $(\widehat{F}^{\mathbf{v}})^{-1} \circ F^{\mathbf{v}} = \Phi_{\boldsymbol{\varepsilon}}$, and applying $\hat{f}^{-1} = \Phi_{\mathbf{z}} \circ f^{-1}$ to both sides turns that agreement into

$$\hat{g}^{\mathbf{v}}_{[\hat{\iota}]}\big(\Phi_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon})\big) \;=\; \Phi_{\mathbf{z}}\big(g^{\mathbf{v}}_{[\iota]}(\boldsymbol{\varepsilon})\big) \qquad \text{for every } \boldsymbol{\varepsilon} \in \mathbb{R}^n, \qquad \hat{\iota} = do\big(\hat{z}_{\sigma'(i)} = \tau_i(c)\big). \tag{27}$$

Read Equation (27) at the coordinate $\sigma'(i)$. On the left the *do*-intervention fixes the coordinate at the constant $\tau_i(c)$, independent of $\boldsymbol{\varepsilon}$. On the right the coordinate is $h_{z,j}\big(g^{\mathbf{v}}_{[\iota]}(\boldsymbol{\varepsilon})_j\big)$ with $j := \sigma^{-1}(\sigma'(i))$. If $j \neq i$, node j keeps its mechanism, $g^{\mathbf{v}}_{[\iota]}(\boldsymbol{\varepsilon})_j = m^{\mathbf{v}}_j(\cdot) + \varepsilon_j$ moves with ε_j over its full support, and $h_{z,j}$ is injective, so the right side is not constant, a contradiction. Hence $\sigma'(i) = \sigma(i)$ for every i: the correspondence uses the coordinate pairing of the construction.

Step 1 (the per-unit Pexider identity). Take instead the two-world family with the *perfect* intervention ι with intervention value c at node i; the correspondence answers with the perfect value $\rho_i(c)$ at $\sigma(i)$, and Equation (27) holds with this pair. At the coordinate $\sigma(i)$: on the right, $g^{\mathbf{v}}_{[\iota]}(\boldsymbol{\varepsilon})_i = c + \varepsilon_i$, giving $h_{z,i}(c+\varepsilon_i)$; on the left, node $\sigma(i)$ of the derived hatted model retains its exogenous variable, giving $\rho_i(c) + \hat{\varepsilon}_{\sigma(i)} = \rho_i(c) + h_{\varepsilon,i}(\varepsilon_i)$. Hence

$$h_{z,i}(c+\varepsilon_i) \;=\; \rho_i(c) + h_{\varepsilon,i}(\varepsilon_i) \qquad \text{for every } c \in \mathbb{R} \text{ and every } \varepsilon_i \in \mathbb{R}. \tag{28}$$

Step 2 (rigidity). Equation (28) is the hypothesis of Lemma A3 with $h_z = h_{z,i}$, $h_\varepsilon = h_{\varepsilon,i}$, and $\rho = \rho_i$, so $h_{z,i}$ and $h_{\varepsilon,i}$ are affine with a common slope, and i was arbitrary. This contradicts the hypothesis that for some node the pair $(h_{\varepsilon,i}, h_{z,i})$ is not affine with a common slope. No correspondence exists, and $\mathcal{M} \not\sim_{\mathrm{CF}} \widehat{\mathcal{M}}$. □

D Proof of Proposition 3: causal effects under the isomorphism

The effects of Definition 6 are those of Pearl (2001), the direct effect written in the notation of Correa et al. (2021, Eq. 2); the indirect effect is the total indirect effect of Robins & Greenland (1992), which is why $\mathrm{TE} = \mathrm{NDE} + \mathrm{NIE}$ holds exactly; it is the indirect term of the decomposition in Pearl (2001, Thm. 3), whose own indirect effect holds the treatment at x instead.

Proposition 3 (Isomorphism preserves causal effects up to scale)**.** *Let $\mathcal{M} \sim_{\mathrm{LSI}} \widehat{\mathcal{M}}$ with the permutation σ, scales s_i, and offsets c_i of Definition 7; let z_X, z_M, z_Y be causal variables of $\mathcal{M}$ and x, x' two intervention values for z_X, with $\hat{x} := \tau_X(x) = s_X x + c_X$ and $\hat{x}' := \tau_X(x') = s_X x' + c_X$ the corresponding values for $\hat{z}_{\sigma(X)}$. Then, for every $\mathbf{u}$ and $\mathbf{v}$, the effects of Definition 6 for $\hat{z}_{\sigma(X)}, \hat{z}_{\sigma(M)}, \hat{z}_{\sigma(Y)}$ in $\widehat{\mathcal{M}}$ equal s_Y times those for z_X, z_M, z_Y in $\mathcal{M}$, with s_Y the scale of Definition 7 at Y:*

$$\widehat{\mathrm{TE}}_{\hat{x}\to\hat{x}'} = s_Y\,\mathrm{TE}_{x\to x'}, \qquad \widehat{\mathrm{NDE}}_{\hat{x}\to\hat{x}'} = s_Y\,\mathrm{NDE}_{x\to x'}, \qquad \widehat{\mathrm{NIE}}_{\hat{x}\to\hat{x}'} = s_Y\,\mathrm{NIE}_{x\to x'}. \tag{6}$$

The three effects are thus identified up to the common factor s_Y, and their proportions exactly.

Proof. Write (σ, s_i, b_i, c_i) and $\Phi_{\boldsymbol{\varepsilon}}, \Phi_{\mathbf{z}}$ for the data of Definition 7, fix the regime $\mathbf{v}$ of Definition 6 and a value of $\mathbf{u}$, and write $\hat{X} := \sigma(X)$, $\hat{M} := \sigma(M)$, $\hat{Y} := \sigma(Y)$. The proof has three steps: the single-treatment worlds correspond, the nested world corresponds, and the expectations follow.

Step 1 (single-treatment worlds). Let ι be the do-intervention $do(z_X{=}x)$ and $\hat{\iota}$ its image under the correspondence of Equation (13), $do(\hat{z}_{\hat{X}}{=}\tau_X(x))$ with $\tau_X(x) = s_X x + c_X = \hat{x}$. Lemma A1 and the decoder clause give $f \circ \Phi_{\mathbf{z}}^{-1} \circ \hat{g}^{\mathbf{v}}_{[\hat{\iota}]} \circ \Phi_{\boldsymbol{\varepsilon}} = f \circ g^{\mathbf{v}}_{[\iota]}$, and f is injective, so

$$\hat{g}^{\mathbf{v}}_{[\hat{\iota}]} \circ \Phi_{\boldsymbol{\varepsilon}} \;=\; \Phi_{\mathbf{z}} \circ g^{\mathbf{v}}_{[\iota]}. \tag{29}$$

Reading Equation (29) at the coordinates $\hat{Y}$ and $\hat{M}$ of the element-wise $\Phi_{\mathbf{z}}$, for every $\boldsymbol{\varepsilon}$,

$$\hat{Y}_{\hat{x}}\big(\Phi_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon})\big) = s_Y Y_x(\boldsymbol{\varepsilon}) + c_Y, \qquad \hat{M}_{\hat{x}}\big(\Phi_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon})\big) = s_M M_x(\boldsymbol{\varepsilon}) + c_M = \tau_M\big(M_x(\boldsymbol{\varepsilon})\big), \tag{30}$$

and likewise at x'.

Step 2 (the nested world). For a fixed value m of z_M (a vector when M is a set of nodes, τ_M then acting coordinate-wise) the do-intervention $do(z_X{=}x',\ z_M{=}m)$ is an intervention of Definition 2, and the correspondence carries it to $do(\hat{z}_{\hat{X}}{=}\hat{x}',\ \hat{z}_{\hat{M}}{=}\tau_M(m))$; the argument of Step 1 applied to this pair gives, for every $\boldsymbol{\varepsilon}$ and every m,

$$\hat{Y}_{\hat{x}',\tau_M(m)}\big(\Phi_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon})\big) = s_Y Y_{x',m}(\boldsymbol{\varepsilon}) + c_Y. \tag{31}$$

Put $m := M_x(\boldsymbol{\varepsilon})$, the unit's own mediator value under $do(z_X{=}x)$. By Equation (30), $\tau_M(m) = \hat{M}_{\hat{x}}(\Phi_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}))$, so the left side of Equation (31) is the nested counterfactual of $\widehat{\mathcal{M}}$ at the unit $\Phi_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon})$:

$$\hat{Y}_{\hat{x}',\hat{M}_{\hat{x}}}\big(\Phi_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon})\big) = s_Y Y_{x',M_x}(\boldsymbol{\varepsilon}) + c_Y, \tag{32}$$

which is the one nested counterfactual that Equation (5) uses; the remaining outcomes $Y_{x,M_x} = Y_x$ and $Y_{x',M_{x'}} = Y_{x'}$ are the single-treatment worlds of Step 1: holding z_M at the value it already takes under the same treatment changes no variable, since the solution is unique.

Step 3 (expectations). The exogenous-variable clause of Definition 7 gives $\hat{p}(\,\cdot \mid \mathbf{u}) = \Phi_{\boldsymbol{\varepsilon}\#}\, p(\,\cdot \mid \mathbf{u})$, so the expectation under $\hat{p}$ of each left side in Equations (30)–(32) equals the expectation under p of its right side:

$$\mathbb{E}\big[\hat{Y}_{\hat{x}}\big] = s_Y\mathbb{E}[Y_x] + c_Y, \quad \mathbb{E}\big[\hat{Y}_{\hat{x}'}\big] = s_Y\mathbb{E}[Y_{x'}] + c_Y, \quad \mathbb{E}\big[\hat{Y}_{\hat{x}',\hat{M}_{\hat{x}}}\big] = s_Y\mathbb{E}[Y_{x',M_x}] + c_Y. \tag{33}$$

Substituting Equation (33) into the total effect and into Equation (5), with $Y_{x,M_x} = Y_x$ and $Y_{x',M_{x'}} = Y_{x'}$, the offset c_Y cancels in every difference and Equation (6) follows; dividing, the proportions agree whenever $\mathrm{TE}_{x\to x'} \neq 0$. □

E Proof of Theorem 4: reaching the isomorphism

Theorem 4 (Reaching latent SCM isomorphism)**.** *Let $\mathcal{M}, \widehat{\mathcal{M}}$ be latent SCMs satisfying Assumptions 1 and 2. If $\mathcal{M}$ and $\widehat{\mathcal{M}}$ are observationally equivalent, then $\mathcal{M} \sim_{\mathrm{LSI}} \widehat{\mathcal{M}}$.*

Roadmap. The proof splits along the two assumptions, one subsection each. Appendix E.1 consumes Assumption 1 and none of the interventions of Assumption 2. Two facts independent of the graph set it up: the decoder is shared by every regime, so the reparametrization $\Phi := \hat{f}^{-1} \circ f$ is *regime-independent*; and the additive-noise solution map $g_A^{\mathbf{v}}$ is triangular with unit Jacobian, so the SCM contributes no regime-dependent volume. Conditioning on a regime that carries the index contrast of Assumption 1(ii) therefore leaves a nonlinear-ICA problem with the single mixing $F^{\mathbf{v}}$, solved up to a per-slice element-wise affine map $\psi_{\mathbf{v}}$ (Lemma A4, following the proof of Theorem 1 of Khemakhem et al., 2020 and Appendix B of Sorrenson et al., 2020); rigidity of the contrast (Lemma A5) then glues the per-slice maps into one global exogenous-variable map (Corollary A6). The gluing is where the argument departs from its nonlinear-ICA progenitors: there the mixing is a single fixed map, whereas here $F^{\mathbf{v}}$ moves with the regime and only f is shared, so the per-slice conclusions must additionally be made consistent. Appendix E.2 consumes Assumption 2 on top of the exogenous-variable matching: the designated regimes turn the exogenous-variable map into the causal-variable map, the mechanisms, and the graph, all with the *same* scales (Proposition A7), locating the unrecorded targets first and then reading each non-root coordinate off at its designated regime. The closing proof assembles the clauses of Definition 7 from these pieces. Appendix E.3 ends the section by locating the exact step that fails when the index also moves the mechanisms: the separation of Definition 1 is a condition of the argument, not a convention.

Throughout, by Assumption 1(i) the prior is Gaussian with sufficient statistics $\mathbf{T}(\boldsymbol{\varepsilon}) := (\varepsilon_i, \varepsilon_i^2)_{i\in[n]} \in \mathbb{R}^{2n}$ and natural parameters $\boldsymbol{\lambda}(\mathbf{u}) := (\lambda_{i,1}(\mathbf{u}), \lambda_{i,2}(\mathbf{u}))_{i\in[n]}$,

$$\log p(\boldsymbol{\varepsilon} \mid \mathbf{u}) \;=\; \mathbf{T}(\boldsymbol{\varepsilon})^\top \boldsymbol{\lambda}(\mathbf{u}) - \textstyle\sum_i \log Z_i(\mathbf{u}), \tag{34}$$

and hatted objects are the candidate's analogues. Fix the index values $\mathbf{u}^0, \dots, \mathbf{u}^{2n}$ of Assumption 1(ii) and call a regime whose slice carries all of them a *contrast regime*; Assumption 1(ii) makes every regime a contrast regime.

E.1 Identifying the exogenous variables

This subsection consumes Assumption 1 and observational equivalence only; the structure of the interventions does not enter.

Lemma A4 (Per-slice element-wise affine identification of the exogenous variables)**.** *Let $\mathcal{M}, \widehat{\mathcal{M}}$ be observationally equivalent and let $\mathbf{v}$ be a contrast regime. Then the slice reparametrization $\psi_{\mathbf{v}} := (\widehat{F}^{\mathbf{v}})^{-1} \circ F^{\mathbf{v}}$ is element-wise affine: there are a permutation $\sigma_{\mathbf{v}}$ of $[n]$ and constants $a_{\mathbf{v},i} \neq 0$, $d_{\mathbf{v},i}$ with $(\psi_{\mathbf{v}}(\boldsymbol{\varepsilon}))_{\sigma_{\mathbf{v}}(i)} = a_{\mathbf{v},i}\varepsilon_i + d_{\mathbf{v},i}$.*

Proof. The argument has four steps. Step 0 sets up $\psi_{\mathbf{v}}$ on the observed support; Step 1 turns observational equivalence on the slice into a log-density matching; Step 2 differences the matching across the index values and inverts the contrast matrix, leaving an affine relation between the two statistics vectors; Step 3 uses the two-parameter structure of the statistics to force that relation to be element-wise.

Step 0 (a change of exogenous variables on the observed support). Both $F^{\mathbf{v}} = f \circ g_A^{\mathbf{v}}$ and $\widehat{F}^{\mathbf{v}} = \hat{f} \circ \hat{g}_{\widehat{A}}^{\mathbf{v}}$ are injective C^1 immersions (Definition 1: the decoder is a diffeomorphism onto its image and $g_A^{\mathbf{v}}$ is a bijection of $\mathbb{R}^n$), and the exogenous variables have a positive density on $\mathbb{R}^n$ by Assumption 1(i). Equal laws give images that agree up to a null set, not equal images, so the argument runs on the observed support S, their common part (Appendix A): $\psi_{\mathbf{v}} = (\widehat{F}^{\mathbf{v}})^{-1} \circ F^{\mathbf{v}}$ is a bijection from $U := (F^{\mathbf{v}})^{-1}(S)$ onto $\widehat{U} := (\widehat{F}^{\mathbf{v}})^{-1}(S)$, both of full measure in $\mathbb{R}^n$, with $\hat{\boldsymbol{\varepsilon}} = \psi_{\mathbf{v}}(\boldsymbol{\varepsilon})$. Steps 1–3 hold at every $\mathbf{x} \in S$, and the end of Step 3 extends the conclusion to all of $\mathbb{R}^n$.[1]

Step 1 (change of variables; the SCM contributes no volume). Observational equivalence on the slice reads $p(\mathbf{x} \mid \mathbf{u}, \mathbf{v}) = \hat{p}(\mathbf{x} \mid \mathbf{u}, \mathbf{v})$ for every $\mathbf{u}$ carried by the slice. Pushing each prior through its fixed slice mixing $\mathbf{x} = F^{\mathbf{v}}(\boldsymbol{\varepsilon})$,

$$\log p(\mathbf{x} \mid \mathbf{u}, \mathbf{v}) = \log p(\boldsymbol{\varepsilon} \mid \mathbf{u}) + \log\bigl|\det J_{(g_A^{\mathbf{v}})^{-1}}(\mathbf{z})\bigr| + \log\bigl|\det J_{f^{-1}}(\mathbf{x})\bigr|. \tag{35}$$

[1] For $d > n$ the injective C^1 decoder is read as a C^1 embedding; this gives meaning to the volume elements below: $\log|\det J_{f^{-1}}(\mathbf{x})|$ abbreviates the n-dimensional volume element $-\frac{1}{2}\log\det(J_f^\top J_f)$ on $f(\mathbb{R}^n)$, of which only the $\mathbf{u}$-independence is used.

In any topological order of A the solution map $g_A^{\mathbf{v}} : \boldsymbol{\varepsilon} \mapsto \mathbf{z}$ is lower-triangular with $\partial z_i/\partial \varepsilon_i = 1$: by Equation (1), ε_i enters $z_i = m_i^{\mathbf{v}}(\mathbf{z}_{\mathrm{pa}(i)}) + \varepsilon_i$ additively while the parents depend on strictly earlier exogenous coordinates only. Hence $|\det J_{g_A^{\mathbf{v}}}| = 1$ and the middle term of Equation (35) vanishes identically. The decoder term $\log|\det J_{f^{-1}}(\mathbf{x})|$ is $\mathbf{u}$-independent. Crucially, because $F^{\mathbf{v}}$ is fixed within the slice, the realized exogenous variables $\boldsymbol{\varepsilon} = (F^{\mathbf{v}})^{-1}(\mathbf{x})$ are a $\mathbf{u}$-*independent function of* $\mathbf{x}$, which is what the separation of Definition 1 secures: the index enters only the prior and the regime only the mechanisms. Equating the two models at a fixed $\mathbf{x}$ and substituting Equation (34),

$$\begin{aligned}\mathbf{T}(\boldsymbol{\varepsilon})^\top \boldsymbol{\lambda}(\mathbf{u}) - \textstyle\sum_i \log Z_i(\mathbf{u}) + \log|\det J_{f^{-1}}(\mathbf{x})| \\ = \ \widehat{\mathbf{T}}(\hat{\boldsymbol{\varepsilon}})^\top \hat{\boldsymbol{\lambda}}(\mathbf{u}) - \textstyle\sum_i \log \widehat{Z}_i(\mathbf{u}) + \log|\det J_{\hat{f}^{-1}}(\mathbf{x})|.\end{aligned} \tag{36}$$

Step 2 (difference across the index values; invert the contrast). Evaluate Equation (36) at $\mathbf{u} = \mathbf{u}^l$ and subtract its value at $\mathbf{u} = \mathbf{u}^0$, for $l = 1, \dots, 2n$. Every $\mathbf{u}$-independent term (both Jacobian terms) cancels and, because $\boldsymbol{\varepsilon}$ and $\hat{\boldsymbol{\varepsilon}}$ are fixed functions of $\mathbf{x}$, $\mathbf{T}(\boldsymbol{\varepsilon})$ and $\widehat{\mathbf{T}}(\hat{\boldsymbol{\varepsilon}})$ factor out of the differences, leaving

$$\mathbf{T}(\boldsymbol{\varepsilon})^\top \big[\boldsymbol{\lambda}(\mathbf{u}^l) - \boldsymbol{\lambda}(\mathbf{u}^0)\big] - \zeta_l \ = \ \widehat{\mathbf{T}}(\hat{\boldsymbol{\varepsilon}})^\top \big[\hat{\boldsymbol{\lambda}}(\mathbf{u}^l) - \hat{\boldsymbol{\lambda}}(\mathbf{u}^0)\big] - \hat{\zeta}_l, \qquad \zeta_l := \textstyle\sum_i \log \frac{Z_i(\mathbf{u}^l)}{Z_i(\mathbf{u}^0)}. \tag{37}$$

Stacking the $2n$ rows, $\mathbf{L}^\top \mathbf{T}(\boldsymbol{\varepsilon}) = \widehat{\mathbf{L}}^\top \widehat{\mathbf{T}}(\hat{\boldsymbol{\varepsilon}}) + \mathbf{b}$ with $\mathbf{L}$ the contrast matrix of Assumption 1(ii) and $\widehat{\mathbf{L}}$ its analogue. Since $\mathbf{L}$ is invertible, $\mathbf{T}(\boldsymbol{\varepsilon}) = \mathbf{Q}\,\widehat{\mathbf{T}}(\hat{\boldsymbol{\varepsilon}}) + \mathbf{c}$ with $\mathbf{Q} := (\mathbf{L}^\top)^{-1}\widehat{\mathbf{L}}^\top$. Both priors are minimal two-parameter families, so $\{\varepsilon_i, \varepsilon_i^2\}_i$ and $\{\hat{\varepsilon}_i, \hat{\varepsilon}_i^2\}_i$ are each affinely independent, and $\mathbf{T}(\boldsymbol{\varepsilon})$, $\widehat{\mathbf{T}}(\hat{\boldsymbol{\varepsilon}})$ each affinely span $\mathbb{R}^{2n}$ as their arguments range over the full-measure sets U and $\widehat{U}$. As $\hat{\boldsymbol{\varepsilon}} = \psi_{\mathbf{v}}(\boldsymbol{\varepsilon})$ is a bijection, the identity maps one spanning set onto the other; were $\mathbf{Q}$ singular, its image would be a proper affine subspace, contradicting the span. Hence $\mathbf{Q}$ is invertible.

Step 3 (the two-parameter family forces element-wise affine maps). Inverting the identity, $\widehat{\mathbf{T}}(\hat{\boldsymbol{\varepsilon}}) = \mathbf{Q}^{-1}(\mathbf{T}(\boldsymbol{\varepsilon}) - \mathbf{c})$: each $\hat{\varepsilon}_j$ is an affine combination $\sum_i(\alpha_{ji}\varepsilon_i + \beta_{ji}\varepsilon_i^2) + \gamma_j$ of the statistics, and so is $\hat{\varepsilon}_j^2$. Squaring the first expression must give the second; both sides are polynomials in $\boldsymbol{\varepsilon}$ that agree on the full-measure set U, hence identically. The square contains $\beta_{ji}^2\varepsilon_i^4$ and, for $i \neq k$, the cross term $2\alpha_{ji}\alpha_{jk}\varepsilon_i\varepsilon_k$, which no affine combination of the statistics contains, so every β_{ji} vanishes and at most one α_{ji} is nonzero; as $\hat{\varepsilon}_j$ is not constant, exactly one is. Since $\psi_{\mathbf{v}}$ is a bijection, the surviving indices form a permutation $\sigma_{\mathbf{v}}$, and $\hat{\varepsilon}_{\sigma_{\mathbf{v}}(i)} = a_{\mathbf{v},i}\varepsilon_i + d_{\mathbf{v},i}$ with $a_{\mathbf{v},i} \neq 0$ (the Gaussian case of the two-parameter reduction, Sorrenson et al., 2020, App. B). This holds on U, where $F^{\mathbf{v}} = \widehat{F}^{\mathbf{v}} \circ \psi_{\mathbf{v}}$; so $F^{\mathbf{v}}$ and $\widehat{F}^{\mathbf{v}}$ composed with the element-wise affine map, both continuous, agree on the dense set U and hence on $\mathbb{R}^n$: the two images coincide, and $\psi_{\mathbf{v}}$ is this map on all of $\mathbb{R}^n$. No moment or tail condition on the exogenous variables was used. □

A priori the map $\psi_{\mathbf{v}}$, its permutation included, may differ from slice to slice, while Definition 7 asks for one regime-independent exogenous-variable map. The next lemma is the rigidity that closes this gap.

Lemma A5 (Rigidity from an invertible contrast)**.** *Let the index values $\mathbf{u}^0, \dots, \mathbf{u}^{2n}$ make $\mathbf{L}$ invertible, and let χ be a permutation σ followed by element-wise affine maps $\chi_i(t) = \alpha_i t + \delta_i$ ($\alpha_i \neq 0$) with $\chi_{\#}\, p(\cdot \mid \mathbf{u}^l) = p(\cdot \mid \mathbf{u}^l)$ for $l = 0, \dots, 2n$. Then $\chi = \mathrm{id}$.*

Proof. Write $\mu_i(\mathbf{u}), \omega_i^2(\mathbf{u})$ for the Gaussian parameters of coordinate i, and $\Delta\lambda_{i,j} := \big(\lambda_{i,j}(\mathbf{u}^l) - \lambda_{i,j}(\mathbf{u}^0)\big)_{l=1}^{2n} \in \mathbb{R}^{2n}$ for the (i,j) row of $\mathbf{L}$ read as a vector; invertibility of $\mathbf{L}$ says that the $2n$ vectors $\{\Delta\lambda_{i,1}, \Delta\lambda_{i,2}\}_{i\in[n]}$ are linearly independent.

Step 1 (what preservation says about the parameters). Under $p(\cdot \mid \mathbf{u}^l)$ the coordinates are independent with $\varepsilon_i \sim \mathcal{N}(\mu_i(\mathbf{u}^l), \omega_i^2(\mathbf{u}^l))$, and coordinate $\sigma(i)$ of $\chi(\boldsymbol{\varepsilon})$ is $\alpha_i\varepsilon_i + \delta_i$, with law $\mathcal{N}(\alpha_i\mu_i(\mathbf{u}^l) + \delta_i,\ \alpha_i^2\omega_i^2(\mathbf{u}^l))$. The preservation $\chi_{\#}\, p(\cdot \mid \mathbf{u}^l) = p(\cdot \mid \mathbf{u}^l)$ equates this law with the $\sigma(i)$-th marginal of $p(\cdot \mid \mathbf{u}^l)$:

$$\mu_{\sigma(i)}(\mathbf{u}^l) = \alpha_i\mu_i(\mathbf{u}^l) + \delta_i, \qquad \omega_{\sigma(i)}^2(\mathbf{u}^l) = \alpha_i^2\,\omega_i^2(\mathbf{u}^l), \qquad l = 0, \dots, 2n. \tag{38}$$

Substituting $\lambda_{i,1} = \mu_i/\omega_i^2$ and $\lambda_{i,2} = -1/(2\omega_i^2)$ turns Equation (38) into relations linear in the natural parameters with $\mathbf{u}$-free coefficients,

$$\lambda_{\sigma(i),2}(\mathbf{u}^l) = \alpha_i^{-2}\lambda_{i,2}(\mathbf{u}^l), \qquad \lambda_{\sigma(i),1}(\mathbf{u}^l) = \alpha_i^{-1}\lambda_{i,1}(\mathbf{u}^l) - 2\delta_i\alpha_i^{-2}\lambda_{i,2}(\mathbf{u}^l), \tag{39}$$

and differencing each relation at $\mathbf{u}^l$ against its value at $\mathbf{u}^0$, for $l = 1, \dots, 2n$, gives the same relations for the rows of $\mathbf{L}$:

$$\Delta\lambda_{\sigma(i),2} = \alpha_i^{-2}\Delta\lambda_{i,2}, \qquad \Delta\lambda_{\sigma(i),1} = \alpha_i^{-1}\Delta\lambda_{i,1} - 2\delta_i\alpha_i^{-2}\Delta\lambda_{i,2}. \tag{40}$$

Step 2 (the permutation is trivial). In block form, Equation (40) expresses the row pair of coordinate $\sigma(i)$ through the row pair of coordinate i:

$$\big(\Delta\lambda_{\sigma(i),1},\ \Delta\lambda_{\sigma(i),2}\big) \;=\; \big(\Delta\lambda_{i,1},\ \Delta\lambda_{i,2}\big) \begin{pmatrix} \alpha_i^{-1} & 0 \\ -2\delta_i\alpha_i^{-2} & \alpha_i^{-2} \end{pmatrix}, \tag{41}$$

so both rows of coordinate $\sigma(i)$ lie in $\operatorname{span}\{\Delta\lambda_{i,1}, \Delta\lambda_{i,2}\}$. Suppose $\sigma(i) \neq i$ for some i. Then the two pairs are four *distinct* rows of $\mathbf{L}$ confined to a subspace of dimension at most two, and four vectors in such a subspace are linearly dependent; so the rows of $\mathbf{L}$ are dependent and $\mathbf{L}$ is singular, contradicting the invertibility hypothesis. Hence $\sigma(i) = i$ for every i.

Step 3 (the affine part is trivial). With $\sigma = \mathrm{id}$, Equation (40) rearranges to

$$(1 - \alpha_i^{-2})\,\Delta\lambda_{i,2} = 0, \qquad (1 - \alpha_i^{-1})\,\Delta\lambda_{i,1} + 2\delta_i\alpha_i^{-2}\,\Delta\lambda_{i,2} = 0. \tag{42}$$

Both $\Delta\lambda_{i,1}$ and $\Delta\lambda_{i,2}$ are nonzero (a zero row would already make $\mathbf{L}$ singular) and, as a pair, linearly independent. The first relation of Equation (42) is a scalar multiple of the nonzero $\Delta\lambda_{i,2}$, so $\alpha_i^2 = 1$. If $\alpha_i = -1$, the second relation reads $2\Delta\lambda_{i,1} + 2\delta_i\Delta\lambda_{i,2} = 0$, a vanishing linear combination of the independent pair with a nonzero coefficient, which is impossible; so $\alpha_i = 1$. The second relation then reduces to $2\delta_i\Delta\lambda_{i,2} = 0$, and $\Delta\lambda_{i,2} \neq 0$ forces $\delta_i = 0$. Hence $\chi = \mathrm{id}$. □

The two lemmas combine into the subsection's conclusion: one exogenous-variable map serves every contrast regime.

Corollary A6 (One global exogenous-variable map)**.** *Let $\mathcal{M}, \widehat{\mathcal{M}}$ be observationally equivalent and let Assumption 1 hold. Then there are a permutation σ of $[n]$ and constants $s_i \neq 0$, b_i with*

$$\hat{\varepsilon}_{\sigma(i)} \;=\; s_i\varepsilon_i + b_i \qquad \textit{at every contrast regime.} \tag{43}$$

Proof. Lemma A4 makes each contrast-regime slice map $\psi_{\mathbf{v}}$ element-wise affine. We first record that every such map carries the true exogenous-variable family to the candidate's. Fix a contrast regime $\mathbf{v}$ and one of the index values $\mathbf{u}^l$ its slice carries. On the slice the two models generate $\mathbf{x} = F^{\mathbf{v}}(\boldsymbol{\varepsilon})$ with $\boldsymbol{\varepsilon} \sim p(\cdot \mid \mathbf{u}^l)$ and $\mathbf{x} = \widehat{F}^{\mathbf{v}}(\hat{\boldsymbol{\varepsilon}})$ with $\hat{\boldsymbol{\varepsilon}} \sim \hat{p}(\cdot \mid \mathbf{u}^l)$, so observational equivalence there reads

$$F^{\mathbf{v}}_{\#}\, p(\cdot \mid \mathbf{u}^l) \;=\; \widehat{F}^{\mathbf{v}}_{\#}\, \hat{p}(\cdot \mid \mathbf{u}^l), \qquad l = 0, \dots, 2n. \tag{44}$$

Applying $(\widehat{F}^{\mathbf{v}})^{-1}$ to both sides (the inverse of a diffeomorphism onto the shared support, Step 0 of Lemma A4) turns Equation (44) into

$$\psi_{\mathbf{v}\#}\, p(\cdot \mid \mathbf{u}^l) \;=\; \hat{p}(\cdot \mid \mathbf{u}^l), \qquad l = 0, \dots, 2n. \tag{45}$$

In Equation (45) only the map $\psi_{\mathbf{v}}$ depends on the regime: the priors $p(\cdot \mid \mathbf{u})$ and $\hat{p}(\cdot \mid \mathbf{u})$ are indexed by $\mathbf{u}$ alone (Definition 1: a regime changes the mechanisms, not the law of the exogenous variables), and every contrast regime carries the same values $\mathbf{u}^0, \dots, \mathbf{u}^{2n}$. Equation (45) therefore holds at every contrast regime, with one and the same family $\hat{p}(\cdot \mid \mathbf{u}^l)$ on the right. Let $\mathbf{v}_a, \mathbf{v}_b$ be any two contrast regimes; eliminating the shared right-hand side between the two instances of Equation (45),

$$\psi_{\mathbf{v}_a\#}\, p(\cdot \mid \mathbf{u}^l) \;=\; \psi_{\mathbf{v}_b\#}\, p(\cdot \mid \mathbf{u}^l), \qquad l = 0, \dots, 2n. \tag{46}$$

Applying $\big(\psi^{-1}_{\mathbf{v}_b}\big)_{\#}$ to both sides of Equation (46),

$$\chi_{\#}\, p(\cdot \mid \mathbf{u}^l) \;=\; p(\cdot \mid \mathbf{u}^l), \qquad l = 0, \dots, 2n, \qquad \text{where } \chi := \psi^{-1}_{\mathbf{v}_b} \circ \psi_{\mathbf{v}_a}, \tag{47}$$

and χ is a permutation followed by element-wise affine maps, being a composition of two such. Lemma A5 applied to Equation (47) gives $\chi = \mathrm{id}$, that is, $\psi_{\mathbf{v}_a} = \psi_{\mathbf{v}_b}$. The pair was arbitrary, so the slice maps coincide into one element-wise affine map, written as Equation (43). □

E.2 IDENTIFYING THE CAUSAL VARIABLES AND THE GRAPH

This subsection consumes Assumption 2 on top of the exogenous-variable matching of Equation (43), which Corollary A6 supplies in every regime. The targets of the designated regimes are not recorded, and a priori an observationally equivalent candidate might account for a designated regime by modifying a different node than the truth did; Step 0 of the proof shows that it cannot, and the remaining steps read each non-root coordinate off at its designated regime.

Proposition A7 (Causal variables, mechanisms, and the graph)**.** *Assume the setting of Theorem 4 and the exogenous-variable matching of Equation* (43)*, with the permutation σ relabeled to the identity. Then there are offsets β_i such that:*

(i) ***(causal variables)*** *with the same scales s_i as Equation* (43)*,*

$$\hat{z}_i \;=\; s_i z_i + \beta_i \qquad \textit{for every } \mathbf{z} \in \mathbb{R}^n; \tag{48}$$

(ii) ***(graph)*** *$\widehat{A} = A$, orientations included;*

(iii) ***(mechanisms)*** *at every regime $\mathbf{v}$ where Equation* (43) *holds,*

$$\hat{m}_i^{\mathbf{v}}\big(\hat{\mathbf{z}}_{\hat{\mathrm{pa}}(i)}\big) \;=\; s_i\, m_i^{\mathbf{v}}(\mathbf{z}_{\mathrm{pa}(i)}) + (\beta_i - b_i). \tag{49}$$

Proof. Step 0 (the designated regimes target the same nodes). The targets of the designated regimes are not recorded, and a priori an observationally equivalent candidate might account for a designated regime by modifying a *different* node than the truth did. We show that it cannot. Let $\mathbf{v}_j$ be the designated regime with true target j.

(a) Reference pinning. Write the candidate's node-k mechanism in the truth's coordinates: $\hat{z}_k = \Phi_k(\mathbf{z})$ with $\Phi = \hat{f}^{-1} \circ f$ regime-independent (the decoder is shared), and $\hat{\varepsilon}_k = s_k \varepsilon_k + b_k$ at any regime where Equation (43) holds; the reference regime and every designated regime are contrast regimes, so all the regimes compared below qualify. At such a regime $\mathbf{v}$,

$$\hat{m}_k^{\mathbf{v}}\big(\hat{\mathbf{z}}_{\hat{\mathrm{pa}}(k)}\big) \;=\; \hat{z}_k - \hat{\varepsilon}_k \;=\; \Phi_k(\mathbf{z}) - s_k\big(z_k - m_k^{\mathbf{v}}(\mathbf{z}_{\mathrm{pa}(k)})\big) - b_k. \tag{50}$$

Both Φ_k and the affine exogenous-variable map are regime-independent, so the right-hand side depends on the regime only through $m_k^{\mathbf{v}}$. For every node $k \neq j$ the true mechanism is unchanged between $\mathbf{v}_j$ and $\mathbf{v}_0$ (Assumption 2), so Equation (50) is literally the same function of $\mathbf{z}$ at $\mathbf{v}_j$ as at $\mathbf{v}_0$: every unchanged node carries its reference mechanism. (The reference regime is known, so both models compare against the same label $\mathbf{v}_0$; no separate alignment of references is needed.)

(b) The modification lands on the true target. At node j the truth differs: $m_j^{\mathbf{v}_j} \equiv c_j^{\mathbf{v}_j}$ while $m_j^{\mathbf{v}_0}$ is the reference mechanism, so by Equation (50) the candidate's node-j mechanisms at $\mathbf{v}_j$ and $\mathbf{v}_0$ differ by $s_j\big(c_j^{\mathbf{v}_j} - m_j^{\mathbf{v}_0}(\mathbf{z}_{\mathrm{pa}(j)})\big)$, a nonzero function because Assumption 2 separates the designated mechanism from its reference. By (a), node j is the *only* node whose candidate mechanism differs between $\mathbf{v}_j$ and $\mathbf{v}_0$. Assumption 2, which the candidate satisfies with the same designated regimes and the same reference, requires its $\mathbf{v}_j$ to modify exactly one of its mechanisms relative to $\mathbf{v}_0$ and to make that mechanism a constant as in Equation (7); as j is the only differing node, the candidate's modified mechanism at $\mathbf{v}_j$ is $\hat{m}_j^{\mathbf{v}_j}$, and it is a constant function. Hence the pair $(\mathbf{v}_j, j)$ is located on both sides. Ranging over the designated regimes, the candidate's nodes modified by a designated regime are exactly the true non-root nodes; since every non-root node of the candidate is the target of one of its designated regimes (Assumption 2 on the candidate's side), the true root nodes are root nodes of the candidate as well.

Step 1 (one coordinate at its designated regime). Fix a non-root node $i \in [n]$ and its designated regime $\mathbf{v}_i$, located in Step 0. On the true side $m_i^{\mathbf{v}_i}$ is a constant as in Equation (7), say $m_i^{\mathbf{v}_i} \equiv c_i^{\mathbf{v}_i}$, so $z_i = c_i^{\mathbf{v}_i} + \varepsilon_i$ at $\mathbf{v}_i$; on the candidate side the located mechanism is likewise a constant $\hat{c}_i^{\mathbf{v}_i}$, so $\hat{z}_i = \hat{c}_i^{\mathbf{v}_i} + \hat{\varepsilon}_i$ at $\mathbf{v}_i$. Substituting Equation (43) and $\varepsilon_i = z_i - c_i^{\mathbf{v}_i}$ gives $\hat{z}_i = s_i z_i + \beta_i$ with $\beta_i := \hat{c}_i^{\mathbf{v}_i} + b_i - s_i c_i^{\mathbf{v}_i}$, on the range of $\mathbf{z}$ at $\mathbf{v}_i$. That range is all of $\mathbb{R}^n$: $g_A^{\mathbf{v}_i}$ is a bijection of $\mathbb{R}^n$ and the prior has full support. The reparametrization $\Phi = \hat{f}^{-1} \circ f$ is regime-independent because the decoder is shared, so the identity extends off $\mathbf{v}_i$: $\Phi_i(\mathbf{z}) = s_i z_i + \beta_i$ for every $\mathbf{z} \in \mathbb{R}^n$. A root node r needs no designated regime: its mechanism is a constant $c_r^{\mathbf{v}}$ in every regime, and so is the candidate's at r, a root of the candidate by Step 0, say $\hat{c}_r^{\mathbf{v}}$; at any contrast regime, substituting

Equation (43) and $\varepsilon_r = z_r - c_r^{\mathbf{v}}$ into $\hat{z}_r = \hat{c}_r^{\mathbf{v}} + \hat{\varepsilon}_r$ gives $\hat{z}_r = s_r z_r + \beta_r$ with $\beta_r := \hat{c}_r^{\mathbf{v}} + b_r - s_r c_r^{\mathbf{v}}$, regime-independent because Φ_r is. Ranging over the nodes gives Equation (48), with the *same* scales s_i as the exogenous-variable map of Equation (43): the shared-scale clause is derived, not assumed.

Step 2 (mechanisms and the graph). At any regime $\mathbf{v}$ where Equation (43) holds, substituting $\hat{z}_i = s_i z_i + \beta_i$ and $\hat{\varepsilon}_i = s_i \varepsilon_i + b_i$ into $\hat{\varepsilon}_i = \hat{z}_i - \hat{m}_i^{\mathbf{v}}(\hat{\mathbf{z}}_{\widehat{\mathrm{pa}}(i)})$ recovers Equation (49), an identity between functions of $\mathbf{z}$, since $\mathbf{z} \mapsto \hat{\mathbf{z}}$ is an element-wise bijection. The right-hand side depends on z_j exactly where the true $m_i^{\mathbf{v}}$ does, and the change of variables is element-wise, so $\hat{m}_i^{\mathbf{v}}$ depends on $\hat{z}_j$ exactly where $m_i^{\mathbf{v}}$ depends on z_j. Every regime is a contrast regime by Assumption 1(ii), so Equation (49) holds in every regime, and by Definition 1 each edge $j \to i$ of either graph is used by the mechanism of i in some regime: the candidate's parent set at each i is $\mathrm{pa}(i)$, as a set of directed edges. Hence $\widehat{A} = A$, orientations included; no orientation genericity is consumed, since Step 1 delivers the element-wise form for every admitted competitor, leaving no residual family of reversed candidates. □

Proof of Theorem 4. Relabel the permutation of Corollary A6 to the identity, as in Proposition A7; undoing the relabeling restores the σ of Definition 7. By Assumption 1(ii) every regime of $\mathcal{V}$ is a contrast regime, so Equations (43) and (49) hold on all of $\mathcal{V}$. The clauses of Definition 7 now assemble. *Exogenous variables:* Equation (43), holding in every regime, is the element-wise affine correspondence $\Phi_{\boldsymbol{\varepsilon}}$, and $\Phi_{\boldsymbol{\varepsilon}}$ carries $p(\cdot \mid \mathbf{u})$ to $\hat{p}(\cdot \mid \mathbf{u})$ for every $\mathbf{u}$: it is the slice reparametrization of the contrast regimes, and every index value is carried by one of them. *Causal variables:* Equation (48) is the correspondence $\Phi_{\mathbf{z}}$ with the same scales, and Equation (49), read mechanism-wise across all regimes, is the conjugation $\hat{g}^{\mathbf{v}}_{\widehat{A}} \circ \Phi_{\boldsymbol{\varepsilon}} = \Phi_{\mathbf{z}} \circ g^{\mathbf{v}}_A$. *Graph and decoder:* the graph is Proposition A7(ii), and $\hat{f} = f \circ \Phi_{\mathbf{z}}^{-1}$ follows from $\Phi = \hat{f}^{-1} \circ f = \Phi_{\mathbf{z}}$. Hence $\mathcal{M} \sim_{\mathrm{LSI}} \widehat{\mathcal{M}}$. □

Remark A8 (What the proof consumes from the data). The proof uses the index contrast of Assumption 1(ii) in every regime, where it matches the exogenous variables (Corollary A6) and, through them, the mechanisms and the graph (Proposition A7). Whether a single regime carrying the contrast suffices is left open.

E.3 Where the argument needs the index to stay out of the mechanisms

When the index also moves the mechanisms, most simply when a single auxiliary variable drives both the prior and the mechanisms, $\mathbf{v} = \mathbf{u}$, the argument above breaks at an exact point: the passage from Equation (36) to Equation (37) in Step 2 of Lemma A4. To locate it, run Steps 1–2 in this setting. The mixing is now $F^{\mathbf{u}} = f \circ g_A^{\mathbf{u}}$, so the realized exogenous variables $\boldsymbol{\varepsilon}^{\mathbf{u}}(\mathbf{x}) := (F^{\mathbf{u}})^{-1}(\mathbf{x})$ depend on $\mathbf{u}$: the pre-image of a fixed $\mathbf{x}$ moves with the index. The change of variables itself survives (the solution map keeps its unit Jacobian and the decoder term stays $\mathbf{u}$-free), so the analogue of Equation (36) reads

$$\begin{aligned} &\mathbf{T}\big(\boldsymbol{\varepsilon}^{\mathbf{u}}(\mathbf{x})\big)^\top \boldsymbol{\lambda}(\mathbf{u}) - \textstyle\sum_i \log Z_i(\mathbf{u}) + \log\lvert\det J_{f^{-1}}(\mathbf{x})\rvert \\ &\quad = \widehat{\mathbf{T}}\big(\hat{\boldsymbol{\varepsilon}}^{\mathbf{u}}(\mathbf{x})\big)^\top \hat{\boldsymbol{\lambda}}(\mathbf{u}) - \textstyle\sum_i \log \widehat{Z}_i(\mathbf{u}) + \log\lvert\det J_{\hat{f}^{-1}}(\mathbf{x})\rvert. \end{aligned} \tag{51}$$

Differencing at $\mathbf{u} = \mathbf{u}^l$ against $\mathbf{u} = \mathbf{u}^0$ at the fixed $\mathbf{x}$ still cancels the two Jacobian terms and leaves, with ζ_l as in Equation (37),

$$\begin{aligned} &\mathbf{T}\big(\boldsymbol{\varepsilon}^{\mathbf{u}^l}(\mathbf{x})\big)^\top \boldsymbol{\lambda}(\mathbf{u}^l) - \mathbf{T}\big(\boldsymbol{\varepsilon}^{\mathbf{u}^0}(\mathbf{x})\big)^\top \boldsymbol{\lambda}(\mathbf{u}^0) - \zeta_l \\ &\quad = \widehat{\mathbf{T}}\big(\hat{\boldsymbol{\varepsilon}}^{\mathbf{u}^l}(\mathbf{x})\big)^\top \hat{\boldsymbol{\lambda}}(\mathbf{u}^l) - \widehat{\mathbf{T}}\big(\hat{\boldsymbol{\varepsilon}}^{\mathbf{u}^0}(\mathbf{x})\big)^\top \hat{\boldsymbol{\lambda}}(\mathbf{u}^0) - \hat{\zeta}_l. \end{aligned} \tag{52}$$

Compare with Equation (37). There, the separation of Definition 1 makes the mixing fixed within the slice, so $\boldsymbol{\varepsilon}^{\mathbf{u}^l}(\mathbf{x}) = \boldsymbol{\varepsilon}^{\mathbf{u}^0}(\mathbf{x}) =: \boldsymbol{\varepsilon}$ and the two source terms collect into the single vector $\mathbf{T}(\boldsymbol{\varepsilon})$ against the contrast $\boldsymbol{\lambda}(\mathbf{u}^l) - \boldsymbol{\lambda}(\mathbf{u}^0)$; stacking the $2n$ differences produces the linear system $\mathbf{L}^\top \mathbf{T}(\boldsymbol{\varepsilon}) = \widehat{\mathbf{L}}^\top \widehat{\mathbf{T}}(\hat{\boldsymbol{\varepsilon}}) + \mathbf{b}$ that Step 2 inverts. In Equation (52) the two source terms sit at different values of the exogenous variables ($\boldsymbol{\varepsilon}^{\mathbf{u}^l}(\mathbf{x}) \neq \boldsymbol{\varepsilon}^{\mathbf{u}^0}(\mathbf{x})$ whenever the mechanisms genuinely move with $\mathbf{u}$), so no common $\mathbf{T}(\boldsymbol{\varepsilon})$ factors out: Equation (52) is not linear in any single unknown statistics vector, there is no contrast matrix to invert, and the argument stops. Separating the index from the regime is a condition of the argument, not a choice of notation.

F DERIVATION OF THE OBJECTIVE

This appendix derives the bound in Equation (10), identifies the gate $\alpha_{\mathbf{v}}$ (the variational distribution of the target) that makes it tight, and explains the two choices the display leaves implicit: where the entropy enters, and the role of the last term. It ends with the schedule under which the bound is optimized.

The bound. Fix a designated regime $\mathbf{v} \in \mathcal{D}$ with $N_{\mathbf{v}} = |D_{\mathbf{v}}|$ samples. Its target is unknown, and by Assumption 2 it is a single node for the whole regime; we model it as a latent variable $T_{\mathbf{v}} \in \{1, \dots, n\}$ with the uniform prior $\pi_i = 1/n$, shared by every sample of the regime. Since the data of $\mathbf{v}$ are generated by the mechanisms of $\mathbf{v}_0$ with the gain of the target set to zero, the marginal log-likelihood of the regime is that of a mixture with one component per candidate target and the whole regime inside each component,

$$\log p_\psi(D_{\mathbf{v}} \mid \mathbf{v}_0) \;=\; \log \sum_{i=1}^{n} \pi_i \, p_\psi(D_{\mathbf{v}} \mid \mathbf{v}_0, i), \qquad p_\psi(D_{\mathbf{v}} \mid \mathbf{v}_0, i) \;=\; \prod_{(\mathbf{x},\mathbf{u}) \in D_{\mathbf{v}}} p_\psi(\mathbf{x} \mid \mathbf{u}, \mathbf{v}_0, i). \tag{53}$$

For any distribution $\alpha_{\mathbf{v}}$ over the nodes, Jensen's inequality applied to the expectation under $\alpha_{\mathbf{v}}$ gives

$$\begin{aligned} \log \sum_i \pi_i \, p_\psi(D_{\mathbf{v}} \mid \mathbf{v}_0, i) &= \log \sum_i \alpha_{\mathbf{v},i} \, \frac{\pi_i \, p_\psi(D_{\mathbf{v}} \mid \mathbf{v}_0, i)}{\alpha_{\mathbf{v},i}} \\ &\geq \sum_i \alpha_{\mathbf{v},i} \log p_\psi(D_{\mathbf{v}} \mid \mathbf{v}_0, i) + H(\alpha_{\mathbf{v}}) - \log n, \end{aligned} \tag{54}$$

with $H(\alpha_{\mathbf{v}}) = -\sum_i \alpha_{\mathbf{v},i} \log \alpha_{\mathbf{v},i}$ and $\sum_i \alpha_{\mathbf{v},i} \log \pi_i = -\log n$. Summing Equation (54) over the designated regimes and adding the exact log-likelihood of the other regimes gives Equation (10) without its last term, up to the constant $|\mathcal{D}| \log n$. The bound is the evidence lower bound of the mixture in Equation (53) with the gate as the variational distribution of the target, and it is tight when

$$\alpha_{\mathbf{v},i} \;=\; \frac{p_\psi(D_{\mathbf{v}} \mid \mathbf{v}_0, i)}{\sum_j p_\psi(D_{\mathbf{v}} \mid \mathbf{v}_0, j)} \;=\; \mathrm{softmax}_i\big(\log p_\psi(D_{\mathbf{v}} \mid \mathbf{v}_0, i)\big), \tag{55}$$

the posterior of the target given the regime's data, whose logits are the total log-likelihoods of the regime under each candidate target. Without its last term, Equation (10) maximized over α is therefore the marginal log-likelihood of Equation (53) up to the constant, so maximizing it over ψ and α jointly maximizes the marginal likelihood over ψ, with the gate of Equation (55) at the optimum.

Where the entropy enters. The entropy in Equation (10) appears once per regime because the target is one latent variable per regime. The logits of Equation (55) are sums over the $N_{\mathbf{v}}$ samples and grow with $N_{\mathbf{v}}$, whereas $H(\alpha_{\mathbf{v}}) \leq \log n$ does not, so the optimal gate is a softmax of log-likelihood gaps of order $N_{\mathbf{v}}$ and concentrates on the best candidate as the regime grows. In that limit the bound becomes $\max_i \log p_\psi(D_{\mathbf{v}} \mid \mathbf{v}_0, i)$, the criterion of the enumeration estimator used on the pendulum (Appendix G.4), which trains one model per target assignment and keeps the one with the highest likelihood. A model with one target per sample has instead the entropy $N_{\mathbf{v}}$ times in its bound; a gate trained under that weight competes on per-sample rather than total log-likelihood gaps and stays soft however large the regime.

The regularizer $\lambda H(\bar{\alpha})$. The regimes enter Equation (54) independently, whereas Assumption 2 assigns different nodes to different designated regimes. With $\bar{\alpha}$ the average of the gates over $\mathcal{D}$, $H(\bar{\alpha})$ is maximal when the designated regimes spread their mass evenly over the nodes and decreases when two of them settle on the same node; the term is a relaxation of the constraint that the targets form a matching between regimes and nodes, which the enumeration imposes by construction. With $\lambda > 0$ the optimal gate is no longer exactly that of Equation (55), but the term stays below $\lambda \log n$ while the log-likelihood gaps grow with $N_{\mathbf{v}}$, so when the likelihood separates the regimes it leaves the most likely target of each gate unchanged and mainly shapes the early training.

The entropy schedule. The bound is optimized from a random initialization, and a gate that hardens before the mechanisms have separated the candidates can lock the target of a regime on a wrong node, from which the likelihood does not move it. We therefore anneal: training starts with the entropy weighted per sample, $N_{\mathbf{v}} H(\alpha_{\mathbf{v}})$, and lowers the weight to the regime-level form of Equation (10) over the first 60% of training, after which the bound is optimized as written. On Causal3DIdent the schedule raises MCC($\mathbf{z}$) of FlowSCM from 0.923 to the 0.955 of Table 1; the pendulum rows use the enumeration estimator, which needs neither gate nor schedule, at the price of one training run per target assignment.

G Experimental setup

The generating processes, the training and selection rules, the metrics, and the protocols behind Tables 1 and 2 are documented here; Appendix H collects the additional results.

G.1 Data generating processes

Synthetic latent SCMs. Each of the 30 graphs is drawn with $n = 5$ nodes in a fixed topological order, every edge $j \to i$ with $j < i$ present with probability 0.4 and every node after the first given at least one parent, so that the first node is the only root. The mechanisms are

$$m_i^{\mathbf{v}}(\mathbf{z}_{\mathrm{pa}(i)}) \;=\; \gamma_i(\mathbf{v})\, g_i\big(\mathbf{w}_i(\mathbf{v}) \odot \mathbf{z}_{\mathrm{pa}(i)}\big), \qquad \varepsilon_i \mid \mathbf{u} \;\sim\; \mathcal{N}\big(\mu_i(\mathbf{u}), \omega_i^2(\mathbf{u})\big), \tag{56}$$

where g_i is a fixed two-layer leaky-ReLU network (16 hidden units, slope 0.2) drawn once per graph, the coefficients $w_{ij}(\mathbf{v})$ are drawn per regime and per edge with $|w_{ij}(\mathbf{v})| \sim \mathrm{U}[1, 2]$ and a random sign, and the gains $\gamma_i(\mathbf{v}) \in \{0, 1\}$. The index takes $M_u = 20$ values, with $\mu_i(\mathbf{u}) \sim \mathrm{U}[-2, 2]$ and $\omega_i(\mathbf{u}) \sim \mathrm{U}[0.1, 3]$ drawn per value and node. There are $M_v = 24$ regimes: 20 regimes with their own coefficient tables and every gain 1, the first of which is the reference regime $\mathbf{v}_0$, and one designated regime per non-root node i that carries the coefficients of $\mathbf{v}_0$ with $\gamma_i = 0$, the perfect intervention of Assumption 2. The observation is $\mathbf{x} = f(\mathbf{z}, \boldsymbol{\eta}) \in \mathbb{R}^{10}$, with five extra coordinates $\boldsymbol{\eta}$ drawn from a fixed Gaussian independently of $\mathbf{u}$ and $\mathbf{v}$, and f one invertible leaky-ReLU layer whose square weight matrix has singular values in $[0.5, 1.5]$. Every $(\mathbf{u}, \mathbf{v})$ cell holds 104 samples, $49{,}920$ per graph. Three features go beyond the setting of Theorem 4: the 19 regimes besides the reference with coefficient tables of their own, which the theorem does not need and the ablations keep; the extra coordinates, so that $\mathbf{x} = f(\mathbf{z}, \boldsymbol{\eta})$ rather than $f(\mathbf{z})$, which FlowSCM models with the Gaussian prior of §6; and the leaky-ReLU mechanisms and mixing, continuous but not smooth.

Pendulum. The four causal variables, in topological order, are the pendulum angle θ, the light angle φ, the shadow's position s, and the shadow's length ℓ, with the graph $\theta, \varphi \to s$ and $s, \varphi \to \ell$. The two angles are exogenous, $\theta = \varepsilon_1$ and $\varphi = \varepsilon_2$, and the shadow follows from projective geometry: with the pivot at $(p_x, p_y) = (0.5, 0.13)$, the rod of length $L = 0.42$, and the ground line at $y_g = 0.86$ in image coordinates, the ball's shadow and the pivot's shadow fall at

$$g_s(\theta, \varphi) \;=\; p_x + L \sin\theta + \big(y_g - p_y - L\cos\theta\big)\tan\varphi, \qquad q(\varphi) \;=\; p_x + (y_g - p_y)\tan\varphi, \tag{57}$$

and the two children are

$$s \;=\; \gamma_s(\mathbf{v})\big(g_s(\theta, \varphi) - 0.50\big) + \varepsilon_3, \qquad \ell \;=\; \gamma_\ell(\mathbf{v})\big(\, b\,|s - q(\varphi)| - 0.13\big) + \varepsilon_4, \tag{58}$$

so that the length is measured from the pivot's shadow to the realized position of the ball's shadow, the mediator of §7.1, and the coefficient b sets the strength of that mechanism: Table 1 averages over $b \in \{0.85, 1.0, 1.15\}$, each generated with its own seed, so that the index table and the exogenous draws differ across the three as well. The mechanisms are written as deviations from the defaults 0.50 and 0.13, so that a vanishing mechanism leaves the shadow at its default position or length up to the exogenous variable. Each variable is clipped to the renderable range, $\theta \in [-0.70, 0.70]$, $\varphi \in [-0.75, 0.75]$, $s \in [0.08, 0.92]$, $\ell \in [0.02, 0.45]$, with the exogenous variable recorded as the residual, so that Equation (58) holds exactly on the stored data. The clip binds mostly on ℓ, for 11.8% of the samples at $b = 1$ and similarly at the other two strengths; on those samples the recorded exogenous variable of ℓ depends on its parents and is not Gaussian, and with the kink of $|\cdot|$ the pendulum departs there from Assumption 1(i) and from the smooth mechanisms of Definition 1. The index takes $M_u = 16$ values; per value, $\varepsilon_i \mid \mathbf{u} \sim \mathcal{N}(\mu_i(\mathbf{u}), \omega_i^2(\mathbf{u}))$ with $\mu \sim \mathrm{U}[-0.35, 0.35]$,

$\omega \sim \mathrm{U}[0.10, 0.28]$ for θ; $\mu \sim \mathrm{U}[-0.45, 0.45]$, $\omega \sim \mathrm{U}[0.12, 0.28]$ for φ; $\mu \sim 0.50 + \mathrm{U}[-0.10, 0.10]$, $\omega \sim \mathrm{U}[0.03, 0.12]$ for s; and $\mu \sim 0.13 + \mathrm{U}[-0.05, 0.05]$, $\omega \sim \mathrm{U}[0.02, 0.07]$ for ℓ. There are $M_v = 3$ regimes: the reference regime with both gains 1 and one designated regime per child with its gain 0; the two angles have no parents, so Assumption 2 asks for no regime at them, and the estimator enumerates the 12 assignments of two targets to four nodes. Every $(\mathbf{u}, \mathbf{v})$ cell holds 400 samples, 19,200 images of 64×64 RGB pixels per value of b: the rod in blue from the pivot to the ball, the sun in red at the horizontal position $p_x - 0.5 \sin \varphi$, and the shadow in green as an ellipse centered at s whose half-width, shade, and height are affine in ℓ, so that every pixel of the shadow carries its length; the scene is rendered at four times the resolution and downsampled. The scene follows the pendulum of Yang et al. (2021); the renderer, the mechanisms, the index and the regimes are ours, so the true counterfactual image of any intervention is obtained by re-rendering.

Causal3DIdent. The latent SCM, the index, the regimes, and the samples are those of the synthetic data at the first four seeds (four graphs, $n = 5$, $M_u = 20$, $M_v = 24$, 49,920 samples per graph); the decoder is the 3DIdent renderer of Zimmermann et al. (2021), as used for Causal3DIdent (von Kügelgen et al., 2021), in place of f. Each causal variable is standardized, squashed to $t_i = \tanh(z_i/2) \in (-1, 1)$, and placed affinely into one input of the renderer: z_1 and z_2 become the object's ground-plane coordinates $2.4\,t$, and z_3, z_4, z_5 the spotlight's orbit angle, the object's hue, and the spotlight's hue, $3.0 + 2.2\,t$, an arc of the circle that never reaches the wrap point; the object's rotations, its height above the ground plane, and the background hue are fixed. The map is fixed and strictly monotone per coordinate, so it is absorbed into the decoder, and no extra coordinates enter the image. The images are 64×64 RGB, rendered deterministically, so the true counterfactual image of any intervention is obtained by re-rendering.

G.2 Training and cost

Training. On the synthetic data FlowSCM has four flow blocks of 64 hidden units on the ten coordinates and is trained by Adam at learning rate 10^{-3} for 1000 epochs, 16 seeds per graph. On the images a convolutional encoder, four stride-2 convolutions with 32 channels, GroupNorm, and SiLU followed by a linear layer, maps the image to a code $\mathbf{h} \in \mathbb{R}^D$, a mirrored decoder reconstructs the image from $\mathbf{h}$, and FlowSCM runs on the detached code; the loss is the reconstruction error plus a weight times the negative log-likelihood of the code under the flow, with $D = 4$, eight flow blocks, and weight 1 on the pendulum, and $D = 16$ and weight 0.05 on Causal3DIdent; Adam at learning rate 10^{-3}, batches of 256, 400 epochs, 16 seeds per dataset and graph. The β-VAE and iVAE use the same encoder and decoder with a Gaussian posterior and $\beta = 1$.

Parameter counts and training cost. Table 3 gives the size of every model we train and the wall-clock time of one training run. On the images the β-VAE and iVAE share FlowSCM's encoder and decoder and differ from it in what sits on the code, a Gaussian posterior with its KL term to a standard or an index-conditioned prior in place of the flow on the detached code; ILCM and SuaVE run as released, on data from their generators (Appendix G.4). The 16 seeds of a cell run in parallel, one CPU core per synthetic run.

Table 3: Parameter counts and training time per run. Synthetic runs: one CPU core, 1000 epochs; image runs: one GPU, 400 epochs. On the images the count outside the flow is the convolutional autoencoder, shared with the VAEs; the β-VAE and iVAE are not timed separately.

Model (data)	Parameters	Time per run
FlowSCM (synthetic)	47k	1 h
FlowSCM (Causal3DIdent)	1.02M (0.10M flow)	2 h, shared GH200
FlowSCM (pendulum)	0.95M	23 min, A100
β-VAE, iVAE (Causal3DIdent)	0.96M	—
β-VAE, iVAE (pendulum)	0.88M	—
ILCM (its generator, 183,700 rows)	as released	3 h, one CPU core
SuaVE (its generator, 40,000 rows)	as released	1 h, two CPU cores

G.3 METRICS

Identification. MCC($\boldsymbol{\varepsilon}$) and MCC($\mathbf{z}$) are the mean absolute correlations between the true coordinates and the learned coordinates after a one-to-one matching of the coordinates on the observational data, so that the element-wise affine maps of Definition 7 cost nothing and a permutation is not a failure; a method whose latent dimension exceeds n is matched on its n best coordinates. The two matchings are made separately; the counterfactuals and effects below use the matching of the causal variables for both. The floor quoted with the tables is the MCC between the causal and the exogenous variables of the generator after the same matching, the score of returning one set of variables when asked for the other. The graph is read from the learned mechanisms: the strength of an edge $j \to i$ is the mean absolute sensitivity $|\partial m^{\mathbf{v}_0}_{i,\psi}/\partial z_j|$ of the learned mechanism of node i to coordinate j over the observational data in the reference regime, which counts what the mechanism uses (the graph weights $\mathbf{a}_i$ of Equation (8) alone can be nonzero on an input the mechanism leaves unused). SHD is the structural Hamming distance between the true graph and the k strongest learned edges on the matched coordinates, k the true edge count, and the orientation accuracy is the fraction of the recovered adjacencies whose direction is right. Table 2(a) uses ILCM's scores, computed with its routines for every method (Brehmer et al., 2022): DCI is the disentanglement component of Eastwood & Williams (2018), read off the importance matrix of a regressor from the learned to the true causal variables, and the ENCO SHD is the structural Hamming distance of the graph that ENCO (Lippe et al., 2022a) learns from the learned latent variables and the interventional data.

Counterfactual identifiability. A counterfactual is specified in the true latent SCM, a unit's exogenous variables and an intervention of one of the three forms of Definition 2 at one node, and translated into the learned one by the correspondence that the matching and the measured affine maps induce (σ from the matching of the causal variables; the common scale s_i of ρ_i and τ_i from the response of the matched coordinate to a small controlled edit of the true variable, and their offsets from the means of the matched causal and exogenous coordinates; the soft form through the induced mechanism), as in Theorem 1. The learned latent SCM abducts the unit's exogenous variables from its factual observation, applies the translated intervention, and decodes; the error is that of the model's counterfactual observation against the true one, relative to the true change, so that the model that returns the factual observation scores 1. The perfect and the do forms set the mechanism value and the variable value of the target on a grid of values; the soft form redraws the target's mechanism from the generator's prior. On the images the error is computed in pixel space against the rendered truth, and the soft column is a pooled ratio, the sum of errors over the sum of true changes, since a redraw that saturates the renderer changes the image by less than the autoencoder floor. ObsFit is the debiased Sinkhorn divergence between model samples and a fresh draw of the data on the synthetic data and the reconstruction error on the images. The AE floor encodes and decodes the true counterfactual with the model's own autoencoder.

Causal effects. For every treatment X and every outcome Y that X reaches through at least one other node, the mediator set M is the set of nodes on a path from X to Y, and the treatment moves from each unit's natural value to $x' = \mu_X \pm \sigma_X$; the true effects follow Definition 6 in the true latent SCM, and pairs whose true total effect is below 0.05 standard deviations of Y are skipped. The proof of Proposition 3 holds unit by unit, so its scaling covers this contrast from the natural value as well. In the learned latent SCM the treatment value is translated by the scale and offset of the coordinate matched to X (the controlled-edit scale above on the synthetic data; on the images, where an edited image leaves the data, the least-squares slope and intercept of the matched coordinate regressed on the true variable over the training data of every regime, whose designated regimes break the dependence of each variable on its parents), the two worlds (the treatment moved; the treatment moved with the mediators held at their factual values), whose difference gives the total indirect effect, are computed by the learned mechanisms on the learned coordinates, and the effects are read on the coordinate matched to Y and divided by its measured scale s_Y, the quantity Proposition 3 equates to the true effect; the proportions NDE/TE and NIE/TE are compared without any scale, and their errors coincide. Errors are absolute differences, relative to the true total effect for the three effects, averaged over the pairs, the two levels, and the graphs; the proportion error is capped at 5, the score of a proportion left undefined by a model total effect of exactly zero. Every counterfactual and every effect is computed on units drawn anew from the generator, $2{,}000$ per graph or strength on the images, rendered anew, and $4{,}000$ on the synthetic data, which no model has seen in training. A pair whose outcome coordinate is not identified cannot be scored on the latent level, since the effect

Table 4: The Oracle and FlowSCM at the pendulum strengths b of Table 1: training loss (nats), MCC($\boldsymbol{\varepsilon}$) and MCC($\mathbf{z}$) of the seed the training loss selects.

	Oracle			FlowSCM		
b	loss	MCC($\boldsymbol{\varepsilon}$)	MCC($\mathbf{z}$)	loss	MCC($\boldsymbol{\varepsilon}$)	MCC($\mathbf{z}$)
0.85	2.73	0.76	0.72	2.17	0.98	0.99
1.0	2.31	0.87	0.92	1.99	0.96	0.99
1.15	3.11	0.56	0.55	2.46	0.92	0.99

is divided by that coordinate's scale: pairs with $|s_Y|$ below one tenth of the median scale of the model's coordinates are excluded and counted. No pair of Table 1 or of Appendix H.1 falls under it: the smallest relative scale is 0.23 on Causal3DIdent, 0.38 on the synthetic data, and 0.50 on the pendulum.

G.4 Protocols of Tables 1 and 2

Details of Table 1. Selection: every model, on both datasets, is the best of 16 seeds by the training loss, without ground truth. On the pendulum the Oracle, whose mechanisms are frozen in the true units, reproduces the true latent SCM only at the true coordinates themselves, whereas FlowSCM, whose mechanisms are learned, reproduces it anywhere in the isomorphism class of Definition 7. On this dataset, 2 of the Oracle's 48 runs reach those coordinates, and the seeds the training loss selects do not (Table 4); the frozen mechanisms pass the coordinate errors into every counterfactual and effect, so the Oracle row of Table 1 (†) sits at the null, with $\Delta_{\mathrm{NDE/TE}}$ undefined because the selected seeds' total effect is zero at two strengths. Estimator variants: on the synthetic data and on Causal3DIdent, FlowSCM optimizes Equation (10) with the soft gate, its entropy weight annealed from the per-sample to the regime-level form over the first 60% of training (Appendix F); on the pendulum it enumerates the 12 target assignments and selects by likelihood. ObsFit: reconstruction RMSE. The baselines receive the transported true target value for every counterfactual. The soft form uses a pooled ratio on Causal3DIdent and is undefined on the pendulum, whose physics has no mechanism prior to redraw from. Floors: the MCC obtained by returning the wrong variables is 0.52 on Causal3DIdent and 0.78 on the pendulum; encoding and decoding the true counterfactual (AE floor) costs 0.11 / 0.10 / 0.23 and 0.13 / 0.11 on the perfect / *do* / soft columns. Effect cells: the metric of §7.1 (Appendix G.3); the two VAEs, which have no latent SCM, have no cell.

Reproductions of ILCM and SuaVE (Table 2). In its published setting, ILCM's generator with a single Gaussian exogenous distribution ($M_u = 1$), ILCM returns DCI 0.965 (0.986 under its own evaluation protocol), an ENCO SHD of 0.4–0.5, and an intervention-target accuracy of 0.98, with 466 of the 480 runs above a DCI of 0.95; the $M_u = 1$ column of Table 2(a) is this reproduction. SuaVE's released generator as shipped (40 coupled segments of $1{,}000$ rows, no interventions, its own coefficient law) gives, at $n = 5$ over 30 data seeds and 16 model seeds, MCC($\boldsymbol{\varepsilon}$)/MCC($\mathbf{z}$) $0.742/0.865$ and a graph of SHD 3.39 and orientation 0.54. Block (b) of Table 2 runs SuaVE's generator with the coefficient law of our synthetic data, $|w| \sim \mathrm{U}[1, 2]$ (Appendix G.1).

The generators of Table 2 and what each method reads. Block (a) runs a port of ILCM's released generator, a linear additive-noise latent SCM with $n = d = 5$ behind a random rotation, whose regimes are the reference and one perfect intervention per node that removes the target's mechanism and redraws its exogenous variable from the same law; on unpaired data such a regime has the distribution of Assumption 2. At $M_u = 1$ the exogenous variables are $\mathcal{N}(0, 1)$, ILCM's published setting; at $M_u = 20$ and 100 they follow the index table of Appendix G.1, $\mu_i(\mathbf{u}) \sim \mathrm{U}[-2, 2]$ and $\omega_i(\mathbf{u}) \sim \mathrm{U}[0.1, 3]$. ILCM trains on its $100{,}200$ counterfactual pairs, FlowSCM on the $183{,}700$ rows of the same budget read as single observations with their index value and regime: the reference and the non-root regimes designated, their targets unknown, and the roots' regimes entering as further regimes. Block (b) keeps every law of SuaVE's generator except its coefficient law and how the segment parameters are assigned: the exogenous parameters follow the index ($M_u = 20$) and the coefficients the regime, 20 draws plus one regime per non-root node with the reference coefficients and the target's column set to zero, under the coefficient law $|w| \sim \mathrm{U}[1, 2]$ above. SuaVE reads

the index and the regime as one label; FlowSCM reads them separately, with the same designated regimes as in block (a).

H ADDITIONAL RESULTS

H.1 FIGURE 1 IN NUMBERS

Table 5 lists the numbers plotted in Figure 1 and adds the effect columns of Table 1, which the figure omits; the metrics are those of Appendix G.3. Every cell is read at the seed the training loss selects on each graph and averaged over the 30 graphs, so the table also covers the 11 of 360 per-graph counterfactual points, all of them β-VAE and iVAE points, that fall outside the frame of Figure 1C and are not drawn there.

Table 5: Synthetic data, 30 graphs: the numbers behind Figure 1, and the causal effects of Proposition 3. Means over the graphs with the standard error beneath; the columns and the bold rule are those of Table 1. ObsFit is the Sinkhorn divergence between model samples and a fresh draw, median over graphs. Medians of the effect columns over the graphs: 0.25 / 0.09 / 0.25 / 0.15 for FlowSCM and 0.03 / 0.00 / 0.03 / 0.00 for the Oracle. The last row gives the MCC obtained by returning the wrong variables and the scores of the model that predicts no change.

	Identification				Counterfactual				Effects			
Model	MCC($\boldsymbol{\varepsilon}$) ↑	MCC($\mathbf{z}$) ↑	SHD ↓	Orient ↑	ObsFit ↓	CF error ↓ perfect	CF error ↓ *do*	CF error ↓ soft	Δ_{TE} ↓	Δ_{NDE} ↓	Δ_{NIE} ↓	$\Delta_{\mathrm{NDE/TE}}$ ↓
FlowSCM	**0.979** (0.004)	**0.956** (0.010)	0.73 (0.21)	0.99 (0.01)	**0.58**	**0.257** (0.039)	**0.203** (0.029)	**0.323** (0.038)	0.377 (0.058)	0.160 (0.048)	0.363 (0.066)	0.348 (0.099)
β-VAE	0.439 (0.019)	0.692 (0.014)	—	—	0.67	1.131 (0.052)	0.966 (0.012)	1.148 (0.092)	—	—	—	—
iVAE	0.513 (0.020)	0.756 (0.015)	—	—	0.63	1.017 (0.050)	0.915 (0.017)	1.113 (0.099)	—	—	—	—
Oracle	0.992 (0.000)	0.997 (0.000)	*given*	*given*	0.68	0.140 (0.017)	0.105 (0.011)	0.129 (0.012)	0.055 (0.012)	0.010 (0.003)	0.051 (0.012)	0.021 (0.009)
floor / null	*0.538*					*1*	*1*	*1*	*1*	*0.36*	*0.98*	

H.2 ABLATION TABLES

Table 6 gives the complete rows behind Figure 2, whose $k = 5$ point is given in the text below, and Table 8 ablates the components of the estimator, both on the synthetic data of Figure 1 under its conventions: the same generator, graphs, and seeds, every model its best of 16 seeds by the training loss, and every column of a row read at that seed, so that the rows of Table 5 reappear as the reference rows. The columns are those of Table 5; the exclusion rule of Appendix G.3 removes no pair on these rows except where marked.

Assumptions. Each row of Table 6 removes or weakens one hypothesis of Theorem 4. *Index only* drops the designated regimes of Assumption 2: the model is trained on the 20 non-designated regimes alone, so no regime realizes a perfect intervention and no graph is returned. The k rows give a designated regime to k of the four non-root nodes, from index only ($k = 0$) to the full set ($k = 4$). The M_u *ladder* weakens the sufficient variability of Assumption 1(ii): the index is truncated to its first M_u values after the full draw, so that the graph, the mechanisms, the mixing, the sample budget, and the seeds of the $M_u = 20$ row are unchanged and only the index shrinks; $M_u = 1$ removes the index (*interventions only*), and *neither* removes both resources. A designated regime at the root, added to the full set ($k = 5$ in Figure 2; the root's mechanism, zero in this generator, is replaced by a constant of one standard deviation of the root, 2.1 on average), leaves every column within its standard error: MCC($\boldsymbol{\varepsilon}$) 0.978 against 0.979, MCC($\mathbf{z}$) 0.963 against 0.956, the perfect-form counterfactual error 0.224 against 0.257, the effect errors 0.381/0.180 against 0.377/0.160, and the graph unchanged; a root is identified by Assumption 1 alone, and Assumption 2 asks for no regime there. Table 7 gives the index-only and interventions-only rows on Causal3DIdent and the pendulum, where index only drops the designated regimes from the training data and interventions only withholds the index from the prior.

Table 6: Assumption ablation on the synthetic data: one hypothesis of Theorem 4 removed or weakened per row. Means over 30 graphs with the standard error beneath; the effect columns are those of Equation (12), as in Table 1. Index only: no designated regime, so no graph is returned. k: the number of non-root nodes with a designated regime, out of four, on the full index (the full set is $k = 4$). M_u: the number of index values; $M_u = 1$ is interventions only. Neither: no index variation and no designated regime, so no graph is returned. ^{o}: ObsFit computed on the row's own standardized data. ^{s}: median over graphs; the error exceeds 2 on 5 (index only), 5 ($M_u = 1$), and 8 (neither) of the 30 graphs. ^{x}: effect pairs excluded by the rule of Appendix G.3 (index only: 2 of 177, $M_u = 11$: 2 of 177, neither: 8 of 177). The last row gives the MCC of returning the wrong variables and the scores of the model that predicts no change. Bold: the best value per column among the FlowSCM rows, the Oracle plain.

	Identification				Counterfactual				Effects			
Row	MCC($\boldsymbol{\varepsilon}$) ↑	MCC($\mathbf{z}$) ↑	SHD ↓	Orient ↑	ObsFit ↓	CF error ↓			Δ_{TE} ↓	Δ_{NDE} ↓	Δ_{NIE} ↓	$\Delta_{\mathrm{NDE/TE}}$ ↓
						perfect	*do*	soft				
Oracle	0.992 (0.000)	0.997 (0.000)	*given*	*given*	0.68	0.140 (0.017)	0.105 (0.011)	0.129 (0.012)	0.055 (0.012)	0.010 (0.003)	0.051 (0.012)	0.021 (0.009)
FlowSCM, full set	**0.979** (0.004)	**0.956** (0.010)	**0.73** (0.21)	**0.99** (0.01)	**0.58**	**0.257** (0.039)	**0.203** (0.029)	**0.323** (0.038)	**0.377** (0.058)	**0.160** (0.048)	**0.363** (0.066)	**0.348** (0.099)
$k = 3$	0.966 (0.005)	0.915 (0.008)	1.90 (0.25)	0.96 (0.02)	0.69	0.405 (0.036)	0.281 (0.022)	0.557 (0.058)	0.689 (0.080)	0.308 (0.066)	0.790 (0.108)	0.830 (0.181)
$k = 2$	0.966 (0.004)	0.881 (0.010)	2.27 (0.22)	0.97 (0.02)	0.67	0.526 (0.036)	0.386 (0.032)	0.751 (0.085)	0.764 (0.039)	0.277 (0.051)	0.788 (0.058)	0.860 (0.104)
$k = 1$	0.963 (0.005)	0.857 (0.012)	2.63 (0.22)	0.96 (0.02)	0.73	0.670 (0.078)	0.475 (0.054)	0.838 (0.077)	0.855 (0.064)	0.334 (0.066)	0.842 (0.065)	1.045 (0.134)
Index only ($k = 0$)^x	0.975 (0.005)	0.842 (0.013)	—	—	0.61	0.604 (0.093)	0.430 (0.060)	0.73^s	1.037 (0.171)	0.842 (0.241)	1.131 (0.243)	0.772 (0.167)
$M_u = 11$ ($= 2n + 1$)^x	0.970 (0.006)	0.954 (0.010)	1.03 (0.19)	0.98 (0.01)	0.71^o	0.301 (0.043)	0.208 (0.016)	0.435 (0.085)	0.387 (0.065)	0.275 (0.097)	0.508 (0.101)	0.380 (0.110)
$M_u = 6$	0.956 (0.011)	0.941 (0.010)	1.37 (0.27)	0.97 (0.02)	0.92^o	0.296 (0.032)	0.234 (0.024)	0.427 (0.054)	0.477 (0.075)	0.259 (0.072)	0.551 (0.105)	0.633 (0.147)
$M_u = 3$	0.943 (0.010)	0.923 (0.011)	1.67 (0.28)	0.91 (0.03)	1.15^o	0.338 (0.035)	0.249 (0.023)	0.484 (0.064)	0.499 (0.071)	0.330 (0.081)	0.496 (0.069)	0.631 (0.118)
$M_u = 1$ (interventions only)	0.748 (0.031)	0.809 (0.020)	3.40 (0.37)	0.81 (0.04)	1.10^o	0.735 (0.078)	0.492 (0.036)	0.80^s	0.879 (0.101)	0.502 (0.092)	0.843 (0.091)	0.848 (0.093)
Neither^x	0.718 (0.029)	0.730 (0.018)	—	—	1.18^o	1.279 (0.148)	0.757 (0.056)	1.37^s	2.067 (0.772)	1.628 (0.772)	1.195 (0.122)	0.857 (0.107)
floor / null	*0.538*					*1*	*1*	*1*	*1*	*0.36*	*0.98*	

Estimator components. Each row of Table 8 changes one component of FlowSCM at the same data and seeds, along the ladder from the target search to the estimator of the paper: the first rung locates the targets by an explicit penalty on the mechanisms, which pushes one node's parent term to zero and the other mechanisms to those of the reference regime; the second replaces the penalty by the marginal likelihood over the gate and the entropy terms of Equation (10); the third anneals the entropy weight (Appendix F), the FlowSCM of the paper. Replacing the penalty lowers every error column at a slightly higher ObsFit; the anneal then improves the graph and leaves the counterfactual and effect columns within their standard errors.

H.3 Counterfactuals on ILCM's and SuaVE's generators

Table 9 scores the methods of Table 2 on the CF error of Equation (11) and the causal effects of Equation (12): FlowSCM and ILCM on ILCM's generator at $M_u = 100$ (block (a)) and FlowSCM and SuaVE on SuaVE's generator (block (b)), each at the checkpoint that table reports, on the same test units of every dataset and with the protocol of Appendix G.3. ILCM's latent SCM is implicit: each causal variable is a function of its own exogenous variable and those of its ancestors. Its *do* intervention on a node replaces the node's exogenous variable by the value that sets the node to the intervention value and recomputes the descendants, the way ILCM's own training pairs arise. ILCM's mechanisms are not additive, so the perfect and soft forms, which keep the unit's exogenous variable at the target, are not interventions of ILCM, and the *do* form is the one the two methods share. SuaVE's latent SCM is linear with additive exogenous variables, so all three forms are interventions of its own, with its mechanisms read at the label of the factual units' cell. The effects of Definition 6 need only *do* interventions, on the treatment and on the treatment with the

Table 7: Assumption ablation on the images: FlowSCM of Table 1 with one resource removed, each model at the seed the training loss selects out of 16. Index only: the designated regimes dropped from the training data, so no graph is returned; interventions only: the index withheld from the prior. Mean (standard error) over the four graphs of Causal3DIdent and the three mechanism strengths of the pendulum; columns as in Table 1. Bold: the best value per column and dataset.

	Identification				Counterfactual				Effects			
Row	MCC($\boldsymbol{\varepsilon}$) ↑	MCC($\mathbf{z}$) ↑	SHD ↓	Orient ↑	ObsFit ↓	CF error ↓ perfect	CF error ↓ *do*	CF error ↓ soft	Δ_{TE} ↓	Δ_{NDE} ↓	Δ_{NIE} ↓	$\Delta_{\mathrm{NDE/TE}}$ ↓
Causal3DIdent												
FlowSCM	**0.951** (0.003)	**0.955** (0.012)	**0.50** (0.50)	**1.00** (0.00)	**0.0078** (0.0008)	**0.355** (0.033)	**0.323** (0.037)	**0.612** (0.132)	**0.124** (0.031)	**0.123** (0.078)	**0.165** (0.047)	**0.103** (0.060)
Index only	0.947 (0.004)	0.868 (0.018)	—	—	0.0083 (0.0009)	0.484 (0.076)	0.446 (0.076)	1.082 (0.371)	0.603 (0.202)	0.384 (0.127)	0.405 (0.220)	1.028 (0.444)
Interventions only	0.806 (0.058)	0.863 (0.037)	0.75 (0.48)	0.95 (0.05)	0.0080 (0.0008)	0.553 (0.062)	0.544 (0.065)	0.999 (0.229)	0.559 (0.392)	0.284 (0.102)	0.759 (0.472)	0.323 (0.189)
Pendulum												
FlowSCM	**0.954** (0.020)	**0.993** (0.001)	**0.00** (0.00)	**1.00** (0.00)	0.0085 (0.0003)	**0.552** (0.017)	**0.521** (0.010)	—	**0.071** (0.017)	**0.130** (0.027)	**0.085** (0.013)	**0.074** (0.002)
Index only	0.843 (0.054)	0.646 (0.049)	—	—	0.0189 (0.0074)	1.005 (0.078)	1.193 (0.130)	—	1.013 (0.013)	0.744 (0.104)	0.881 (0.055)	0.878 (0.053)
Interventions only	0.575 (0.055)	0.642 (0.050)	3.67 (0.88)	0.44 (0.29)	**0.0079** (0.0001)	1.485 (0.129)	1.496 (0.131)	—	1.135 (0.083)	0.830 (0.177)	0.957 (0.050)	1.100 (0.181)

Table 8: Estimator ablation on the synthetic data: one component of FlowSCM changed per row at the same data and seeds; columns and conventions as in Table 6. The rows are the ladder from the target search to the FlowSCM of the paper: the target located by an explicit penalty on the mechanisms; the penalty replaced by the marginal likelihood over the gate and the entropy terms of Equation (10); the entropy weight annealed (the FlowSCM of the paper). ^{x}: 20 of 177 effect pairs excluded by the rule of Appendix G.3, one graph without an included pair. Bold: the best value per column.

	Identification				Counterfactual				Effects			
Row	MCC($\boldsymbol{\varepsilon}$) ↑	MCC($\mathbf{z}$) ↑	SHD ↓	Orient ↑	ObsFit ↓	CF error ↓ perfect	CF error ↓ *do*	CF error ↓ soft	Δ_{TE} ↓	Δ_{NDE} ↓	Δ_{NIE} ↓	$\Delta_{\mathrm{NDE/TE}}$ ↓
Target located by a penalty^x	0.969 (0.003)	0.904 (0.011)	2.70 (0.29)	0.97 (0.02)	**0.57**	0.446 (0.053)	0.287 (0.023)	0.612 (0.061)	0.840 (0.118)	0.485 (0.123)	0.869 (0.112)	0.757 (0.175)
Target located by marginal likelihood	0.978 (0.004)	0.951 (0.009)	0.93 (0.21)	**0.99** (0.01)	0.59	**0.250** (0.023)	**0.185** (0.014)	0.334 (0.034)	0.423 (0.054)	0.197 (0.040)	0.422 (0.058)	0.425 (0.106)
+ annealing (= FlowSCM)	**0.979** (0.004)	**0.956** (0.010)	**0.73** (0.21)	**0.99** (0.01)	0.58	0.257 (0.039)	0.203 (0.029)	**0.323** (0.038)	**0.377** (0.058)	**0.160** (0.048)	**0.363** (0.066)	**0.348** (0.099)
floor / null	*0.538*					*1*	*1*	*1*	*1*	*0.36*	*0.98*	

mediators held at their factual values, so every method returns them. ILCM and SuaVE are trained on the raw observations and FlowSCM on standardized ones, and all are scored in the raw space. On ILCM's generator the perfect form at the value 0 leaves a root unchanged, so the error at that value alone is zero over zero; every cell pools the errors and the true changes over the values of the grid. At $M_u = 100$ ILCM's variables and graph are close to FlowSCM's (Table 2), yet its do-form CF error is higher on 29 of the 30 datasets. On SuaVE's generator, where SuaVE recovers the causal variables but neither the exogenous variables nor the orientation of the graph (Table 2(b)), its CF error is higher than FlowSCM's on every dataset in the perfect and do forms.

H.4 Counterfactuals under a non-affine reparametrization

Lemma 2 states that element-wise diffeomorphisms, the class in which causal representation learning leaves the causal variables, do not carry every counterfactual of Definition 4. Figure 3 measures it on the synthetic data. For each of the 30 graphs, the root coordinate of the FlowSCM model of Table 5 is bent by $\varphi_a(t) = m + s\,(w + a \tanh w)$ with $w = (t - m)/s$, m and s the mean and standard deviation of the coordinate over the factual test units, and the model is transported with it: the mechanism and the exogenous variable of the root, the children's reading of the root, and the

Table 9: CF error and causal effects on the generators of Table 2, for the methods of its two blocks: mean (standard error) over 30 datasets, CF errors and true changes pooled over the intervention values 0, ± 0.5, ± 1 standard deviations of the target. —: not an intervention of ILCM's latent SCM. Medians of the soft column: 0.23 (FlowSCM) in (a), 0.42 (FlowSCM) and 0.90 (SuaVE) in (b). The effect columns are those of Equation (12), as in Table 1, over the 25 datasets of each block that have a pair with a mediator and a true total effect above the floor of Appendix G.3; SuaVE returns a total effect of exactly zero on 44 of its 146 pairs, where $\Delta_{\mathrm{NDE/TE}}$ is undefined and scored 5, the cap of the column. ^{x}: effect pairs excluded by the rule of Appendix G.3 (ILCM: 4 of 106). The *null* rows are the model that predicts no change. Bold: the best value per column in each block.

	CF error ↓			Effects			
Method	perfect	*do*	soft	Δ_{TE} ↓	Δ_{NDE} ↓	Δ_{NIE} ↓	$\Delta_{\mathrm{NDE/TE}}$ ↓
(a) ILCM's generator, $M_u = 100$							
FlowSCM	**0.179** (0.019)	**0.121** (0.012)	**0.376** (0.118)	**0.112** (0.020)	**0.213** (0.057)	**0.283** (0.062)	**0.245** (0.064)
ILCM^x	—	0.814 (0.088)	—	0.568 (0.135)	0.881 (0.119)	0.612 (0.083)	0.842 (0.114)
null	*1*	*1*	*1*	*1*	*0.48*	*0.87*	
(b) SuaVE's generator, $M_u = 20$							
FlowSCM	**0.380** (0.027)	**0.310** (0.023)	**0.804** (0.379)	**0.418** (0.052)	0.456 (0.139)	**0.599** (0.166)	**1.028** (0.273)
SuaVE	0.971 (0.015)	0.923 (0.015)	1.039 (0.093)	1.004 (0.016)	**0.397** (0.090)	1.168 (0.102)	2.230 (0.236)
null	*1*	*1*	*1*	*1*	*0.35*	*1.15*	

decoder are composed with φ_a or its inverse, so that the reparametrized model is an additive-noise latent SCM with the same observational distribution, related to the original by an element-wise map that is not affine. Only a root admits such a map: at a node with parents, whose mechanism is not constant (Definition 1), the argument of Lemma A3 with the mechanism's value in place of the intervention value makes the map affine. The counterfactual error is the CF error of Equation (11), computed with the protocol of Table 5 but for interventions at the bent root only, with intervention values ± 0.5 and ± 1 standard deviations of the root and averaged over the four; at $a = 0$ it is the unbent model's error on these interventions, not the all-node error of Table 5. The soft form needs no curve of its own: a root has no parents, so its intervention function is a constant and the soft form is the perfect form. The perfect form is scored under the natural correspondence $\tilde{\rho}(c) = \varphi_a(\rho(c))$ and under the most favourable one, the single constant per intervention value closest to the original model's perfect world over the units, the best any correspondence of Definition 3 can do. With the observational fit unchanged, the perfect form degrades from 0.32 to 0.49 at a bend of a third of a standard deviation and to 0.77 at 1.3 even under the most favourable correspondence, and reaches the null under the natural one, as Lemma 2 predicts; the *do* form, which keeps no exogenous variable at the target, is unchanged by the bend.

H.5 Exogenous variables outside the Gaussian family

Assumption 1(i) asks for a Gaussian exogenous law, and FlowSCM fits one. Table 10 trains the FlowSCM of Table 5 on the same graphs, index tables, and seeds with the standardized law of every exogenous variable replaced: Student-t with $\nu = 10, 5, 3$ (excess kurtosis 1, 6, and infinite) and a symmetric two-component Gaussian mixture with modes at ± 0.8 standard deviations (excess kurtosis -0.8), the index still setting the mean and the scale. Both laws lie outside the exponential family on which Appendix E.1 rests and outside FlowSCM's Gaussian prior; both are location-scale families, so the affine class remains the target of the metrics. The bimodal law identifies and transports as the Gaussian does, within the standard errors of the MCC, orientation, perfect and *do* columns, and the SHD rises from 0.87 to 1.27. Under the Student-t laws, $\nu = 10$ is indistinguishable from the Gaussian, $\nu = 5$ loses 0.01 of MCC and up to 0.1 of counterfactual error, and $\nu = 3$ loses 0.04 of MCC($\boldsymbol{\varepsilon}$) and MCC($\mathbf{z}$) while the graph is unchanged and the counterfactual error rises on every form, most on the soft form. Only the heaviest tails ($\nu = 3$) thus cost noticeably, mostly in the counterfactuals.

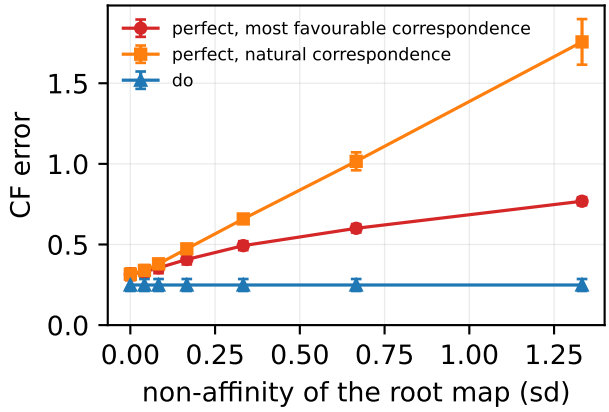


Figure 3: Counterfactual error of FlowSCM on the synthetic data after a non-affine bend of one root coordinate, against the bend's distance from its best affine fit in standard deviations of the coordinate; mean and standard error over the 30 graphs, for the *do* form and for the perfect form under the two correspondences.

Table 10: FlowSCM on the synthetic data with the exogenous law replaced; means over the 30 graphs with the standard error beneath. ^{s}: median over the graphs; the error exceeds 2 on 1 and 4 of the 30 graphs for $\nu = 5$ (*do* and soft forms), on 10, 9, and 18 for $\nu = 3$ (the three forms), and on 1 for the bimodal law (soft form). ^{o}: ObsFit computed on each law's own standardized data. ^{r}: the checkpoints of Table 5, scored like the other rows.

Exogenous law	Identification				Counterfactual			
	MCC($\boldsymbol{\varepsilon}$) ↑	MCC($\mathbf{z}$) ↑	SHD ↓	Orient ↑	ObsFit ↓	CF error ↓		
						perfect	*do*	soft
Gaussian^{r}	0.976 (0.004)	0.957 (0.010)	0.87 (0.22)	0.99 (0.01)	0.58	0.257 (0.039)	0.203 (0.029)	0.323 (0.038)
Student-t, $\nu = 10$	0.976 (0.004)	0.952 (0.010)	0.93 (0.24)	0.99 (0.01)	0.61^{o}	0.234 (0.018)	0.173 (0.012)	0.372 (0.037)
Student-t, $\nu = 5$	0.964 (0.010)	0.944 (0.011)	0.97 (0.23)	0.97 (0.01)	0.77^{o}	0.328 (0.057)	0.20^{s}	0.41^{s}
Student-t, $\nu = 3$	0.935 (0.005)	0.913 (0.009)	0.93 (0.24)	0.99 (0.01)	1.46^{o}	0.61^{s}	0.59^{s}	2.8^{s}
Bimodal, modes ± 0.8	0.967 (0.008)	0.953 (0.011)	1.27 (0.23)	0.99 (0.01)	0.59^{o}	0.277 (0.026)	0.230 (0.020)	0.31^{s}

H.6 Factual and counterfactual image samples

Figures 4 and 5 show factual and counterfactual images on Causal3DIdent and on the pendulum for the rendered truth and every model; the captions name the intervention targets.

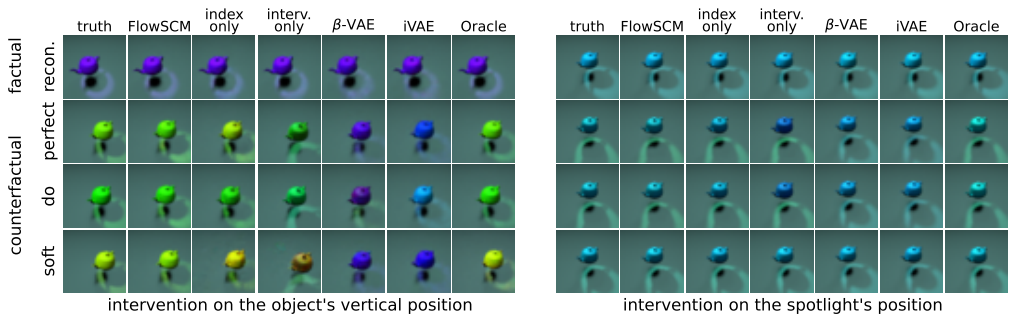


Figure 4: Factual and counterfactual images on Causal3DIdent. Rows: the factual image (for the models, their reconstruction) and the counterfactual under the perfect, do, and soft interventions; columns: the rendered truth, then every model, the two ablations of Table 7 included. The intervention target is the object's vertical position in the image (its second ground-plane coordinate) on the left and the spotlight's position on the right.

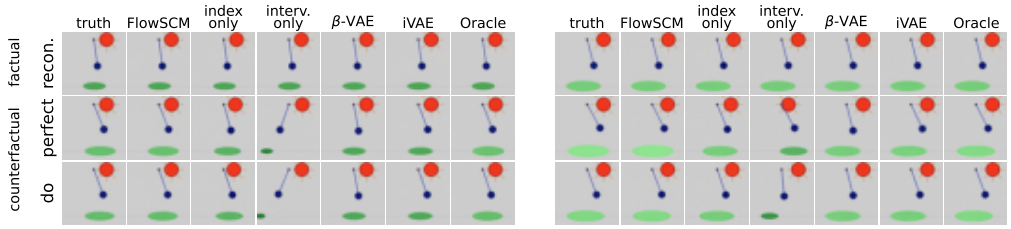


Figure 5: Factual and counterfactual images on the pendulum, two units. Rows: the factual image (for the models, their reconstruction) and the counterfactual under the perfect and do interventions on the pendulum angle at $+1$ standard deviation; columns: the rendered truth, then every model, the two ablations of Table 7 included. The angle moves the rod and, through the mechanism, the shadow's position and length; the soft form is undefined on this dataset (Appendix G.4).