LATENT STRUCTURAL CAUSAL MODELING FOR HIGH-DIMENSIONAL COUNTERFACTUAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual image generation requires propagating low-dimensional causal interventions to complex images while preserving unit-specific variation. Existing approaches typically couple causal propagation with high-dimensional image synthesis, requiring the generative model to serve both causal reasoning and image rendering. We introduce a manifold-guided factorization that separates these two roles: causal interventions are propagated entirely through a low-dimensional latent structural causal model (SCM), while a separate conditional generator renders the resulting latent state to a high-dimensional image. It enables simple graph-constrained structural mechanisms and distinguishes unit-specific semantic exogenous variation. We instantiate the framework with an additive latent SCM and conditional flow matching, and provide identification, error decomposition, and generalization results. The advantages of the proposed framework are demonstrated through extensive empirical studies.

1 INTRODUCTION

Counterfactual inference asks how the outcome of an observed individual would have changed under an intervention. Structural causal models (SCMs) answer this through abduction, action, and prediction, preserving unit-specific exogenous variables across counterfactual worlds (Pearl, 1995; 2009). For image outcomes, however, an intervention on a few causal variables must be propagated into millions of correlated voxels while preserving unaffected characteristics. High-dimensional observations concentrate near lower-dimensional structure rather than filling the ambient space, suggesting that causal mechanisms need not operate directly on every observed coordinate. The key difficulty is therefore not simply high-dimensional generation, but propagating a low-dimensional causal intervention into a high-dimensional image while preserving unit-specific variation.

Existing causal generative approaches combine causal reasoning with powerful high-dimensional synthesis models, but typically couple the two tasks within the same generative architecture. As a result, the high-dimensional generative model must encode not only how an image is synthesized, but also how causal interventions propagate to the image. This coupling creates two challenges. First, structural mechanisms must connect low-dimensional causal variables to high-dimensional outcomes despite the large dimensional mismatch. Second, counterfactual inference must preserve unit-specific semantic exogenous variation, whereas image generation also involves observation-level variability. Without explicitly separating these roles, semantic variation that should remain fixed under an intervention can become conflated with stochastic variation associated with image synthesis.

We address these challenges with a manifold-guided factorization that separates causal propagation from image synthesis. Although an image has a large ambient dimension, its causally structured variation can often be represented in a much lower-dimensional space. We therefore propagate interventions entirely through a latent SCM whose graph-constrained mechanisms operate among states of comparable dimension, and use a separate conditional generator only to synthesize the high-dimensional observation. Placing all structural states on comparable scales reduces the dimensional burden of learning parent-to-child mechanisms. The latent-SCM exogenous residuals preserve unit-specific semantic variation in non-intervened assignments, whereas generator noise controls observation-level rendering variability. The principle is generator-agnostic. We instantiate it with an additive latent SCM and conditional flow matching (Lipman et al., 2023; Albergo et al., 2025).

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.

Given factual observations, residual abduction, intervention, and prediction produce a counterfactual terminal latent state, from which the conditional generator samples the corresponding image distribution. Conditionally fixing generator noise additionally gives a model-dependent paired factual-counterfactual realization. Under explicit representation, structural, invariance, and overlap assumptions, we establish identification of the counterfactual observation law and decompose its error into latent structural and conditional-generation components. We also characterize how low-dimensional structural modeling reduces mechanism complexity within stated model classes.

We evaluate the framework on Morpho-MNIST, where the known data-generating SCM enables direct assessment of structural recovery and graph-consistent counterfactuals. We further demonstrate its applicability to OASIS brain MRI, assessing subject preservation and anatomically meaningful changes under interventions within a working causal graph.

Our contributions are:

- We introduce a structural-generative factorization that completes causal propagation in a low-dimensional latent SCM before high-dimensional synthesis, while separating unit-specific semantic exogenous variation from observation-level rendering noise.
- We establish counterfactual identification and error decomposition, and characterize structural generalization, conditional on a valid learned representation and stated causal assumptions.
- We demonstrate direct recovery of counterfactual structural propagation and graph-consistent image counterfactuals on Morpho-MNIST, and subject-preserving, anatomically meaningful counterfactual generation on OASIS brain MRI under a working causal graph.

2 Literature Review

2.1 Causal Inference and Structural Causal Models

Counterfactual inference asks how an outcome would have differed under an alternative intervention for the same observed unit (Rubin, 1974; Holland, 1986; Pearl, 2009; Hernán & Robins, 2020). Structural causal models (SCMs) represent a data-generating system through a directed graph, structural assignments, and exogenous variables (Pearl, 1995; 2009). Unlike conditioning, an intervention replaces selected structural assignments and propagates the resulting changes through the graph. Individual counterfactuals are constructed through abduction, action, and prediction: infer the factual exogenous variables, modify the intervened assignments, and evaluate the resulting system while preserving the remaining exogenous variables (Pearl, 2009; Pearl & Mackenzie, 2018). Identification from observational data further requires assumptions on causal structure, confounding, regime invariance, and support (Shpitser & Pearl, 2006; King & Zeng, 2006; Crump et al., 2009). Recent work also studies counterfactual identifiability through dynamic optimal transport (De Sousa Ribeiro et al., 2026).

These principles imply that a counterfactual generator needs to do more than condition on a target attribute: it must propagate an intervention through the causal structure while preserving the factual unit's non-intervened exogenous state. The validity of this construction further depends on the assumed causal structure, regime invariance, and adequate support. The remaining challenge is how to implement this logic when the outcome is a high-dimensional image.

2.2 Deep Generative Counterfactual Models

Deep generative models, including latent-variable, adversarial, diffusion, and flow-based models, provide flexible tools for modeling high-dimensional conditional distributions (Louizos et al., 2017; Kocaoglu et al., 2018; Ho et al., 2020; Lipman et al., 2023; Albergo et al., 2025; Tong et al., 2024).

Several approaches combine causal structure with deep generative models for counterfactual generation. Deep SCMs use neural structural assignments and abduction-action-prediction to support counterfactual inference for image observations (Pawlowski et al., 2020), while CausalGAN and CausalVAE incorporate prescribed causal structure into adversarial or latent-variable generative models (Kocaoglu et al., 2018; Yang et al., 2021). Probabilistic and variational approaches similarly combine causal inference with high-dimensional generative modeling (Sauer & Geiger, 2021; De Sousa Ribeiro et al., 2023; Wu et al., 2025). Diffusion-based approaches include diffusion causal

models, causal diffusion autoencoders, lesion-localization methods, latent-drifting approaches, and semantic abduction (Sanchez & Tsaftaris, 2022; Komanduri et al., 2024; Sanchez et al., 2022; Yeganeh et al., 2025; Rasal et al., 2025). Related work applies counterfactual diffusion to brain images, incorporates anatomical priors, and studies classifier-free guidance for counterfactual diffusion (Fontanella et al., 2025; Li et al., 2026; Xia et al., 2026). These methods differ in how they represent interventions, abduct factual information, and couple factual evidence to generated counterfactuals. Our framework is distinguished along four dimensions: where causal propagation occurs, how factual unit-specific variation is abducted, whether structural exogenous variation is separated from rendering noise, and whether the high-dimensional terminal mechanism is explicitly factorized into low-dimensional causal propagation and conditional observation synthesis. We give more details in Section D.

3 METHODOLOGY

We consider a Directed Acyclic Graph (DAG) with low dimensional causal variables $\boldsymbol{Y} = (Y_1, \ldots, Y_{K-1})$ and a high dimensional outcome $Y_K = \boldsymbol{X} \in \mathbb{R}^D$. Although $\boldsymbol{X}$ may contain millions of coordinates, its causally structured variation occupies far fewer degrees of freedom. Thus, the ambient dimension should govern synthesis but need not govern causal propagation.

3.1 COUNTERFACTUAL TARGET

For nodes $A \subseteq \{1, \ldots, K-1\}$, conditioning on $\boldsymbol{Y}_A = \boldsymbol{a}$ selects units that naturally realized $\boldsymbol{a}$, whereas $do(\boldsymbol{Y}_A = \boldsymbol{a})$ replaces their structural assignments and propagates the intervention through the graph. Thus, $P(\boldsymbol{X} \mid \boldsymbol{Y}_A = \boldsymbol{a})$ generally differs from $P(\boldsymbol{X} \mid do(\boldsymbol{Y}_A = \boldsymbol{a}))$.

Our target is the individual counterfactual distribution for the unit that produced factual evidence $(\boldsymbol{X}, \boldsymbol{Y}) = (\boldsymbol{x}, \boldsymbol{y})$:

$$P(\boldsymbol{X}^{\boldsymbol{a}} \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}), \tag{1}$$

where the superscript $\boldsymbol{a}$ denotes the world induced by $do(\boldsymbol{Y}_A = \boldsymbol{a})$. Computing this distribution requires abduction of unit-specific exogenous variables, action that replaces the intervened assignments, and prediction that preserves the remaining factual exogenous variables. Directly implementing these structural assignments is difficult when the terminal node is a very high-dimensional observation. We therefore factor the terminal mechanism through a common low-dimensional structural space.

3.2 MANIFOLD-GUIDED FACTORIZATION OF THE TERMINAL MECHANISM

The manifold hypothesis is used here as a modeling principle rather than as a claim that the complete noisy data distribution lies on one smooth manifold. We posit that a terminal state $\boldsymbol{h}_K$ of intrinsic dimension $m \ll D$ captures the low-dimensional directions along which causal changes propagate, whereas observation-level variability is left to a separate noise variable. This distinction is essential: compressibility of images motivates the latent bottleneck, but does not imply that every pixel-level fluctuation is causal.

Let $\boldsymbol{h}_k \in \mathcal{H} \subseteq \mathbb{R}^m$ denote the latent state of node k, and let $\boldsymbol{h} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_K)$. Let $\mathcal{C}_k(\boldsymbol{h})$ collect the latent states of its parents, $\mathcal{C}_k(\boldsymbol{h}) = \{\boldsymbol{h}_j : j \in \mathrm{pa}(k)\}$. We use the population factorization

$$\boldsymbol{h}_k = f_k(\mathcal{C}_k(\boldsymbol{h}), \boldsymbol{U}_k), \quad k = 1, \ldots, K, \tag{2}$$

$$\boldsymbol{X} = G(\boldsymbol{h}_K, \boldsymbol{\epsilon}) \tag{3}$$

where f_k is the structural mechanism of node k, $\boldsymbol{U}_k$ is node-level exogenous noise, G is the observation generator, and $\boldsymbol{\epsilon}$ is observation-level generator noise. We impose no particular forms on f_k or G at this stage, retaining a general latent factorization. The key factorization is that all causal propagation is completed through Equation (2), whereas Equation (3) performs observation synthesis only and does not itself propagate interventions.

More specifically, all structural assignments in Equation (2) operate in the latent space. Using a shared dimension m therefore does not assert that scalar age, binary sex, and an MRI have equal intrinsic complexity. Rather, it gives every parent-to-child mechanism a uniform interface for modeling interactions among compatible latent states, much as language models embed heterogeneous tokens before modeling their contextual interactions. Node-specific latent dimensions m_k are possible in principle, but we use a common m as a practical implementation choice.

In contrast, Equation (3) maps the terminal latent state to the high-dimensional observation space: for fixed $\boldsymbol{\epsilon}$, varying $\boldsymbol{h}_K$ traces a low-dimensional subset of the observation space, as formalized in Section C.1.

Taken together, the latent SCM and conditional generator separate causal propagation from observation synthesis. The shared representation places the terminal observation on the same structural scale as its causal ancestors, so structural learning can focus on how an intervention changes the terminal latent state while the generator learns its rendering into correlated pixels or voxels. The method exploits low intrinsic dimension without requiring geodesics, charts, or an explicit estimate of $\mathcal{M}$. The benefit of this factorization therefore requires more than compressibility: the learned representation must preserve intervention-relevant semantics and admit sufficiently simple structural mechanisms for causal propagation.

3.3 Model Estimation and Counterfactual Inference

We now implement the factorization in Equations (2) and (3) in three stages: first learn the conditional observation model, then fit latent structural propagation, and finally perform counterfactual inference.

Stage 1: Learning the conditional observation model. Let $y_K = \boldsymbol{x}$. A single encoder family maps every observed node to $\mathcal{H}$:

$$\boldsymbol{h}_k = E_k(y_k) \in \mathcal{H}, \qquad k = 1, \dots, K, \tag{4}$$

where $E_K = E_{K,\phi_K}$ is the terminal encoder and $\boldsymbol{h}_K$ conditions the generator. We implement the conditional observation model $P(\boldsymbol{X} \mid \boldsymbol{h}_K)$ with conditional flow matching (Tong et al., 2024; Lipman et al., 2023); flow matching is an implementation of the observation model, not a restriction of the general framework. For a data sample $\boldsymbol{x}_1$, base noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and $t \sim \mathrm{Uniform}(0, 1)$, we form the interpolation $\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_1 + \sigma_t \boldsymbol{\epsilon}$ with target velocity $v_t^\star = \dot{\alpha}_t \boldsymbol{x}_1 + \dot{\sigma}_t \boldsymbol{\epsilon}$. The conditional velocity field is trained by

$$\mathcal{L}_{\mathrm{CFM}}(\theta, \phi_K) = \mathbb{E}_{\boldsymbol{x}_1, \boldsymbol{\epsilon}, t}\left[\| v_\theta(\boldsymbol{x}_t, t, E_K(\boldsymbol{x}_1)) - v_t^\star \|_2^2 \right]. \tag{5}$$

Thus, the terminal encoder and flow network are optimized jointly so that $\boldsymbol{h}_K = E_K(\boldsymbol{x}_1)$ summarizes the observation-level condition used throughout the transport. Given a terminal state $\boldsymbol{h}_K$, generation transports fresh base noise to the observation space through

$$\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = v_\theta(\boldsymbol{x}_t, t, \boldsymbol{h}_K), \qquad \boldsymbol{x}_0 = \boldsymbol{\epsilon}. \tag{6}$$

The interpolation schedule, endpoint conditions, and conditional inversion are detailed in Section A.2. Only after this joint stage do we freeze E_K, so no structural-loss gradients update the terminal encoder. Thus this stage learns the conditional observation model before any structural regressions are fitted. The flow-matching objective trains E_K for conditional image synthesis but does not establish its validity as a structural representation. Our theoretical results therefore assume representation consistency (Assumption 1), and Section 4 examines the structural information retained by E_K empirically.

Stage 2: Learning the latent structural model. We then specialize the general latent mechanisms to the additive SCM

$$\boldsymbol{h}_k = m_k(\mathcal{C}_k(\boldsymbol{h})) + \boldsymbol{U}_k, \qquad k = 1, \dots, K, \tag{7}$$

where m_k is the systematic structural response of node k and $\boldsymbol{U}_k$ is its exogenous noise. This additive-noise specialization makes counterfactual abduction transparent through fitted residuals, at the cost of restricting structural exogenous variation to enter additively on the learned latent scale. For a root node, $\mathcal{C}_k(\boldsymbol{h}) = \varnothing$, and $m_k(\varnothing)$ is a learnable constant that represents its mean latent state. Thus roots and nonroots share the same additive SCM; the former simply have an empty parent context.

Let $\phi = \{\phi_1, \phi_2, \dots, \phi_K\}$ collect the latent-SCM parameters, including the node encoders and mechanisms, with ϕ_K denoting the terminal-encoder component. For $\boldsymbol{h}_{n,k} = E_k(y_{n,k})$, we fit the latent SCM by the simple regression objective

$$\mathcal{L}_{\mathrm{SCM}}(\phi) = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} \mathcal{D}_k(\boldsymbol{h}_{n,k}, m_k(\mathcal{C}_k(\boldsymbol{h}_n))), \tag{8}$$

During this stage, ϕ_K remains fixed after flow-matching training, whereas the remaining latent-SCM parameters are fitted using $\mathcal{L}_{\mathrm{SCM}}$. Squared error is the default choice for $\mathcal{D}_k$. Each m_k reads only $\mathcal{C}_k(\boldsymbol{h})$ and predicts $\boldsymbol{h}_k$, so lightweight graph-constrained regressors suffice. Thus $\mathcal{L}_{\mathrm{SCM}}$ learns causal propagation, whereas $\mathcal{L}_{\mathrm{CFM}}$ learns observation synthesis; critically, $\boldsymbol{U}_k \neq \boldsymbol{\epsilon}$. The former is semantic, node-level variation preserved in non-intervened assignments across counterfactual worlds, whereas the latter is observation-level generator noise. The low-dimensional bottleneck encourages this separation, but does not by itself guarantee that E_K excludes observation-level variation; the intended semantics is formalized through the representation-consistency condition in Assumption 1, conditional on which the structural objective learns causal propagation.

Stage 3: Counterfactual inference. Given factual $(\boldsymbol{x}, \boldsymbol{y})$, we first encode $\hat{\boldsymbol{h}}_k = E_k(y_k)$, with $y_K = \boldsymbol{x}$, and then apply abduction, action, prediction, and generation. *Abduction* computes the factual residuals:

$$\hat{\boldsymbol{U}}_k = \hat{\boldsymbol{h}}_k - m_k\Big(\mathcal{C}_k(\hat{\boldsymbol{h}})\Big), \qquad k = 1, \ldots, K, \tag{9}$$

Action replaces every intervened assignment with its encoded intervention value:

$$\boldsymbol{h}_k^{\boldsymbol{a}} = E_k(a_k), \qquad k \in A. \tag{10}$$

Prediction evaluates the remaining assignments in topological order:

$$\boldsymbol{h}_k^{\boldsymbol{a}} = m_k(\mathcal{C}_k(\boldsymbol{h}^{\boldsymbol{a}})) + \hat{\boldsymbol{U}}_k, \qquad k \notin A. \tag{11}$$

The factual residual enters each non-intervened assignment in the forward SCM recursion, whereas an intervention removes the residual from the assignment it replaces. At this point, causal propagation is complete: the intervention has been propagated through the latent SCM to obtain $\boldsymbol{h}_k^{\boldsymbol{a}}$, before any high-dimensional image is generated. Finally, *generation* draws $\boldsymbol{X}^{\boldsymbol{a}} \sim P_\theta(\boldsymbol{X} \mid \boldsymbol{h}_K^{\boldsymbol{a}})$ by solving the flow ODE in Equation (6). Resampling $\boldsymbol{\epsilon}$ yields the counterfactual observation distribution under a fresh-noise coupling, where observation-level noise is treated as world-specific acquisition variability conditionally independent of the factual acquisition noise given the counterfactual terminal state. Alternatively, conditional ODE inversion at $\hat{\boldsymbol{h}}_K = E_K(\boldsymbol{x})$ yields $\hat{\boldsymbol{\epsilon}}$, which can be held fixed to produce a fixed-noise paired counterfactual realization $\boldsymbol{X}^{\boldsymbol{a}}(\hat{\boldsymbol{\epsilon}})$. In both cases, the semantic latent-SCM residuals $\hat{\boldsymbol{U}}_k$ are preserved in non-intervened assignments across worlds; the two constructions differ only in the coupling of observation-level noise. The fixed-noise paired realization is model-dependent and is not uniquely identified by the observational distribution.

Algorithms 1 and 2 summarize the complete training and counterfactual inference procedures.

3.4 Identification and Statistical Guarantees

We study *validity* of residual abduction, latent propagation, and conditional synthesis, and *estimation* error in the latent SCM and generator. A final result relates latent-mechanism complexity to generalization conditional on the representation.

Assumption 1 (Identification conditions)**.** *The latent DAG includes relevant common causes and follows the additive assignments in Equation* (7)*; the encoders recover a valid structural-scale representation, including $E_K(\boldsymbol{X}) = \boldsymbol{h}_K$ almost surely; $P_{X|h_K}$ is invariant across regimes; and the counterfactual parent configurations and terminal states of interest lie in identified support.*

These conditions respectively support residual abduction and propagation, structural interpretation of the learned representation, transport of the observation law, and identification at the intervention target; their scope is discussed in Section C.2. Under the first two conditions, the residuals in Equation (9) identify a unique latent counterfactual state through the recursion in Stage 3; Proposition 2 gives the formal statement.

Theorem 1 (Identification of the counterfactual observation law)**.** *Under Assumption 1, the counterfactual observation distribution conditional on the factual latent state is identified by*

$$P\big(X^a \in B \mid \boldsymbol{h} = \boldsymbol{h}^{\mathrm{obs}}\big) = P_{X|h_K}(B \mid \boldsymbol{h}_K^a) \tag{12}$$

for every measurable $B \subseteq \mathcal{X}$. Thus, counterfactual generation exactly composes residual abduction, latent propagation, and conditional synthesis.

Theorem 1 identifies the counterfactual observation distribution $P(X^a \mid \boldsymbol{h} = \boldsymbol{h}^{\text{obs}})$, rather than a unique observation-level pairing with the factual image. A fixed-noise paired realization additionally holds the conditionally inverted generator noise fixed, which defines a model-dependent coupling not identified by the observational distribution alone.

Proposition 1 (Counterfactual error decomposition)**.** *Let μ_a and $\widehat{\mu}_a$ be the true and fitted distributions of $\boldsymbol{h}_K^a$. Suppose $G(\cdot, \boldsymbol{\epsilon})$ is uniformly L_G Lipschitz and $\sup_{\boldsymbol{h}_K, \boldsymbol{\epsilon}} \|\widehat{G}(\boldsymbol{h}_K, \boldsymbol{\epsilon}) - G(\boldsymbol{h}_K, \boldsymbol{\epsilon})\|_2 \leq \varepsilon_G$. If the true and fitted counterfactuals use the same distribution of $\boldsymbol{\epsilon}$, then*

$$W_1\big(P_{\widehat{X}^a}, P_{X^a}\big) \leq L_G W_1(\widehat{\mu}_a, \mu_a) + \varepsilon_G. \tag{13}$$

The bound separates latent causal error, scaled by generator sensitivity, from conditional-generation error. Under uniform consistency, Proposition 4 gives counterfactual consistency.

Generalization illustration from structural simplicity. The following result concerns mechanism generalization conditional on a representation, not counterfactual validity or representation recovery.

Theorem 2 (Description length generalization bound)**.** *Let q be a prefix program encoding a candidate family of node mechanisms $\boldsymbol{m}_q = (m_{q,1}, \ldots, m_{q,K})$, with code length $\ell(q)$ in bits, and let $L_q(\boldsymbol{h})$ be the per-sample structural loss in Equation* (22)*, which lies in $[0, 1]$. Let $R(q) = \mathbb{E}_P L_q(\boldsymbol{h})$ and let $\widehat{R}_N(q)$ be its empirical mean over N independent observations. Then, with probability at least $1 - \delta$, every prefix program q satisfies*

$$R(q) \leq \widehat{R}_N(q) + \sqrt{\frac{\ell(q) \log 2 + \log(1/\delta)}{2N}}. \tag{14}$$

For coordinatewise mechanism classes with at most c description bits per scalar response, Theorem 2 yields $O(\sqrt{mc/N})$ for an m-dimensional latent mechanism and $O(\sqrt{Dc/N})$ for a direct D-dimensional mechanism. Thus, for $m \ll D$, the bound scales with intrinsic rather than ambient dimension within these matched classes; Proposition 3 states the specialization formally, excluding encoder complexity and arbitrary architectures.

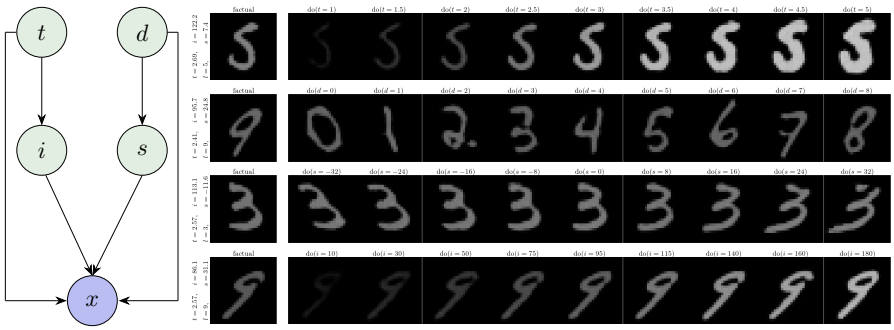
Conditional on a valid structural representation, the factorization identifies the counterfactual law, separates structural from generative error, and provides latent-mechanism generalization guarantees. The proofs are given in Section C.3. We also note that representation recovery is a distinct problem and is not established here.

4 Experiments

We evaluate the structural model-generator factorization in two complementary settings. The Morpho-MNIST simulation, for which the SCM is known, evaluates whether low-dimensional structural propagation recovers the causal mechanisms and produces graph-consistent counterfactuals. The OASIS brain-MRI application evaluates the full framework in a realistic setting, including its ability to generate semantically meaningful, subject-preserving counterfactual images under a working graph. Together, these experiments assess both the latent structural component and its composition with high-dimensional image synthesis. We follow the setting of Rasal et al. (2025), comparing our method with VAE, HVAE, VCI, and DiffSA using the same counterfactual-soundness metrics.

4.1 Simulation Study on MNIST Dataset

Data Preparation. We construct a Morpho-MNIST simulation from MNIST images (Deng, 2012; Castro et al., 2019) using the known causal mechanism of Rasal et al. (2025). Each sample begins with an MNIST source image, whose digit class determines slant, while an independently drawn thickness determines intensity. The resulting thickness, intensity, digit class, and slant jointly determine the terminal image through the morphology transformation of the known SCM in Equation (16), whose graph is shown in Figure 1a. Consequently, the simulated images exhibit an overall tendency for thicker digits to appear brighter and for larger digit classes to slant further to the right. Further dataset and graph details are provided in Section B.1.



(a) (b)

Figure 1: Morpho-MNIST simulation. The causal graph (left) has thickness and digit class as roots and the image as terminal node. The counterfactual sequences (right) show root interventions on thickness and digit class, followed by direct-child interventions on slant and intensity, each propagating to descendants while preserving non-descendants and unit-specific characteristics.

Qualitative Results. Figure 1b addresses two empirical questions for individual-level counterfactuals: whether a requested intervention propagates correctly to its descendants and whether non-descendants and unit-specific characteristics remain invariant. The first two visually prominent rows show the root interventions $do(t)$ and $do(d)$, which provide the strongest propagation checks because they should change intensity and slant, respectively, in addition to their targeted roots. The remaining $do(s)$ and $do(i)$ rows provide complementary invariance checks: they change the intervened child while leaving the other semantic branch unchanged. Across the rows, the generated samples exhibit the requested changes while retaining unit-specific visual characteristics, so each row represents counterfactuals for the same factual instance rather than independent population samples.

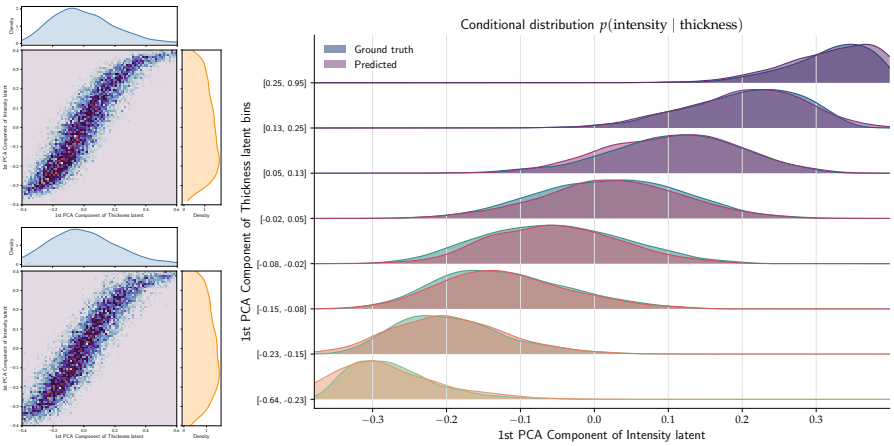
Quantitative Results. Table 1 reports counterfactual soundness results following Rasal et al. (2025). Effectiveness measures target attainment using mean absolute percentage error (MAPE) for continuous attributes and accuracy for digit class. Reversibility measures recovery after undoing an intervention, while null-intervention compatibility measures reconstruction without an intervention. Section B.4 gives the full protocol. Our method yields lower thickness MAPE than the compared methods under both interventions, and under $do(s)$ it also attains low intensity MAPE while preserving digit-class accuracy. The results are more mixed for reversibility and null-intervention compatibility, so the table reflects a trade-off across the different soundness criteria. Although the slant MAPE is relatively high, it is sensitive to targets close to zero; the corresponding median absolute percentage errors are 0.95 for $do(s)$ and 0.66 for $do(d)$.

Figure 2 evaluates the latent structural component of the counterfactual error decomposition in Proposition 1 by assessing recovery of the thickness–intensity mechanism in the first PCA-component coordinates of their latent states. The fitted model reproduces the positive joint dependence induced by the SCM and the rightward shift of $p(\text{intensity} \mid \text{thickness})$ as thickness increases, rather than merely matching the marginal distributions. The match is less precise in extreme thickness bins with

Table 1: Soundness of Morpho-MNIST image counterfactuals under $do(s)$ and $do(d)$ followed the same evaluation protocol described in (Rasal et al., 2025). All MAPEs are scaled by $\times 10^{-1}$, all L_1 metrics are scaled by $\times 10^{-2}$, and Acc for digit class remains unscaled.

	Slant Intervention $do(s)$					**Class Intervention** $do(d)$					**Null**
	Effectiveness				Rev.	Effectiveness				Rev.	Comp.
Mechanism	MAPE $(t)\downarrow$	MAPE $(i)\downarrow$	MAPE $(s)\downarrow$	Acc $(d)\uparrow$	$L_1\downarrow$	MAPE $(t)\downarrow$	MAPE $(i)\downarrow$	MAPE $(s)\downarrow$	Acc $(d)\uparrow$	$L_1\downarrow$	$L_1\downarrow$
VAE (2020)	4.63	6.98	2.91	97.27	2.14	5.90	8.03	2.10	94.92	2.44	1.81
HVAE (2023)	3.39	0.49	3.88	95.02	0.62	4.39	0.51	1.23	95.31	1.62	0.01
VCI (2025)	3.08	0.65	0.91	90.04	1.60	2.63	0.63	0.91	94.62	0.99	0.66
DiffSA (2025):											
$\{\omega = 1.50,\ p_\varnothing = 0.10\}$	2.83	1.20	1.51	98.44	2.98	3.88	1.13	1.43	97.66	4.33	0.94
$\{\omega = 3.00,\ p_\varnothing = 0.10\}$	2.14	1.00	1.48	99.80	4.41	2.15	0.94	0.70	99.80	5.35	2.17
$\{\omega = 4.50,\ p_\varnothing = 0.10\}$	2.13	0.80	1.78	99.80	5.35	7.67	0.94	0.76	99.93	6.15	2.87
Ours	**0.97**	**0.39**	3.85*	**99.83**	<u>1.22</u>	**1.09**	0.67	3.12*	99.12	1.41	0.87

Note: For some test examples, the denominator in MAPE of slant is close to zero, which can inflate the reported values. The corresponding Median APE is 0.95 for $do(s)$ and 0.66 for $do(d)$.



(a) (b)

Figure 2: Evaluation of the latent structural component of the counterfactual error decomposition through recovery of the thickness–intensity mechanism on MNIST data.

limited empirical support, but the overlap in the central regions indicates that the fitted mechanism recovers the local conditional distributions needed for counterfactual propagation. We have also evaluated the latent distribution statistics of the fitted model against the ground truth, including Maximum Mean Discrepancy (MMD), Fréchet distance, and Cross-covariance cosine similarity, which are reported in Table 7 in the appendix.

Table 2 directly compares the predicted intensity descendant with its ground-truth counterfactual under a thickness intervention, before image generation. Preserving the abducted intensity residual reduces root mean squared error (RMSE) from 18.558 to 5.561 and normalized RMSE (NRMSE) from 0.384 to 0.114 relative to propagation without residual preservation; NRMSE is defined in Equation (18) of the appendix. This comparison isolates the role of unit-specific semantic variation in structural propagation; the evaluation protocol is detailed in Section B.5.

Table 2: Counterfactual intensity recovery before image generation; NRMSE uses ground-truth standard deviation.

	$do(t = t') \to i^{\mathrm{CF}}$	
Model	**RMSE** ↓	**NRMSE** ↓
OURS (full)	5.561	0.114
w/o residual preservation	18.558	0.384

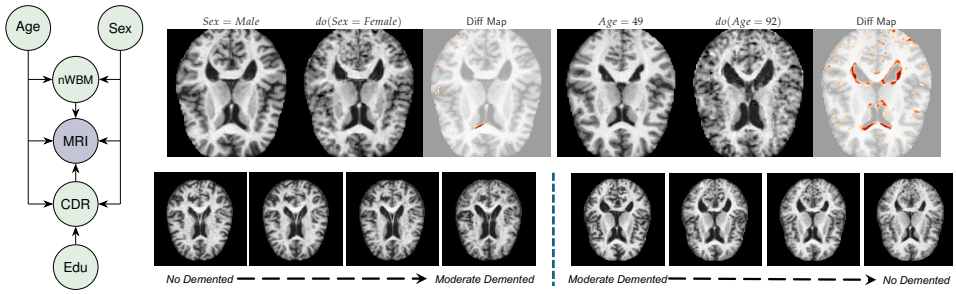
Ablation Study. We ablate structural propagation, preservation of the latent-SCM exogenous residuals $\boldsymbol{U}_k$, and the common latent structural representation. Table 3 shows that removing structural propagation sharply increases slant MAPE and reduces digit-class accuracy under both $do(s)$ and $do(d)$. Propagating conditional means without preserving residuals likewise degrades slant attainment and digit-class accuracy, especially under $do(d)$. Replacing the common structural bottleneck also consistently worsens the effectiveness metrics for both interventions.

Table 3: Ablation study of Morpho-MNIST image counterfactuals under $do(s)$ and $do(d)$.

	SLANT INTERVENTION $do(s)$				**CLASS INTERVENTION** $do(d)$			
MECHANISM	MAPE (t) ↓	MAPE (i) ↓	MAPE (s) ↓	Acc (d) ↑	MAPE (t) ↓	MAPE (i) ↓	MAPE (s) ↓	Acc (d) ↑
OURS (baseline)	**0.97**	**0.39**	**3.85**	**99.83**	**1.09**	**0.67**	**3.12**	**99.12**
w/o structural propagation	1.37	3.93	61.26	43.38	1.18	0.87	29.21	67.33
w/o residual preservation	1.07	0.74	9.79	84.37	1.30	4.53	25.15	68.92
w/o common structural bottleneck	1.07	0.80	16.12	78.32	1.40	1.82	27.01	72.12

4.2 REAL-WORLD APPLICATION

We use axial MRI slices and associated age, sex, education, normalized whole-brain volume (nWBM), and Clinical Dementia Rating (CDR) from the longitudinal OASIS cohort as a real-world showcase (Marcus et al., 2010). Dataset details and the distributions of the variables used by the model are provided in Section B.2. The parsimonious working graph in Figure 3a places age and sex upstream



(a) (b)

Figure 3: OASIS working graph (left) and subject-level counterfactual MRIs (right), generated while holding the abducted exogenous variables fixed for non-intervened assignments. The examples show sex and age interventions with difference maps (top) and bidirectional CDR interventions (bottom), where the leftmost image in each sequence is factual.

of nWBM, CDR, and MRI; education upstream of CDR; and nWBM and CDR upstream of MRI, reflecting the available variables and the role of education as a cognitive-reserve proxy (Stern, 2012). This graph supports an interpretable methodological demonstration but is neither uniquely identified from OASIS nor presented as a clinical causal model. We intervene on each covariate to test whether the generator alters relevant image structure while preserving subject-specific anatomy. Figure 3 shows representative sex, age, and CDR interventions, with additional results given in the appendix.

The counterfactuals in Figure 3b visually preserve global subject anatomy while showing minor localized changes under the sex intervention, ventricular enlargement with age, and greater ventricular and sulcal prominence with increasing CDR that attenuates along the reverse path (Blinkouskaya & Weickenmeier, 2021; Furtner & Prayer, 2021). We further quantify counterfactual changes under CDR interventions using three measurements derived from pretrained MRI segmentation, with the evaluation protocol detailed in Section B.3. Across increasing CDR levels, ventricular area increases and parenchymal fraction decreases monotonically, consistent with clinical imaging patterns associated with greater dementia severity. Ventricle width shows a similar overall trend but decreases at Moderate CDR. This category has very limited empirical support in OASIS, making the corresponding intervention particularly dependent on model extrapolation. With observation noise fixed, these trends support the anatomical responsiveness in Figure 3b, but do not establish clinical causal effects or individual disease trajectories.

5 Conclusion

High-dimensional counterfactual generation must translate low-dimensional interventions into complex observations while preserving unit-specific variation. We address this challenge by separating causal propagation from image synthesis: an additive latent SCM propagates interventions entirely through low-dimensional structural states, while a separate conditional generator synthesizes the high-dimensional image. This factorization further distinguishes unit-specific semantic exogenous variation from observation-level rendering noise.

Under stated representation, structural, invariance, and overlap assumptions, the framework identifies the counterfactual observation law and separates structural from generative error. Morpho-MNIST experiments demonstrate recovery of simulated structural mechanisms and graph-consistent counterfactuals, while OASIS results show anatomically ordered, visually subject-preserving MRI changes under a working causal graph. More broadly, the framework suggests that high-dimensional counterfactual generation need not require high-dimensional causal propagation: causal reasoning can be performed in a compact structural space, leaving the generative model to handle synthesis.

AI Use Statement

In preparing this manuscript, we used generative AI tools to assist with language editing and LaTeX formatting. We also used generative AI tools to assist with implementing portions of the evaluation code so that they matched standard evaluation protocols. We did not use generative AI tools to generate research ideas, design experiments, analyze data, interpret results, produce figures

or numerical results, or make scientific claims. The authors reviewed and tested the AI-assisted evaluation code against the relevant evaluation protocols, and reviewed and revised all AI-assisted text for accuracy, originality, and consistency with the reported methods and results. We take responsibility for the final content of this work, including any text produced with the aid of generative AI.

Ethics Statement

This work introduces a methodological framework and does not involve new data collection or interventions with human participants. Morpho-MNIST is a simulation with a known SCM, whereas the OASIS study is a real-world methodological showcase based on a working graph. Accordingly, the OASIS results should not be interpreted as identified clinical causal effects, individual disease trajectories, or diagnostic recommendations. The cohort has uneven support across demographic and clinical strata, and counterfactuals in sparsely supported regions are especially sensitive to modeling assumptions and extrapolation. Any clinical use would require independent validation, careful assessment of subgroup performance and potential bias, and appropriate governance for sensitive medical data.

Reproducibility Statement

The manuscript specifies the model factorization, estimation and counterfactual-inference procedures, identification assumptions, and evaluation protocols. The appendix provides pseudocode, dataset and causal-graph details, additional quantitative results, and the architecture, optimization, and sampling configurations for both Morpho-MNIST and OASIS experiments. These details are intended to enable reproduction of the reported experimental protocol and facilitate scrutiny of the assumptions underlying the counterfactual claims.

References

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025. URL https://www.jmlr.org/papers/v26/23-1605.html.

Ananya Anand and Namrata Anand. Fast brain volumetric segmentation from t1 mri scans. In *Advances in Computer Vision*, Advances in Intelligent Systems and Computing, pp. 402–415. Springer International Publishing, 2019. doi: 10.1007/978-3-030-17795-9_30.

Yana Blinkouskaya and Johannes Weickenmeier. Brain shape changes associated with cerebral atrophy in healthy aging and alzheimer's disease. *Frontiers in Mechanical Engineering*, 7:705653, 2021. doi: 10.3389/fmech.2021.705653.

Daniel C Castro, Jeremy Tan, Bernhard Kainz, Ender Konukoglu, and Ben Glocker. Morpho-mnist: quantitative assessment and diagnostics for representation learning. *Journal of Machine Learning Research*, 20(178):1–29, 2019.

Richard K. Crump, V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009. doi: 10.1093/biomet/asn055.

Fabio De Sousa Ribeiro, Tian Xia, Miguel Monteiro, Nick Pawlowski, and Ben Glocker. High fidelity image counterfactuals with probabilistic causal models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7390–7425. PMLR, 2023. URL https://proceedings.mlr.press/v202/de-sousa-ribeiro23a.html.

Fabio De Sousa Ribeiro, Ainkaran Santhirasekaram, and Ben Glocker. Counterfactual identifiability via dynamic optimal transport. *Advances in Neural Information Processing Systems*, 38:147855–147903, 2026.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012. doi: 10.1109/MSP.2012.2211477.

Alessandro Fontanella, Grant Mair, Joanna Wardlaw, Emanuele Trucco, and Amos Storkey. Diffusion models for counterfactual generation and anomaly detection in brain images. *IEEE Transactions on Medical Imaging*, 44(9):3574–3585, 2025. doi: 10.1109/TMI.2024.3460391.

Julia Furtner and Daniela Prayer. Neuroimaging in dementia. *Wiener Medizinische Wochenschrift*, 171(11–12):274–281, 2021. doi: 10.1007/s10354-021-00825-x.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 2020. URL https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81 (396):945–960, 1986. doi: 10.1080/01621459.1986.10478354.

Gary King and Langche Zeng. The dangers of extreme counterfactuals. *Political Analysis*, 14(2): 131–159, 2006. doi: 10.1093/pan/mpj004.

Murat Kocaoglu, Christopher Snyder, Alexandros G. Dimakis, and Sriram Vishwanath. CausalGAN: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJE-4xW0W.

Aneesh Komanduri, Chen Zhao, Feng Chen, and Xintao Wu. Causal diffusion autoencoders: Toward counterfactual generation via diffusion probabilistic models. In *ECAI 2024: 27th European Conference on Artificial Intelligence, 19–24 October 2024, Santiago de Compostela, Spain–Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)*, pp. 2516–2523. SAGE Publications Pvt. Ltd 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 2024.

Binxu Li, Wei Peng, Mingjie Li, Ehsan Adeli, and Kilian M. Pohl. Integrating anatomical priors into a causal diffusion model. *IEEE Transactions on Medical Imaging*, 45(7):3934–3947, 2026. doi: 10.1109/TMI.2026.3688515.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Christos Louizos, Uri Shalit, Joris M. Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/7223-causal-effect-inference-with-deep-latent-variable-models.

Daniel S. Marcus, Anthony F. Fotenos, John G. Csernansky, John C. Morris, and Randy L. Buckner. Open access series of imaging studies (OASIS): Longitudinal MRI data in nondemented and demented older adults. *Journal of Cognitive Neuroscience*, 22(12):2677–2684, 2010. doi: 10.1162/jocn.2009.21407.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0987b8b338d6c90bbedd8631bc499221-Abstract.html.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995. doi: 10.1093/biomet/82.4.669.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009. doi: 10.1017/CBO9780511803161.

Judea Pearl and Dana Mackenzie. The book of why. 2018.

Rajat R Rasal, Avinash Kori, Fabio De Sousa Ribeiro, Tian Xia, and Ben Glocker. Diffusion counterfactual generation with semantic abduction. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 51201–51228. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/rasal25a.html.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.

Pedro Sanchez and Sotirios A. Tsaftaris. Diffusion Causal Models for Counterfactual Estimation. *arXiv*, 2022. doi: 10.48550/arXiv.2202.10166.

Pedro Sanchez, Antanas Kascenas, Xiao Liu, Alison Q O'Neil, and Sotirios A Tsaftaris. What is healthy? generative counterfactual diffusion for lesion localization. In *MICCAI workshop on deep generative models*, pp. 34–44. Springer, 2022.

Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=BXewfAYMmJw.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semimarkovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conference*, AAAI-06/IAAI-06, pp. 1219–1226, 2006.

Yaakov Stern. Cognitive reserve in ageing and Alzheimer's disease. *The Lancet Neurology*, 11(11): 1006–1012, 2012. doi: 10.1016/S1474-4422(12)70191-6.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=CD9Snc73AW.

Yulun Wu, Louis McConnell, and Claudia Iriondo. Counterfactual generative modeling with variational causal inference. In *International Conference on Learning Representations*, volume 2025, pp. 86921–86955, 2025.

Tian Xia, Fabio De Sousa Ribeiro, Rajat R. Rasal, Avinash Kori, Raghav Mehta, and Ben Glocker. Factored classifier-free guidance. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2506.14399.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. CausalVAE: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9593–9602, 2021. doi: 10.1109/CVPR46437.2021.00947.

Yousef Yeganeh, Azade Farshad, Ioannis Charisiadis, Marta Hasny, Martin Hartenberger, Björn Ommer, Nassir Navab, and Ehsan Adeli. Latent drifting in diffusion models for counterfactual medical image synthesis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7685–7695, 2025. doi: 10.1109/cvpr52734.2025.00720.

A Reproducibility Details

This section specifies the complete estimation pipeline, the conditional flow construction, and the experimental settings needed to reproduce the reported results.

A.1 Estimation and Counterfactual Inference

Algorithms 1 and 2 give the complete procedures for sequentially estimating the generator and latent SCM, abducting factual residuals, propagating an intervention, and generating a counterfactual observation.

Algorithm 1: Sequential training

Data: DAG $\mathcal{G}$ and data $\{(\boldsymbol{x}_i, y_{i,1}, \ldots, y_{i,K-1})\}_{i=1}^N$.
Result: Trained flow model v_θ and latent SCM ϕ, including E_{K,ϕ_K} and $\{m_k\}_{k=1}^K$.
foreach *flow-matching training step* **do**
  Sample a minibatch $\{\boldsymbol{x}_{1,i}\}_{i=1}^b$, times $t_i \sim \mathrm{Uniform}(0,1)$, and $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$;
  Compute $\boldsymbol{x}_{t_i} = \alpha_{t_i}\boldsymbol{x}_{1,i} + \sigma_{t_i}\boldsymbol{\epsilon}_i$ and $v_{t_i}^\star = \dot{\alpha}_{t_i}\boldsymbol{x}_{1,i} + \dot{\sigma}_{t_i}\boldsymbol{\epsilon}_i$;
  Jointly update θ and ϕ_K by minimizing the minibatch empirical mean of $\mathcal{L}_{\mathrm{CFM}}$ in eq. (5);
Freeze $E_K = E_{K,\phi_K}$;
foreach *latent-SCM training step* **do**
  Sample a minibatch $\{(\boldsymbol{x}_i, y_{i,1}, \ldots, y_{i,K-1})\}_{i=1}^b$ from the data;
  Compute $\boldsymbol{h}_{i,j} = E_j(y_{i,j})$ for all i, j, with $y_{i,K} = \boldsymbol{x}_i$;
  Update the unfrozen components of ϕ, including $\{E_j\}_{j=1}^{K-1}$ and $\{m_k\}_{k=1}^K$, by minimizing the minibatch empirical mean of $\mathcal{L}_{\mathrm{SCM}}$ in eq. (8);
return v_θ *and* ϕ;

Algorithm 2: Counterfactual inference

Data: Trained flow model v_θ and latent SCM ϕ; factual $(\boldsymbol{x}, \boldsymbol{y})$; intervention $do(\boldsymbol{Y}_A = \boldsymbol{a})$.
Result: Counterfactual observation $\boldsymbol{X}^{\boldsymbol{a}}$.
Encode the factual observation as $\hat{\boldsymbol{h}}_k = E_k(y_k)$, with $y_K = \boldsymbol{x}$;
foreach *node* $k = 1, \ldots, K$ **do**
  Abduct $\hat{\boldsymbol{U}}_k = \hat{\boldsymbol{h}}_k - m_k(\mathcal{C}_k(\hat{\boldsymbol{h}}))$;
Set $\boldsymbol{h}_k^{\boldsymbol{a}} = E_k(a_k)$ for $k \in A$;
foreach *node* $k \notin A$ *in topological order* **do**
  Set $\boldsymbol{h}_k^{\boldsymbol{a}} = m_k(\mathcal{C}_k(\boldsymbol{h}^{\boldsymbol{a}})) + \hat{\boldsymbol{U}}_k$;
Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and solve Equation (6) conditional on $\boldsymbol{h}_K^{\boldsymbol{a}}$ to obtain $\boldsymbol{X}^{\boldsymbol{a}}$;
return $\boldsymbol{X}^{\boldsymbol{a}}$;

A.2 Conditional Flow Matching

Section 3.3 instantiates the conditional observation model $P(\boldsymbol{X} \mid \boldsymbol{h}_K)$ with conditional flow matching (Tong et al., 2024; Lipman et al., 2023). For $\boldsymbol{x}_1 \sim P(\boldsymbol{X})$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and $t \sim \mathrm{Uniform}(0,1)$, the interpolation path and its target velocity are

$$\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_1 + \sigma_t \boldsymbol{\epsilon}, \qquad v_t^\star = \dot{\alpha}_t \boldsymbol{x}_1 + \dot{\sigma}_t \boldsymbol{\epsilon}, \tag{15}$$

where $(\alpha_0, \sigma_0) = (0, 1)$ and $(\alpha_1, \sigma_1) = (1, 0)$. The objective in Equation (5) regresses the conditional velocity field along this path to estimate $P_\theta(\boldsymbol{X} \mid \boldsymbol{h}_K)$. For fixed-noise pairing, we integrate the ODE in Equation (6) backward from a factual image at fixed $\boldsymbol{h}_K$, recover its base sample, and reuse that sample under the counterfactual terminal state.

A.3 Implementation Details

Table 4 lists the architectures, optimization settings, and sampling configurations used for Morpho-MNIST and OASIS.

Table 4: Architectures, optimization settings, and sampling configurations used in the Morpho-MNIST and OASIS experiments.

	Morpho-MNIST	**OASIS**
data		
Samples (train/test)	50000/10000	8720/1280
Sample size	32×32 (padding with zero)	160×160
Causal factors	digit, thickness, intensity, slant	age, sex, education, nWBM, CDR
architecture		
Flow Model type	DiT with 6 layers	DiT with 12 layers
SCM Model type	1-layer transformer at each node	2-layer transformer at each node
Latent dim.	64	384
No. of parameters in the Flow model	2.29M	72.33M
No. of parameters in the SCM model	1.59M	28.40M
training		
Optimizer	AdamW	AdamW
Learning rate scheduler	Cosine with linear warmup	Cosine with linear warmup
Warmup ratio	0.1	0.1
Batch size	512	128
Learning rate	2×10^{-4}	2×10^{-4}
Steps (Stage 1 / 2)	50000/5000	150000/15000
sampling		
ODE solver	Discrete Euler	Discrete Euler
ODE steps	50	50
Noise coupling	$\alpha_t = t, \sigma_t = 1 - t$	$\alpha_t = t, \sigma_t = 1 - t$
Noise type in inference	Random sampled ϵ	Random sampled ϵ

B Supplementary Experimental Analysis

The two studies play different evidential roles: Morpho-MNIST tests recovery under a known SCM, whereas OASIS assesses behavior under a working graph in an observational imaging cohort.

B.1 Morpho-MNIST Simulation and Causal Structure

Morpho-MNIST augments MNIST digits with controllable morphological attributes, producing high-dimensional images from a known semantic SCM (Deng, 2012; Castro et al., 2019). Following Rasal et al. (2025), we use each source image to supply digit identity and unit-specific visual variation, then generate thickness, intensity, and slant explicitly. Each simulated sample contains four observed semantic variables—thickness t, intensity i, digit class d, and slant s—together with the transformed image x. The graph contains two root–child paths, $t \to i$ and $d \to s$, and the image x is the terminal node with all four semantic variables as parents. We use four interventions to test complementary structural behaviors. The root interventions $do(t)$ and $do(d)$ test downstream propagation to intensity and slant, respectively. The direct-child interventions $do(i)$ and $do(s)$

test targeted changes while requiring non-descendant semantic variables to remain invariant. For individual-level counterfactuals, we hold semantic exogenous residuals fixed in non-intervened assignments across worlds to preserve unit-specific variation. Separate observation-level generator noise captures image variation conditional on the counterfactual semantic state. Thickness is sampled independently and affects intensity through a nonlinear mechanism, whereas digit class is extracted from the source image and affects slant. The data-generating SCM is

$$
\begin{aligned}
d &:= f_d(U_x) = \texttt{CLS}(U_x), && U_x \sim \text{MNIST} \\
s &:= f_s(d, U_s) = -27 + d \cdot 6 + 3 \cdot U_s, && U_s \sim \mathcal{N}(0,1) \\
t &:= f_t(U_t) = 0.5 + U_t, && U_t \sim \Gamma(10,5) \\
i &:= f_i(t, U_i) = 191 \cdot \sigma(0.5 \cdot U_i + 2t - 5), && U_i \sim \mathcal{N}(0,1) \\
x &:= f_x(i, t, d, s, U_x) = \text{SetMorphology}(i, t, s \mid X), && U_x \sim \text{MNIST}.
\end{aligned} \tag{16}
$$

The final assignment renders the image from the four semantic variables while retaining source-image variation. Because both the mechanisms and exogenous variables are known, this simulation permits direct evaluation of structural recovery and counterfactual soundness.

B.2 OASIS Cohort and Covariate Support

We use axial MRI slices and associated age, sex, education, normalized whole-brain volume (nWBM), and Clinical Dementia Rating (CDR) from the longitudinal OASIS cohort (Marcus et al., 2010). Figure 4 summarizes the marginal distributions of the variables used by the model. The cohort is concentrated among older participants, contains more female than male volumes, and is dominated by CDR values of 0 and 0.5, with few volumes at CDR 2. Consequently, some intervention targets and joint covariate configurations have limited support, and counterfactuals in those regions depend more heavily on extrapolation.

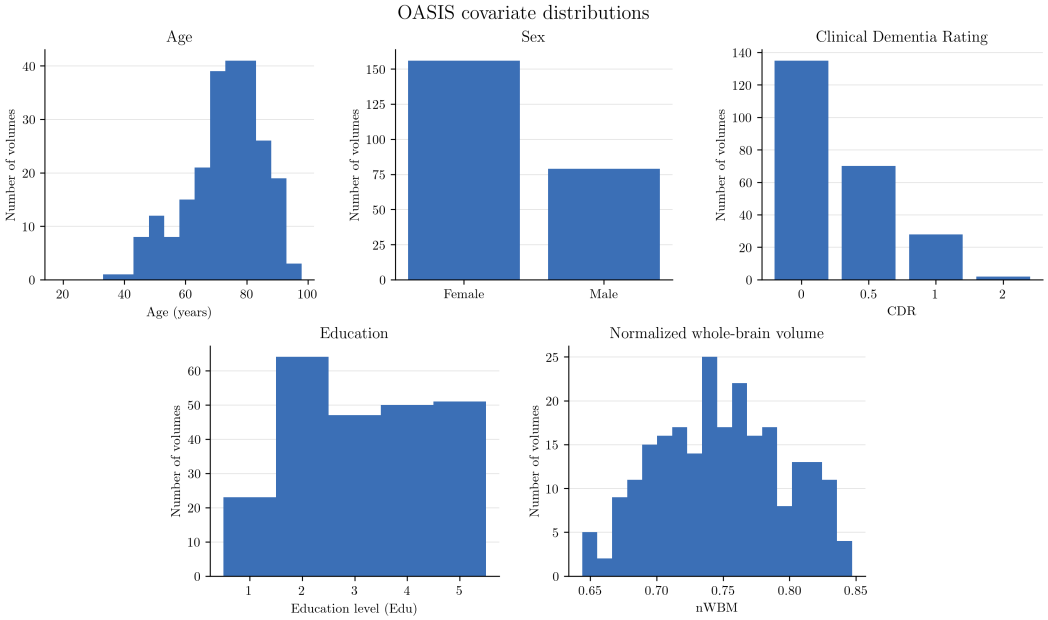


Figure 4: Marginal distributions of age, sex, Clinical Dementia Rating (CDR), education, and normalized whole-brain volume (nWBM) in the OASIS cohort used for evaluation. Coverage is uneven across demographic and clinical strata, particularly for sex and CDR, which limits support for some counterfactual targets.

B.3 OASIS Anatomical Validation

We segment each axial brain image with the pretrained, fine-tuned DeepLab model released by Anand & Anand (2019) at `https://github.com/nanand2/deeplab_mri_seg`. From each mask,

Table 5: Anatomical measurements for real NonDemented reference images and same-noise generated images across dementia severity levels. Values are reported as mean ± standard deviation.

Cohort	Ventricular Area Ratio ↑	Ventricle Width Ratio ↑	Parenchymal Fraction ↓
Real NonDemented reference	0.062 ± 0.006	0.180 ± 0.051	0.938 ± 0.006
Generated VeryMild	0.072 ± 0.006	0.372 ± 0.133	0.928 ± 0.006
Generated Mild	0.110 ± 0.008	0.632 ± 0.088	0.889 ± 0.008
Generated Moderate	0.120 ± 0.007	0.526 ± 0.096	0.879 ± 0.007

we compute the ventricular area ratio, ventricle width ratio, and parenchymal fraction reported in Table 5. These measurements test whether generated changes follow coarse anatomical patterns associated with the intervention; they do not validate the working graph or establish clinical causality.

B.4 Counterfactual-Soundness Metrics

For Table 1, *effectiveness* measures whether a generated image attains the semantic values implied by the intervention, using MAPE for continuous attributes and accuracy for digit class. For a continuous attribute, let a_i^{cf} denote the target counterfactual value for sample i, and let $\hat{a}_i^{\mathrm{cf}}$ denote the value estimated from the generated counterfactual image. We compute the mean absolute percentage error as

$$\mathrm{MAPE}(a) = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{\hat{a}_i^{\mathrm{cf}} - a_i^{\mathrm{cf}}}{a_i^{\mathrm{cf}}} \right|, \tag{17}$$

where multiplying by 100% expresses the value as a percentage. *Reversibility* intervenes and then restores the factual value, reporting the image-space L_1 distance from the recovered image to the factual image. *Null-intervention compatibility* reports the same distance when the intervention leaves the factual assignment unchanged. Together, the metrics distinguish semantic effectiveness from preservation of the factual unit.

B.5 Additional Morpho-MNIST Results

Table 6 complements the main results with interventions on thickness and intensity.

Table 6: Soundness of Morpho-MNIST image counterfactuals under $do(t)$ and $do(i)$ using DSCMs that model the SCM in fig. 1a. Effectiveness for digit class (d) is measured using accuracy (Acc) from a pre-trained classifier and mean absolute percentage error (MAPE) for slant (s), thickness (t), and intensity (i). Counterfactuals are normalised to $[0, 1]$ to measure composition (Comp.) and reversibility (Rev.). All MAPEs are scaled by $\times 10^{-1}$, all L_1 metrics are scaled by $\times 10^{-2}$, and Acc for digit class remains unscaled.

	Thickness Intervention $do(t)$					**Intensity Intervention** $do(i)$				
	Effectiveness				Rev.	Effectiveness				Rev.
Mechanism	MAPE (t) ↓	MAPE (i) ↓	MAPE (s) ↓	Acc (d) ↑	L_1 ↓	MAPE (t) ↓	MAPE (i) ↓	MAPE (s) ↓	Acc (d) ↑	L_1 ↓
VAE (2020)	4.48	6.76	3.88	97.75	4.26	8.06	7.55	4.83	97.85	4.31
HVAE (2023)	3.05	0.675	1.82	95.61	0.678	3.92	0.471	1.60	94.92	0.580
VCI (2025)	3.08	0.626	0.913	92.97	2.10	6.99	2.61	3.97	82.52	3.75
DiffSA (2025):										
$\{\omega = 1.5,\ p_\varnothing = 0.1\}$	2.36	1.29	2.30	98.54	1.86	3.38	1.40	2.12	96.39	1.70
$\{\omega = 3,\ p_\varnothing = 0.1\}$	2.16	1.01	1.56	99.51	3.52	3.41	1.45	2.11	97.75	3.27
$\{\omega = 4.5,\ p_\varnothing = 0.1\}$	2.24	0.76	2.73	99.61	4.72	4.21	1.55	2.56	98.24	4.43
Ours	**1.01**	0.83	2.53*	99.00	<u>1.09</u>	**1.52**	<u>1.02</u>	2.61*	98.12	<u>1.09</u>

** Note:* For some test examples, the denominator in MAPE of slant is close to zero, which can inflate the reported values. The corresponding Median APE is 0.86 for $do(t)$ and 0.89 for $do(i)$.

We separately assess structural recovery by comparing ground-truth thickness–intensity latent states with predictions from the learned $t \to i$ mechanism. Table 7 reports Fréchet distance and RBF–

Table 7: Recovery of the Morpho-MNIST thickness–intensity mechanism in latent space. Lower is better for Fréchet distance, RBF–MMD, and normalized direction error; higher is better for cross-covariance cosine similarity.

Scope	Metric	Value
Thickness	Fréchet distance	0.0238
	RBF–MMD	0.0574
Intensity	Fréchet distance	0.0143
	RBF–MMD	0.0432
Joint	Fréchet distance	0.0279
	RBF–MMD	0.0464
	Cross-covariance cosine similarity	0.999998
	Normalized direction error	0.001824

MMD for the marginal and joint distributions, together with cross-covariance cosine similarity and normalized direction error. The near-unit cross-covariance cosine similarity (0.999998) and small normalized direction error (0.001824) indicate that the fitted mechanism recovers the direction of thickness–intensity dependence in latent space. The nonzero marginal and joint discrepancies show that this agreement is not exact, so the result supports structural alignment rather than perfect distribution recovery.

Direct counterfactual propagation recovery. Because the Morpho-MNIST SCM is known, we also evaluate the counterfactual descendant state directly, before passing the propagated state to the image generator. For each test instance, we intervene with $do(t = t')$, preserve its abducted intensity residual, and compare the predicted descendant $\hat{i}^{\mathrm{CF}}$ with the ground-truth i^{CF} obtained by evaluating the known mechanism in Equation (16) under the same intervention and exogenous realization. For the baseline without residual preservation, we propagate the conditional mean under the same intervention instead of retaining the abducted intensity residual. We report root mean squared error (RMSE) and its scale-normalized version

$$\mathrm{NRMSE} = \frac{\mathrm{RMSE}\left(\hat{i}^{\mathrm{CF}}, i^{\mathrm{CF}}\right)}{\mathrm{SD}(i^{\mathrm{CF}})}. \tag{18}$$

Table 2 reports both models; the full model's NRMSE of 0.114 corresponds to 11.4% of the empirical standard deviation of the true counterfactual intensity. This pre-generation comparison isolates structural propagation error and provides a direct diagnostic of the structural component that contributes to the first term in the counterfactual error decomposition of Proposition 1. Together with the image-level counterfactual results in Tables 1 and 6, this experiment follows the full error pathway in Proposition 1: an intervention first incurs structural propagation error in the counterfactual descendant states, the conditional generator then introduces its own approximation error when mapping the propagated terminal state to the observation space, and their composition determines the final counterfactual image error. In the bound, these contributions correspond respectively to the propagated latent discrepancy $L_G W_1(\widehat{\mu}_a, \mu_a)$ and the conditional-generation discrepancy ε_G. The descendant-state RMSE diagnoses the former pathway upstream of image synthesis, while the image-level effectiveness, composition, and reversibility metrics evaluate the resulting counterfactual after both components have been composed. This alignment connects the experimental evidence to the structural-propagation–generation–final-error chain formalized by the proposition; separate estimation of every term in the bound remains outside the scope of this experiment.

C Geometric and Theoretical Details

This section makes precise the fixed-noise geometry, expands the identification assumptions, and provides the proofs omitted from the main text.

C.1 Fixed-Noise Geometry and Generator Regularity

For a fixed realization $\boldsymbol{\epsilon}$, the generator maps terminal latent states to

$$\mathcal{M}_{\boldsymbol{\epsilon}} = \{G(\boldsymbol{h}_K, \boldsymbol{\epsilon}) : \boldsymbol{h}_K \in \mathcal{H}\} \subseteq \mathbb{R}^D. \tag{19}$$

Equation (19) therefore collects the images reachable by varying only the terminal latent state. When $G(\cdot, \boldsymbol{\epsilon})$ is an embedding, $\mathcal{M}_{\boldsymbol{\epsilon}}$ has intrinsic dimension at most m even though its elements are D-dimensional. The full generated support is the closure of $\bigcup_{\boldsymbol{\epsilon}\in\mathrm{supp}(\boldsymbol{\epsilon})} \mathcal{M}_{\boldsymbol{\epsilon}}$ and need not itself be low-dimensional. Causal changes therefore move within a fixed-noise slice, whereas $\boldsymbol{\epsilon}$ indexes observation-level variation across slices. The method does not estimate a manifold explicitly; it learns this geometry implicitly through the encoders and conditional generator. We formalize the regularity needed for this geometry and for the conditional inversion used in fixed-noise paired counterfactuals.

Assumption 2 (Latent support and conditional generator regularity)**.** *The space $\mathcal{H}$ is compact. For every $\boldsymbol{\epsilon} \in \mathcal{E}$, $G(\cdot, \boldsymbol{\epsilon})$ is a continuously differentiable, L_G-Lipschitz embedding. For every fixed $\boldsymbol{h}_K \in \mathcal{H}$, the conditional flow map $G_{\boldsymbol{h}_K} : \boldsymbol{\epsilon} \mapsto G(\boldsymbol{h}_K, \boldsymbol{\epsilon})$ is invertible on its image.*

Assumption 2 supports fixed-noise geometry and conditional inversion, not a joint inverse from $\boldsymbol{X}$ to $(\boldsymbol{h}_K, \boldsymbol{\epsilon})$.

C.2 Identification Assumptions and Supporting Results

The combined assumption in Assumption 1 contains four distinct requirements. Structural validity requires that the latent DAG includes the relevant common causes of the modeled semantic nodes and that the factual residual realization is preserved for every non-intervened node. Representation consistency places encoder outputs on the structural scale; it does not require uniqueness of that representation and is not established by the flow-matching objective alone. Observation invariance is a regime-invariance condition conditional on the terminal latent state, rather than an assertion that the observed image itself is invariant. Overlap limits claims to parent configurations and terminal latent states supported by the observational distribution. Together they are sufficient for Theorem 1: structural validity and representation consistency make the factual residual realization recoverable, observation invariance transports the terminal latent state to the observation law, and overlap restricts the claim to supported intervention targets.

For an intervention on nodes $A \subseteq \{1, \ldots, K-1\}$ with values $\boldsymbol{a}$, define the population counterfactual recursion

$$\boldsymbol{h}_k^a = \begin{cases} E_k(a_k), & k \in A, \\ m_k\left(\mathcal{C}_k(\boldsymbol{h}^a)\right) + \boldsymbol{U}_k, & k \notin A, \end{cases} \tag{20}$$

evaluated in topological order, where $\boldsymbol{U}_k = \boldsymbol{h}_k - m_k(\mathcal{C}_k(\boldsymbol{h}))$.

Proposition 2 (Residual abduction)**.** *Under Assumption 1, the factual representation identifies*

$$\boldsymbol{U}_k = \boldsymbol{h}_k - m_k(\mathcal{C}_k(\boldsymbol{h})), \qquad k = 1, \ldots, K, \tag{21}$$

and Equation (20) *yields a unique latent counterfactual state.*

The per-sample structural loss used in Theorem 2 is

$$L_q(\boldsymbol{h}) = \frac{1}{K} \sum_{k=1}^{K} \mathcal{D}_k(\boldsymbol{h}_k, m_{q,k}(\mathcal{C}_k(\boldsymbol{h}))), \tag{22}$$

which matches the training objective in Equation (8) and is assumed to lie in $[0, 1]$. Because the bound holds simultaneously over all prefix programs, it controls the whole candidate family rather than a single fitted mechanism, and it yields a rate of order $\sqrt{\ell(q)/N}$ for a program of code length $\ell(q)$.

Proposition 3 (Output-dimension scaling for coordinatewise mechanism classes)**.** *Suppose each scalar response is independently selected from a prefix-coded class with descriptions of at most c bits. Then a terminal latent mechanism has budget $C_h = mc + O(m)$, whereas a coordinatewise D-dimensional observation-space mechanism has $C_X = Dc + O(D)$, giving generalization terms $O(\sqrt{mc/N})$ and $O(\sqrt{Dc/N})$, respectively.*

The following plug-in result complements the finite-sample error decomposition in the main text by giving an asymptotic consistency statement.

Proposition 4 (Plug-in consistency)**.** *Suppose the fitted encoders and structural mechanisms converge uniformly to their population counterparts, the fitted conditional flow law converges uniformly in Wasserstein distance to $P_{X|h_K}$, the structural functions are Lipschitz, and $P_{X|h_K}$ is continuous in $\boldsymbol{h}_K$ under W_1. Then, for any fixed supported intervention,*

$$W_1\Big(P_{\theta,X|h_K}(\cdot \mid \widehat{\boldsymbol{h}}_K^a), P(X^a \in \cdot \mid \boldsymbol{h} = \boldsymbol{h}^{\mathrm{obs}})\Big) \xrightarrow{p} 0. \tag{23}$$

C.3 Proofs

Proof of Theorem 2. The per-sample loss is $L_q(\boldsymbol{h}) = \frac{1}{K}\sum_{k=1}^{K} \mathcal{D}_k(\boldsymbol{h}_k, m_{q,k}(\mathcal{C}_k(\boldsymbol{h})))$ and lies in $[0,1]$ by assumption. For a fixed prefix program q, Hoeffding's inequality gives

$$P\left(R(q) - \widehat{R}_N(q) > \sqrt{\frac{\ell(q)\log 2 + \log(1/\delta)}{2N}}\right) \le \delta 2^{-\ell(q)}.$$

Because the programs form a prefix code, Kraft's inequality gives $\sum_q 2^{-\ell(q)} \le 1$. A union bound over all programs therefore bounds the probability that any candidate violates Equation (14) by δ. □

Proof of Proposition 3. Under the coordinatewise construction, a vector-valued mechanism concatenates one scalar-function description per output coordinate, together with a fixed delimiter or coordinate index. The terminal latent mechanism therefore has code length at most $mc + O(m)$, whereas the direct observation-space mechanism has code length at most $Dc + O(D)$. Substituting these budgets into Equation (14) gives the claimed orders. The comparison applies only to the matched coordinatewise classes and does not compare approximation error or architectures with parameter sharing. □

Proof of Proposition 2. For every node k, structural validity gives $\boldsymbol{h}_k = m_k(\mathcal{C}_k(\boldsymbol{h})) + \boldsymbol{U}_k$. Semantic consistency makes the factual node states observable on the structural scale, so rearranging the assignment gives Equation (21). Because $\mathcal{G}$ is acyclic, its nodes admit a topological ordering. After replacing assignments in A with their encoded intervention values, every non-intervened node is a deterministic function of previously evaluated parent states and its fixed factual residual. Induction over the topological order gives existence and uniqueness of all $\boldsymbol{h}_k^a$. □

Proof of Theorem 1. By Proposition 2, the factual latent state identifies the residual realization and hence the counterfactual terminal state $\boldsymbol{h}_K^a$. Structural validity makes this recursion the action and prediction step of the modified SCM. Representation consistency and observation invariance imply that the observation assignment remains unchanged across regimes conditional on its latent input. Therefore, for every measurable B,

$$P(X^a \in B \mid \boldsymbol{h} = \boldsymbol{h}^{\mathrm{obs}}) = P_{X|h_K}(B \mid \boldsymbol{h}_K^a).$$

Overlap ensures that both the structural functions along the recursion and the conditional observation law at $\boldsymbol{h}_K^a$ are determined by the observational distribution. □

Proof of Proposition 1. Let π be an optimal coupling of $\widehat{\mu}_a$ and μ_a, and let $(\widehat{\boldsymbol{h}}_K^a, \boldsymbol{h}_K^a) \sim \pi$. Draw one common $\boldsymbol{\epsilon}$ from its population distribution and use it for both counterfactuals. By the triangle inequality,

$$\begin{aligned}\mathbb{E}\|\widehat{\boldsymbol{X}}^a - \boldsymbol{X}^a\|_2 &\le \mathbb{E}\|\widehat{G}(\widehat{\boldsymbol{h}}_K^a, \boldsymbol{\epsilon}) - G(\widehat{\boldsymbol{h}}_K^a, \boldsymbol{\epsilon})\|_2 \\ &\quad + \mathbb{E}\|G(\widehat{\boldsymbol{h}}_K^a, \boldsymbol{\epsilon}) - G(\boldsymbol{h}_K^a, \boldsymbol{\epsilon})\|_2 \\ &\le \varepsilon_G + L_G W_1(\widehat{\mu}_a, \mu_a).\end{aligned}$$

Since the Wasserstein distance is the infimum of this expected transportation cost over all observation level couplings, Equation (13) follows. □

Proof of Proposition 4. Let r denote the maximum length of a directed path ending at the terminal node. Uniform convergence of the encoders and structural regressions gives consistency of the abducted residuals. For an intervened root, consistency of its fitted counterfactual state follows from consistency of $E_k(a_k)$; for a non-intervened root, it follows from consistency of its learned constant and residual. Suppose consistency holds for all nodes at depth at most j. Uniform convergence and Lipschitz continuity of each mechanism imply consistency for every child at depth $j+1$. Induction to depth r yields $\widehat{\boldsymbol{h}}_K^a \xrightarrow{p} \boldsymbol{h}_K^a$.

By the triangle inequality,

$$
\begin{aligned}
&W_1\Big(P_{\theta,X|h_K}(\cdot \mid \widehat{\boldsymbol{h}}_K^a), P_{X|h_K}(\cdot \mid \boldsymbol{h}_K^a)\Big) \\
&\quad \le W_1\Big(P_{\theta,X|h_K}(\cdot \mid \widehat{\boldsymbol{h}}_K^a), P_{X|h_K}(\cdot \mid \widehat{\boldsymbol{h}}_K^a)\Big) + W_1\Big(P_{X|h_K}(\cdot \mid \widehat{\boldsymbol{h}}_K^a), P_{X|h_K}(\cdot \mid \boldsymbol{h}_K^a)\Big).
\end{aligned}
$$

The first term converges to zero by uniform consistency of the conditional flow law. The second converges to zero by continuity of the population conditional law in $\boldsymbol{h}_K$ and consistency of $\widehat{\boldsymbol{h}}_K^a$. The identification result in Theorem 1 completes the proof. □

D Extended Discussion of the Structural–Generative Factorization

The high-dimensional counterfactual methods reviewed in Section 2 combine causal modeling and deep generation in different ways. Their relationship to our framework can be clarified along four dimensions.

Location of causal propagation. Existing methods may propagate interventions through observed structural assignments, learned latent representations, or a generative architecture that jointly models causal variables and images. In our framework, propagation is completed in the low-dimensional latent SCM, including the mechanism producing $\boldsymbol{h}_K$, before a separate generator synthesizes the observation.

Abduction of unit-specific variation. Depending on the model, factual variation may be represented by inferred exogenous variables, an encoded latent state, or the source image, and not all approaches preserve node-wise factual information across counterfactual worlds. Our additive latent SCM abducts residuals $\boldsymbol{U}_k$ and reuses them in every non-intervened structural assignment, while fixed-noise pairing preserves the factual generator noise $\boldsymbol{\epsilon}$ when required.

Structural variation versus rendering noise. In integrated generative models, causally meaningful heterogeneity and image-level stochasticity may share a latent or noise representation, or their distinction may remain implicit. We instead assign them separate roles: $\boldsymbol{U}_k$ captures unit-specific semantic variation in the latent SCM, whereas $\boldsymbol{\epsilon}$ captures observation-level variation conditional on $\boldsymbol{h}_K$.

Factorization of the terminal mechanism. The presence of a latent representation does not itself separate the high-dimensional outcome mechanism from causal propagation. Our framework places an explicit boundary at $\boldsymbol{h}_K$: the latent SCM learns parent-to-terminal propagation, while a replaceable conditional model learns $P(\boldsymbol{X} \mid \boldsymbol{h}_K)$ solely for observation synthesis. Thus, our contribution is not a new image generator, but an explicit structural–generative factorization that confines causal propagation to comparably low-dimensional node states.

E Limitations and Future Directions

The framework relies on a learned representation that preserves intervention-relevant semantics and makes causal mechanisms simpler than direct observation-space modeling. Because the conditional-generation objective learns the terminal representation, these properties are assumptions rather than guarantees of the training procedure. The theoretical results are therefore conditional on representation consistency and do not establish representation recovery or uniqueness. The additive latent SCM also

assumes that structural exogenous variation enters additively on the learned latent scale; nonlinear or non-additive mechanisms would require richer abduction procedures. Counterfactual identification further requires a valid graph, regime-invariant observation mechanisms, and overlap. Violations of these conditions can produce misleading counterfactuals, and interventions outside observed support remain model-based extrapolations. Likewise, reusing conditionally inverted generator noise defines a model-dependent individual pairing, not one identified from observational data alone.

The OASIS study has a modest effective sample size after preprocessing and uneven covariate coverage. The sex distribution is imbalanced, and CDR is concentrated at 0 and 0.5, with few observations at CDR 2. Joint covariate strata can be sparser than these marginal distributions suggest, making rare intervention targets especially dependent on extrapolation. Moreover, the cross-sectional observational design and working graph do not identify a clinical causal model. Generated MRIs should therefore be interpreted as a methodological demonstration, not as estimates of individual disease progression or treatment effects. Future work should target identifiable intervention-relevant representations, extend abduction beyond additive mechanisms, quantify sensitivity to graph misspecification and limited overlap, and evaluate larger longitudinal, interventional, or multimodal cohorts.