A CAUSAL BOOTSTRAP FOR DIVERSE IMAGE GENERATION UNDER UNOBSERVED CONFOUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep generative models reproduce the observational distribution underlying the training data, which, in turn, inherit any spurious associations it contains. A common source is an *unobserved confounder* that simultaneously affects both a treatment attribute the user wants to control/edit (e.g., age, sex) and an observed outcome (e.g., an image) expected to vary in response. Existing causal inference methods resolve the resulting ambiguity by imposing structural assumptions strong enough to single out one interventional distribution. However, in practical domains like image generation, such assumptions are rarely warranted, and the interventional distribution is generally *non-identifiable*: distinct causal models can agree on the observed data yet imply different causal predictions. This paper proposes CAUVADE (Causal Variational Deep Embedding), a novel family of causal generative models in which the unobserved confounder collapses into a discrete latent cluster with bounded support without loss of generality. We prove that this family can approximate both observational and interventional image distributions with arbitrary accuracy in a canonical latent causal diagram. Building on this model representation, we introduce a novel causal bootstrap that varies the strength of confounder entropy to generate a diverse family of causal explanations, each defining a distinct interventional image distribution. Such a procedure produces distinct interventional images while maintaining a good fit to the observed data.

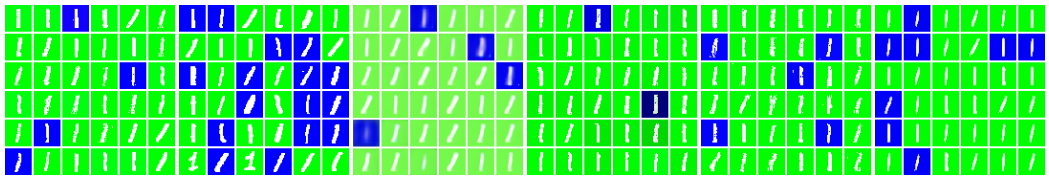
1 INTRODUCTION

Deep generative models, including variational autoencoders (VAEs), generative adversarial networks (GANs), and diffusion models, synthesize images by learning the observational distribution of a training corpus (Kingma & Welling, 2014; Goodfellow et al., 2014; Ho et al., 2020). Such models generalize well when the training corpus faithfully represents the target environment, but they can also reproduce and amplify its biases. Confounding bias is one such failure: when an unobserved confounder simultaneously affects both an attribute X (e.g., age, smile) and image I, the learned conditional $P(I \mid X{=}x)$ can differ from the interventional distribution $P(I \mid \mathrm{do}(X = x))$ obtained by externally setting X to x. Generating potential images under such interventions from confounded observational data is the task we call *deconfounded image generation*.

Causal inference provides a framework for identifying interventional distributions from confounded observational data under explicit assumptions. Causal image models such as CausalGAN, CausalVAE, and DeepSCM encode relationships among semantic factors using structural causal models (SCMs) (Kocaoglu et al., 2018; Yang et al., 2021; Pawlowski et al., 2020); CausalVAE, for instance, assumes causal sufficiency, a linear SCM with independent Gaussian noises, and semantic labels. Such assumptions are difficult to satisfy or even specify in image domains, where the relevant factors may be unobserved and their causal relationships uncertain. Absent sufficient identifying assumptions or data, the interventional distribution is generally non-identifiable: multiple SCMs can produce the same observations yet predict different interventional responses (Pearl, 2000).

The problem of inferring non-identifiable causal effects has been studied within the framework of *partial identification*, which characterizes the range of causal predictions compatible with observations and maintained assumptions. Early work established nonparametric treatment-effect bounds and sharp bounds under instrumental variables (Manski, 1990; Balke & Pearl, 1997). For finite observed domains, discretizing exogenous variables into finitely many response types yields canonical

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.



(a) Observational (b) Interventional (c) VAE (d) DDPM (e) GAN (f) ANCM

Figure 1: Samples from (a) confounded and (b) unconfounded Color-MNIST; and generated by conditional (c) VAE, (d) DDPM, (e) GAN, and (f) ANCM.

SCMs that preserve causal distributions and reduce sharp-bound computation to polynomial optimization (Zhang et al., 2022); neural approaches instead optimize causal queries under observational constraints (Xia et al., 2021; 2023). These discrete-domain methods do not directly provide a practical solution for images: enumerating response types over image-valued variables is prohibitive, and bounding a causal statistic does not itself produce diverse image generators.

More recently, the Augmented Neural Causal Model (ANCM) (Pan & Bareinboim, 2024) models images with hidden confounding while acknowledging non-identifiability, requiring fully labeled discrete semantic factors and a supplied causal diagram. Fitting an ANCM selects one causal explanation consistent with its constraints rather than enumerating the compatible ones, and repeated optimization from different random initializations yields very similar interventional predictions (Fig. 2). Random restarts thus do not adequately explore causal ambiguity to learn diverse interventional image distributions compatible with the observed data, particularly when some image attributes are unlabeled and their causal relationships are not fully specified. The following example illustrates these challenges.

Example 1 (Confounded Color-MNIST). Consider images with binary digit $X \in \{0, 1\}$ and background color $Y \in \{0, 1\}$, encoding green and blue, respectively. An unobserved binary variable U influences both the digit and the color, creating a confounded training distribution (Fig. 1a). The interventional reference breaks the dependence of X on U while preserving the color mechanism (Fig. 1b); it need not make digit and color independent. We train conditional VAE, DDPM, GAN, and ANCM baselines on the confounded data. Their samples appear in Figs. 1c to 1f, and Fig. 2 compares their implied color probabilities with the reference $P(Y{=}1 \mid \mathrm{do}(X{=}1))$. The conventional generators retain observational bias. ANCM is closer to the reference, but its estimates across random starts concentrate near one point inside the Manski bounds (Manski, 1990). These bounds admit a wider range of causal predictions compatible with observations: fitting the observed data alone does not determine which causal prediction is correct. ■

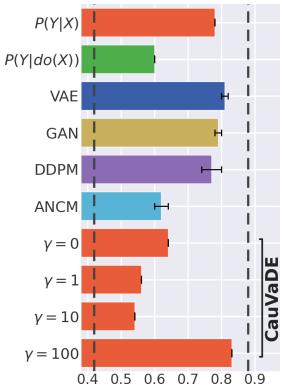


Figure 2: Estimated $P(Y{=}1 \mid \mathrm{do}(X{=}1))$. Dashed lines: Manski bounds from observations (Manski, 1990).

This paper addresses these challenges with a causal bootstrap that generates diverse candidate interventional image distributions under unobserved confounding. We propose a canonical causal diagram for deconfounded image generation, comprising a treatment attribute, latent pre-treatment covariates, and post-treatment image attributes; only the treatment attribute requires full labels, while the remaining attributes may be partially labeled or unlabeled. Our contributions are twofold. *(1) Causal Variational Deep Embedding (*CauVaDE*).* We introduce a causal generative framework whose canonical representation collapses exogenous confounding into a finite-support discrete variable, with independent exogenous noises carrying continuous variation; under standard regularity conditions, this model class approximates the observational and interventional image distributions of compatible augmented SCMs. *(2) Causal bootstrap.* We then gradually vary a cluster-posterior entropy weight to produce diverse candidates, and prove that under local conditions varying the confounder entropy yields distinct interventional image distributions while preserving observational admissibility within a fixed tolerance, though without guaranteeing exhaustive coverage or sharp bounds. We evaluate on Color-MNIST, CelebA, and MIMIC-CXR-JPG (Liu et al., 2015; Johnson et al., 2024): the first two show diverse predictions within the evaluated Manski bounds, and the last shows improved FID against an unconfounded reference. Proofs and implementation details appear in Appendices A–E.

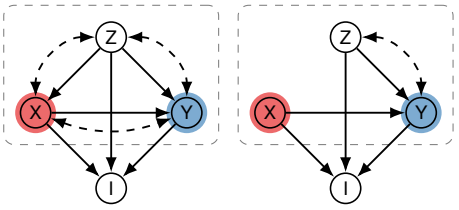
1.1 PRELIMINARIES

We use capital letters to denote variables (X), small letters for their values (x), and Ω_X for their domains. For a set $\boldsymbol{X}$, $|\boldsymbol{X}|$ denotes its cardinality. The probability distribution over variables $\boldsymbol{X}$ is denoted by $P(\boldsymbol{X})$, and we write $P(\boldsymbol{x})$ as shorthand for $P(\boldsymbol{X} = \boldsymbol{x})$. The basic semantical framework of our analysis rests on *structural causal models* (SCMs) (Pearl, 2000; Bareinboim et al., 2022). A SCM is a triple $\mathcal{M} = \langle \boldsymbol{V}, \boldsymbol{U}, \mathcal{F} \rangle$ describing how a set of observed variables $\boldsymbol{V} = \{X, Y, Z\}$ is generated. Throughout, X denotes a *treatment*, Y an *outcome*, and Z a *covariate*. The set $\boldsymbol{U}$ collects unobserved exogenous variables, drawn from an exogenous distribution $P(\boldsymbol{U})$. The structural equations $\mathcal{F} = \{f_X, f_Y, f_Z\}$ specify how each observed variable is generated: $Z \leftarrow f_Z(\boldsymbol{U})$, $X \leftarrow f_X(Z, \boldsymbol{U})$, and $Y \leftarrow f_Y(X, Z, \boldsymbol{U})$. Sampling $\boldsymbol{U} \sim P(\boldsymbol{U})$ and applying these equations induces the *observational distribution* $P(X, Y, Z) \triangleq P(X, Y, Z; \mathcal{M})$.

Each SCM $\mathcal{M}$ is also associated with a causal diagram $\mathcal{G}$ (e.g., Fig. 3), where nodes represent endogenous $\boldsymbol{V}$, and arrows represent the arguments PA_V, U_V of each structural function f_V. Exogenous variables $\boldsymbol{U}$ are not explicitly shown; a bi-directed arrow $V_i \leftrightarrow V_j$ indicates an unobserved confounder (UC), $U_{i,j} \in \boldsymbol{U}$, affecting V_i and V_j. An *intervention* $\mathrm{do}(X{=}x)$ models externally setting X to a constant x, overriding whatever value f_X would have produced. Formally, $\mathrm{do}(X{=}x)$ replaces f_X in $\mathcal{M}$ with the constant assignment $X \leftarrow x$, yielding a submodel $\mathcal{M}_x$. Sampling $\boldsymbol{U} \sim P(\boldsymbol{U})$ in $\mathcal{M}_x$ induces the *interventional distribution* $P(Y \mid \mathrm{do}(x)) \triangleq P(Y_x; \mathcal{M})$, where $Y_x(\boldsymbol{u}) \triangleq Y_{\mathcal{M}_x}(\boldsymbol{u})$ is the potential outcome of Y when X is set to x. For convenience, we also use the subscript notation $P_x(Y) = P(Y \mid \mathrm{do}(x))$ interchangeably for interventional distributions.

2 CANONICAL CAUSAL MODELS FOR IMAGE GENERATION

The augmented structural causal model (ASCM) of Pan & Bareinboim (2024) provides a framework for causal image generation with a pre-specified causal diagram and fully labeled, discrete semantic factors. In practical domains, detailed causal relationships may be unavailable, labels incomplete, and attributes continuous. We address these challenges in two steps: a canonical diagram obtained by grouping semantic variables relative to the treatment, followed by a canonical representation that places discrete structure on exogenous confounding.



(a) $\mathcal{M}$ (b) $\mathcal{M}_x$

Figure 3: (a) Canonical diagram with treatment X, pre-treatment covariates Z, post-treatment attributes Y, and image I. (b) Submodel induced by an intervention $\mathrm{do}(X{=}x)$.

Rather than specify every causal relationship among individual image attributes, we organize the semantic variables of an acyclic ASCM around the attribute X to be manipulated, grouping its descendants into a post-treatment block Y and all other semantic variables into a pre-treatment block Z. Here, pre-treatment means unaffected by interventions on X, so Z may include both causes of X and other non-descendants. Both blocks may be multivariate and latent. For example, when X is a digit, Z can encode the writer's identity and Y a background color generated in response to the digit. For any acyclic causal semantic model with treatment X, composing causal mechanisms within each block yields $Z \leftarrow f_Z(\boldsymbol{U})$, $X \leftarrow f_X(Z, \boldsymbol{U})$, and $Y \leftarrow f_Y(X, Z, \boldsymbol{U})$, where $\boldsymbol{U}$ collects the original exogenous variables. This reduction gives the *canonical diagram* in Fig. 3a while preserving the observational image distribution and interventions on X (Sec. D.1).

Following the ASCM framework of Pan & Bareinboim (2024), we augment the grouped causal semantic model with an image variable I and mechanism $I \leftarrow f_I(X, Y, Z, \mathcal{E}_I)$, where the rendering noise $\mathcal{E}_I$ is independent of $\boldsymbol{U}$ and captures visual features not encoded in the causal semantics. The covariates Z are endogenous causal attributes, distinct from the exogenous variables $\boldsymbol{U}$ that can confound the blocks. Only X requires full labels; the components of Y and Z may be partially labeled or unlabeled. For the representation result below, we assume:

- **(A1) Compact observed domain.** The image domain Ω_I and treatment domain Ω_X are compact.
- **(A2) Continuous counterfactual outcomes.** For every exogenous realization u, the map $x \mapsto Y_x(u) = f_Y(x, f_Z(u), u)$ is continuous on Ω_X.
- **(A3) Continuous mixing function.** f_I is continuous in its endogenous inputs (X, Y, Z).

(A1) models finite-resolution images with bounded pixel values and treatments in closed, bounded ranges; (A2) requires outcomes to vary continuously with treatment for each fixed exogenous realization; (A3) ensures that small changes in semantic inputs produce small changes in images for each fixed rendering-noise realization. Stronger continuity matters for partial identification, as causal bounds on continuous domains can be vacuous without Lipschitz regularization (Tan et al., 2024).

Sampling the exogenous variables and propagating through this model gives the observational image distribution $P(I) \triangleq P(I; \mathcal{M})$. Applying $\mathrm{do}(X{=}x)$ replaces the treatment mechanism by $X \leftarrow x$, giving the submodel in Fig. 3b and the interventional image distribution $P_x(I) \triangleq P(I_x; \mathcal{M}) = P\big(f_I(x, Y_x, Z_x, \mathcal{E}_I)\big)$, where $Z_x = Z$. Under confounding, this distribution can differ from the observational $P(I)$ and conditional $P(I \mid X{=}x)$. Thus, even with a common diagram, fitting the observed pairs (X, I) need not determine the intervention response. Our target is the set of candidate ASCMs compatible with $P(X, I)$ and the canonical diagram, each inducing a possibly distinct $P_x(I)$.

2.1 Canonical Representation with Discrete Confounding

The second challenge is to represent this model class when causal attributes are incompletely labeled and need not have finite discrete domains. We leave unlabeled Y and Z latent, without claiming to identify their values from images alone. Their domains and causal mechanisms can remain continuous. The unknown exogenous structure still makes the compatible class infinite-dimensional: $\boldsymbol{U}$ can have unknown dimension, support, and joint distribution. We therefore introduce a representation with a finite-support shared confounder and independent noises carrying continuous variation.

Definition 2.1 (Canonical ASCM)**.** A *canonical ASCM* (CASCM) is an ASCM satisfying (A1)–(A3) in which: (i) the unobserved confounder is a single discrete variable $U \in \{1, \ldots, d\}$ drawn from $P(U)$; (ii) the semantic variables are generated by continuous structural functions $Z \leftarrow f_Z(U, \mathcal{E}_Z)$, $X \leftarrow f_X(Z, U, \mathcal{E}_X)$, $Y \leftarrow f_Y(X, Z, U, \mathcal{E}_Y)$ with noises $\mathcal{E}_Z, \mathcal{E}_X, \mathcal{E}_Y, \mathcal{E}_I, U$ mutually independent; (iii) images are generated by the mixing function $I \leftarrow f_I(X, Y, Z, \mathcal{E}_I)$ satisfying (A3).

The canonical diagram groups causal mechanisms; the canonical representation additionally replaces their exogenous domain by a discrete variable U. The following theorem justifies the latter step by approximation for continuous counterfactual outcomes and mixing functions.

Theorem 2.2 (Augmented Causal Approximation)**.** *Fix the causal diagram in Fig. 3a and assume (A1)–(A3), with the domain and measurability conventions above. Let $\mathcal{M}$ be a compatible ASCM with exogenous variables $\boldsymbol{U}$ taking values in a Borel subset of $\mathbb{R}^m$ for some finite m. All structural assignments, including interventions in Ω_X, are assumed well-defined in these domains. For every $\epsilon > 0$, there exists a CASCM $\mathcal{N}$ such that:*

$$W_1\big(P(I, X; \mathcal{M}),\, P(I, X; \mathcal{N})\big) < \epsilon, \qquad \int_{\Omega_X} W_1\big(P_x(I; \mathcal{M}),\, P_x(I; \mathcal{N})\big)\, dx < \epsilon, \tag{1}$$

where W_1 is the 1-Wasserstein distance and dx denotes Lebesgue measure for continuous treatments or counting measure for categorical treatments. That is, the class of CASCMs is dense *in this regular class of ASCMs compatible with the canonical diagram Fig. 3a under joint W_1 approximation of the observational and interventional distributions (Eq.* (1)*).*

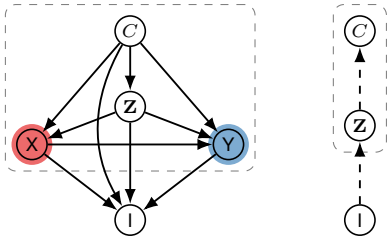
The finite-domain causal representation result of Zhang et al. (2022) motivates this construction, but does not itself establish the image-domain guarantee with latent causal models. Our proof (Sec. D.2) quantizes the joint distribution of natural treatment assignments, covariates, and entire response functions, preserving their dependence, and couples the rendering noise between models. Continuity of the responses and mixing function, together with compact image and treatment domains, yields observational and integrated interventional convergence.

3 Causal Variational Deep Embedding

Under its continuity and exogenous-domain assumptions, Thm. 2.2 motivates CASCMs as a representation of observational and interventional distributions, but this representation alone does not provide a tractable procedure for learning and inference. We therefore introduce Causal Variational Deep Embedding (CauVaDE), a tractable parametric instantiation based on Variational Deep Embedding

(VaDE) (Jiang et al., 2017) that incorporates the canonical causal relationships in Fig. 3a. We then develop a causal bootstrap to explore diverse causal explanations compatible with the observed data and their associated interventional predictions.

VaDE models the latent space of a variational autoencoder as a Gaussian mixture indexed by a discrete cluster $C \in \{1, \dots, K\}$. In CAUVADE, we identify C with the canonical confounder U in Def. 2.1, using its finite mixture structure to parameterize the confounder's influence on the generative mechanisms. The continuous latent collects the remaining exogenous structure of Def. 2.1 into a single bundle, $\mathbf{Z} = (Z, \mathcal{E}_Z, \mathcal{E}_X, \mathcal{E}_Y, \mathcal{E}_I)$, where Z is the pre-treatment attribute and $\mathcal{E}_Z, \mathcal{E}_X, \mathcal{E}_Y, \mathcal{E}_I$ are the independent noises driving the four CASCM mechanisms; structural functions f_Z, f_X, f_Y, f_I are then realized by neural decoders. CAUVADE specifies the generative process described as follows:



(a) Decoder (b) Encoder

Figure 4: CAUVADE: (a) decoder of Eq. (2), where the discrete C instantiates the canonical confounder of Def. 2.1; (b) encoder network $q_\phi(\mathbf{Z}, C \mid I) = q_\phi(\mathbf{Z} \mid I)\, q_\phi(C \mid \mathbf{Z})$.

$$C \sim P(C), \qquad \mathbf{Z} \mid C \sim \mathcal{N}(\mu_c, \sigma_c^2 \mathbf{I}), \qquad X \mid \mathbf{Z}, C \sim p_\theta(X \mid \mathbf{Z}, C),$$
$$Y \mid X, \mathbf{Z}, C \sim p_\theta(Y \mid X, \mathbf{Z}, C), \qquad I \mid X, Y, \mathbf{Z}, C \sim p_\theta(I \mid X, Y, \mathbf{Z}, C), \tag{2}$$

where $P(C)$ is a learned categorical prior, (μ_c, σ_c^2) are cluster-specific Gaussian parameters, and the conditionals over X, Y, and I are amortized by neural decoders θ. Fig. 4(a) shows the corresponding graph: C and $\mathbf{Z}$ jointly play the role of the unobserved exogenous structure of the CASCM, X and Y are the user-controlled and post-treatment semantic attributes, and I is the rendered image. In CAUVADE, Z and the noise variables have no separate latent representation; they are folded into $\mathbf{Z}$ and, along with Y, recovered as unlabeled semantic factors by the decoder.

Under intervention, the CAUVADE submodel induced by $\mathrm{do}(X{=}x)$ proceeds by drawing unobserved exogenous variable $(C, \mathbf{Z})$ from their joint prior $p(C)\, p_\theta(\mathbf{Z} \mid C)$, fixing $X = x$, and propagating through the decoders for Y and I. This yields the parametric interventional image distribution:

$$P_x^\theta(I) \;=\; \sum_{c=1}^{K} p(c) \int p_\theta(\mathbf{z} \mid c) \int p_\theta(y \mid x, \mathbf{z}, c)\, p_\theta(I \mid x, y, \mathbf{z}, c)\, dy\, d\mathbf{z}. \tag{3}$$

With the generative and interventional distributions specified, we turn to fitting CAUVADE from a confounded dataset $\mathcal{D} = \{(x_i, I_i)\}_{i=1}^N$ in which only X is labeled. Following VaDE, we approximate the interventional posterior $p_\theta(\mathbf{Z}, C \mid I, \mathrm{do}(X = x))$ by a factorized variational family $q_\phi(\mathbf{z}, c \mid I) = q_\phi(\mathbf{z} \mid I)\, q_\phi(c \mid \mathbf{z})$ (Fig. 4(b)). The encoder uses only I as an architectural choice; interventional generation uses the latent prior independently of this choice. A variational lower bound on $\log P_x^\theta(I)$, evaluated at the observed $x = X$ (Sec. B), yields the classic ELBO objective:

$$\mathcal{L}_{\mathrm{ELBO}}(\theta, \phi) \;=\; \underbrace{\mathbb{E}_{q_\phi}\big[\log p_\theta(I \mid X, \mathbf{Z}, C)\big]}_{\mathcal{L}_{\mathrm{rec}}} \;-\; \underbrace{D_{\mathrm{KL}}\big(q_\phi(\mathbf{Z}, C \mid I) \,\|\, p_\theta(\mathbf{Z} \mid C)\, P(C)\big)}_{\mathcal{L}_{\mathrm{KL}}}. \tag{4}$$

To incorporate labeled latent components, we anchor treatment X in the latent representation with a discriminative head, yielding the *treatment-consistency term*:[1] $\mathcal{L}_X(\phi, \psi) = \mathbb{E}_{q_\phi}\big[-\log p_\psi(X \mid \mathbf{Z}, C)\big]$, where p_ψ is an auxiliary classifier on top of the encoder. The term encourages $(C, \mathbf{Z})$ to predict X, but does not by itself force the image decoder to use X, or identify the generative treatment mechanism $p_\theta(X \mid \mathbf{Z}, C)$ with the auxiliary head p_ψ. Combining reconstruction, latent regularization, and treatment supervision with fixed weights $\alpha, \beta, \lambda > 0$ gives the *ELBO surrogate*

$$D(\theta, \phi, \psi) \;\coloneqq\; -\alpha\, \mathcal{L}_{\mathrm{rec}} \;+\; \beta\, \mathcal{L}_{\mathrm{KL}} \;+\; \lambda\, \mathcal{L}_X, \tag{5}$$

where the reconstruction and treatment terms measure fitness with the observed images and labeled components, while the KL term regularizes the marginalized latent representation. Optimizing Eq. (5) adapts ANCM's reconstruction and semantic supervision (Pan & Bareinboim, 2024) to partially labeled data. In our Color-MNIST experiments (Fig. 2), repeated ANCM training from different random starts yields similar interventional predictions. This empirical concentration motivates a more directed exploration of candidate models; it does not identify the underlying optimization mechanism, which calls for an alternative training and inference approach.

[1] Labeled subsets of Y or components of $\mathbf{Z}$, when available, are readily incorporated by adding analogous discriminative heads with no change to the rest of the objective.

3.1 A Causal Bootstrap for Diverse Interventional Image Generation

Our goal is to obtain a set of diverse predictions on interventional distributions among causal models compatible with the observed data. To achieve this goal, we introduce a training and inference procedure that varies the fitted model by regularizing its cluster-posterior entropy.

To make this argument more concrete, we introduce a cluster-entropy regularizer that controls uncertainty in the inferred canonical confounder. Formally, define the marginal posterior distribution over the cluster confounder by $q_\phi(c \mid I) = \int q_\phi(c \mid \mathbf{z}) q_\phi(\mathbf{z} \mid I)\, d\mathbf{z}$ and its mean entropy by $H(\phi) \coloneqq -\mathbb{E}_{P(I)} \sum_{c=1}^{K} q_\phi(c \mid I) \log q_\phi(c \mid I) \in [0, \log K]$, which measures uncertainty in the inferred cluster assignment for each image. Let $\eta = (\theta, \phi, \psi)$ collect the model and inference parameters. Adding this regularizer to the surrogate objective in Eq. (5) gives an augmented ELBO regularized by the entropy of the categorical confounder, defined as follows:

$$\mathcal{L}(\eta; \gamma) = -\alpha \mathcal{L}_{\text{rec}} + \beta \mathcal{L}_{\text{KL}} + \lambda \mathcal{L}_X - \gamma H = D(\eta) - \gamma H(\phi), \qquad \eta = (\theta, \phi, \psi), \tag{6}$$

with coefficient $\gamma \geq 0$; the negative signs reward reconstruction log-likelihood and posterior entropy. At $\gamma = 0$, this reduces to the baseline objective D. To generate interventional image distributions, we fix hyperparameters (α, β, λ) and fit the model as the entropy weight γ varies gradually. This yields a sequence of fitted parameters (θ, ϕ); each pair defines an interventional image distribution $P_x^\theta(I)$.

Our analysis provides intuition for how this training procedure could generate a diverse family of interventional images while being consistent with the observational data. Specifically, we consider three separate questions: how much varying the entropy weight increases the ELBO surrogate, whether it changes interventional predictions, and when the resulting models remain observationally admissible. We start at an unregularized baseline η_0 and impose two local assumptions. (A4) specifies the local smoothness and baseline curvature needed to follow fitted parameters as γ varies.

- **(A4) Stable baseline.** D, H are twice differentiable near η_0, with $\nabla_\eta D(\eta_0) = 0$, $\nabla_\eta^2 D(\eta_0) \succ 0$.

To describe the branch's initial direction and entropy response, set

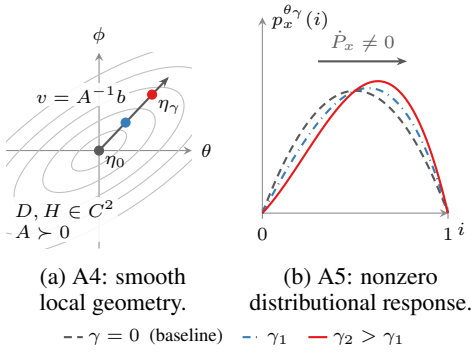
$$A = \nabla_\eta^2 D(\eta_0), \qquad b = \nabla_\eta H(\eta_0), \qquad v = A^{-1} b, \qquad \kappa_0 = b^\top A^{-1} b. \tag{7}$$

Here v is the initial parameter direction induced by entropy regularization: the local curvature A converts the entropy gradient b into a displacement of the fitted optimum. The quantity κ_0 measures the corresponding entropy response. (A4) supplies a stable local branch, but does not require $b \neq 0$; if $b = 0$, its initial direction vanishes. Even when parameters move, the generated distribution need not change. Write $\mathsf{P}_x(\eta) = P_x^\theta(I)$ and let $\|\cdot\|_{\text{KR}}$ denote the Kantorovich–Rubinstein norm on the bounded image domain, with $\|P - Q\|_{\text{KR}} = W_1(P, Q)$. (A5) requires a smooth, nondegenerate response of the interventional image distribution along the induced parameter direction.

- **(A5) Nondegenerate interventional response.** The map P_x is differentiable at η_0 in the KR norm, with $\dot{P}_x = \mathrm{d}\mathsf{P}_x(\eta_0)[A^{-1}b] \neq 0$ and $\rho_x = \|\dot{P}_x\|_{\text{KR}} > 0$.

Fig. 5(a) illustrates the local geometry required by (A4): the gradients and Hessians of both D and H exist and vary continuously near η_0. The center of the loss contours marks the stationary baseline, $\nabla D(\eta_0) = 0$; their positive curvature represents $A \succ 0$. Together, these conditions ensure that a small change in γ admits a nearby local optimum of $D - \gamma H$, illustrated by the colored points. Each point locally minimizes its own regularized objective. Panel (b) illustrates the requirement in (A5): along the induced parameter direction, the interventional distribution is differentiable and has a nonzero derivative. The matching density curves depict probability shifting toward larger values as γ increases.

We first quantify the cost of the observation surrogate along the local branch of the fitted model parameters as the γ weight varies.



(a) A4: smooth local geometry.

(b) A5: nonzero distributional response.

Figure 5: (A4)–(A5): a stationary baseline with positive curvature gives a smooth local branch (left), along which the interventional distribution has a nonzero derivative (right). Colors match parameters to distributions.

Lemma 3.1. *Under (A4), there exist $\bar{\gamma} > 0$ and a unique nearby C^1 branch η_γ of strict local minimizers of $D - \gamma H$ through η_0, for $0 \le \gamma < \bar{\gamma}$. Along this branch, $\eta_\gamma = \eta_0 + \gamma v + o(\gamma)$, and*

$$D(\eta_\gamma) - D(\eta_0) = \tfrac{1}{2}\kappa_0\gamma^2 + o(\gamma^2), \qquad H(\phi_\gamma) - H(\phi_0) = \kappa_0\gamma + o(\gamma). \tag{8}$$

Lem. 3.1 shows that, when the entropy gradient is nonzero, a small step in γ buys a first-order increase in entropy at only a second-order increase in the surrogate cost. The local relation between surrogate cost and achieved entropy characterizes the cost of moving through parameter space. This movement provides a potential route to exploring different interventional image distributions, with (A5) specifying a local condition under which such variation occurs.

Theorem 3.2. *Under (A4)–(A5), there exists $\rho_x > 0$ such that $W_1(P_x^{\theta_\gamma}, P_x^{\theta_0}) = \gamma\rho_x + o(\gamma)$. Thus the interventional image distribution changes at first order while D changes at second order, i.e.,*

$$\lim_{\gamma\downarrow 0} \frac{W_1(P_x^{\theta_\gamma}, P_x^{\theta_0})}{\gamma} = \rho_x > 0, \qquad \lim_{\gamma\downarrow 0} \frac{D(\eta_\gamma) - D(\eta_0)}{W_1(P_x^{\theta_\gamma}, P_x^{\theta_0})} = 0. \tag{9}$$

Thm. 3.2 establishes that varying the entropy weight γ can produce distinct interventional image distributions at a comparatively small observational surrogate cost: near the baseline, the observational cost increase divided by the interventional movement tends to zero. Both changes vanish as γ decreases, but the surrogate cost vanishes at a faster rate. The remaining question is observational admissibility. A small increase in the surrogate D does not ensure that a model is compatible with the observational data, because D is a training surrogate rather than a discrepancy between the full observational distributions. Let $\Delta_j(\theta)$, $j = 1, \ldots, J$, be a fixed finite collection of discrepancy measurements between $P^\theta(X, I)$ and the real data distribution $P(X, I)$, with observational constraints $\Delta_j(\theta) \le \varepsilon_j$. Acceptance refers to these finite checks, not equality of the full observational distributions. The following result establishes that there exists a sequence of entropy-weighted movements satisfying the finite observational checks while yielding distinct interventional image predictions.

Corollary 3.3 (Observationally admissible movement)**.** *Under the conditions of Thm. 3.2, suppose each Δ_j is continuous at θ_0 and the baseline satisfies $\Delta_j(\theta_0) < \varepsilon_j$. Then there exists $\delta > 0$ such that every $0 \le \gamma < \delta$ satisfies all observational constraints, and every $0 < \gamma < \delta$ yields $P_x^{\theta_\gamma} \ne P_x^{\theta_0}$.*

Cor. 3.3 shows that interventional image predictions can change while the same observational constraints remain satisfied. These analyses motivate training candidates across a sequence of entropy weights $\Gamma = \{\gamma_1, \ldots, \gamma_M\}$ and retaining those that satisfy common held-out constraints. Sec. C provides details of the training procedure. For each $\gamma \in \Gamma$, we minimize Eq. (6) with gradient descent, reparameterizing $\mathbf{Z}$ and analytically summing over C. To estimate H, average the cluster posterior over reparameterized draws of $\mathbf{Z}$ before taking its entropy. After unregularized-ELBO pretraining initializes $(P(C), \mu_c, \sigma_c^2)$, jointly optimize the encoder, decoders, and classifier. Each model generates interventional samples according to Eq. (3). Observational admissibility uses a joint image-treatment held-out measurement, such as maximum mean discrepancy (MMD).

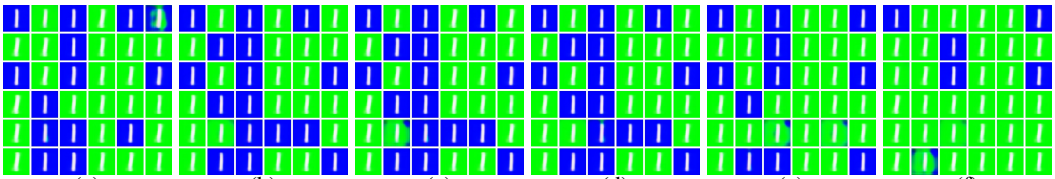
4 Experiments

We evaluate CauVaDE on confounded variants of Color-MNIST, CelebA (Liu et al., 2015), and MIMIC-CXR-JPG (Johnson et al., 2024; 2019; Goldberger et al., 2000). Each construction introduces an unobserved confounder U affecting both the treatment X and outcome Y. We train models on confounded images; unconfounded data serve as a ground truth for evaluation. We compare CauVaDE with a vanilla VAE and ANCM (Pan & Bareinboim, 2024); Table 1 also reports conditional GAN (CGAN) and diffusion baselines on Color-MNIST and CelebA.

The experiments assess two questions: whether varying the entropy weight γ produces diverse interventional predictions, and whether constructed causal explanations are compatible with observed data. For Color-MNIST and CelebA, we use the (X, Y) labels both to check interventional estimates of $P_x(Y = 1)$ against scalar Manski bounds (Manski, 1990) and to assess observational fit through Label RMSE between model-implied joint semantic probabilities $P(X, Y)$ and their empirical counterparts. These complementary checks assess compatibility at the causal-semantic variable level, while reconstruction negative log-likelihood (NLL) and mean squared error (MSE) assess image-level observational fit. We also examine generated samples and report FID against the MIMIC-CXR-JPG reference. Sec. E provides details of the evaluation metrics and training procedures.

Table 1: Reference observational and interventional probabilities, baseline estimates, and CAUVADE predictions on Confounded Color-MNIST and CelebA. Varying γ yields different estimates of $P_x(Y = 1)$ within the reported scalar Manski bounds (Manski, 1990).

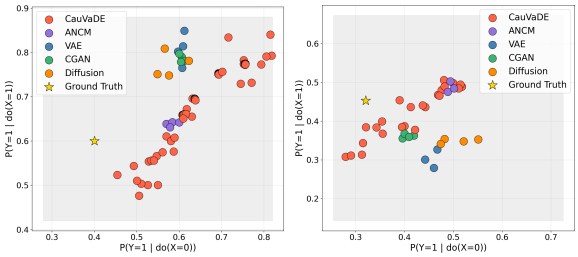
		$P(Y=1 \mid X)$	$P_x(Y=1)$	VAE	CGAN	Diffusion	ANCM	Causal Bounds		CAUVADE			
								Lower	Upper	$\gamma=0$	$\gamma=1$	$\gamma=10$	$\gamma=100$
C-MNIST	$X=0$	0.61	0.40	0.61 ± 0.01	0.61 ± 0.01	0.58 ± 0.03	0.58 ± 0.01	0.28	0.82	0.57	0.54	0.49	0.72
	$X=1$	0.78	0.60	0.81 ± 0.01	0.79 ± 0.01	0.77 ± 0.03	0.62 ± 0.02	0.42	0.88	0.61	0.56	0.54	0.83
CelebA	$X=0$	0.46	0.33	0.46 ± 0.01	0.41 ± 0.01	0.51 ± 0.03	0.49 ± 0.02	0.26	0.73	0.42	0.57	0.48	0.40
	$X=1$	0.31	0.46	0.30 ± 0.02	0.36 ± 0.01	0.35 ± 0.01	0.48 ± 0.02	0.14	0.67	0.44	0.51	0.48	0.39



(a) (b) (c) (d) (e) (f)

Figure 6: CAUVADE samples on Confounded Color-MNIST under $X = 1$ at (a) $\gamma = 1$, (b) $\gamma = 2$, (c) $\gamma = 10$, (d) $\gamma = 20$, (e) $\gamma = 50$, and (f) $\gamma = 100$. The digit–color association varies across entropy weights, illustrating changes in the fitted interventional predictions.

Confounded Color-MNIST. We construct Color-MNIST with digit labels $X \in \{0, 1\}$, background colors $Y \in \{\text{green}, \text{blue}\}$ (encoded as $0, 1$), and a binary confounder U affecting both. Figs. 1a to 1b contrast observational samples from $P(I \mid X = 1)$ with the interventional reference $P_{X=1}(I)$; construction details and additional samples appear in Sec. E. Across five seeds, VAE and ANCM yield concentrated estimates of $P_x(Y = 1)$ (Table 1). VAE produces nearly identical estimates for the two treatment values, whereas ANCM distinguishes them. Varying γ in CAUVADE yields a broader range of predictions than seed variation in either baseline, with all tabulated estimates remaining within the Manski bounds. Figs. 6 and 7a show the corresponding changes in predicted color frequencies and generated images. Across the reconstruction's entropy weights, per-pixel MSE ranges from 0.0303 to 0.0366, NLL from 0.1153 to 0.1362, and Label RMSE for $P(X, Y)$ from 0.0067 to 0.0519 (Table 4 and Sec. E.3.1). The small errors in the labeled distribution, together with stable reconstruction metrics, provide support for observational admissibility as interventional predictions vary.



(a) Color-MNIST (b) Confounded CelebA

Figure 7: Estimated $P_x(Y = 1)$ on Confounded Color-MNIST and CelebA. VAE and ANCM estimates concentrate across seeds, whereas CAUVADE predictions vary with γ. The shaded feasible region denotes the evaluated Manski's causal bounds (Manski, 1990).

Confounded CelebA. We resample images from CelebA (202,599 source images), assigning "Young" to X, "Heavy Makeup" to Y, and "Attractive" to the hidden confounder U. The resampling protocol is given in Sec. E.1.2. Figs. 8a to 8b compare the confounded data and unconfounded reference at $X = 1$. VAE and ANCM show limited variation across three seeds, while changing γ yields a range of CAUVADE predictions within the evaluated Manski bounds (Fig. 7b and Table 1). The tabulated predictions vary nonmonotonically with γ over the evaluated range. For $X = 1$, the tabulated estimates span 0.385–0.506, compared with the reference probability of 0.463, which lies within this range. Thus, increasing γ does not necessarily bring predictions closer to the reference. Samples at $\gamma \in \{1, 10\}$ visually resemble the unconfounded reference (Fig. 8), although this resemblance does not identify the causal mechanism. Reconstruction MSE and NLL remain stable across the entropy weights, while Label RMSE for $P(X, Y)$ ranges from 0.0055 to 0.0387 (Table 5). Together, these diagnostics support the observational admissibility as interventional predictions vary.



Figure 8: (a) Confounded CelebA samples $P(I \mid X = 1)$; (b) unconfounded reference $P(I \mid \mathrm{do}(X{=}1))$; (c) VAE; (d) ANCM; (e) CAUVADE at $\gamma = 0$; (f) CAUVADE at $\gamma = 1$; (g) CAUVADE at $\gamma = 10$; (h) CAUVADE at $\gamma = 100$. Blue circles mark images with $Y = 1$ (Heavy Makeup).

Confounded MIMIC-CXR-JPG. We construct a resampling experiment from MIMIC-CXR-JPG (Johnson et al., 2024; 2019; Goldberger et al., 2000) (377,110 source radiographs), using Pneumonia as X and subsequent outcomes as Y (e.g., Lung Opacity, unlabeled). The hidden confounder U is the patient's age and the radiograph's anteroposterior view. Fig. 9 shows confounded samples. This comparison tests image generation when treatment assignment no longer depends on the age–view grouping. FID summarizes image-distribution agreement rather than the accuracy of individual clinical attributes. For $\gamma \in \{1, 10, 100\}$, CAUVADE achieves FIDs of 276.64, 284.57, and 289.51, respectively. Averaged across these entropy weights, its FID is 283.57, compared with 354.97 for VAE and 407.73 for ANCM (Table 2), reductions of 71.40 and 124.16, respectively. All three settings yield lower FID than both baselines, indicating closer image-distribution agreement with the unconfounded reference. Alongside these changes, we compare observational fit with the $\gamma = 0$ baseline to check whether diverse causal predictions violate admissibility.

Table 2: FID against the unconfounded MIMIC-CXR-JPG reference. The CAUVADE entry averages over $\gamma \in \{1, 10, 100\}$. The lower the better.

	VAE	ANCM	CAUVADE
FID	354.97	407.73	**283.57**

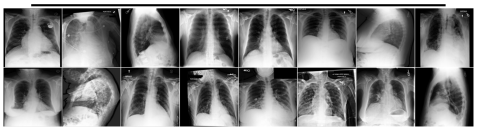


Figure 9: Confounded MIMIC-CXR-JPG samples: $X = 0$ (top row, no Pneumonia) and $X = 1$ (bottom row, Pneumonia).

5 CONCLUSION

Unobserved confounding means that the same observed images can support different predictions about the effects of an intervention. Our augmented causal approximation result (Thm. 2.2) provides a foundation for representing both observed and intervention-generated images using a discrete confounder. CAUVADE instantiates this canonical augmented causal representation as a Gaussian-mixture VAE. A causal bootstrap then explores different causal explanations and their resulting interventional image predictions by varying a training weight. Our analysis explains how the resulting predictions can change while the fit to the observed data remains nearly unchanged (Lem. 3.1 and Thm. 3.2). Experiments on Color-MNIST and CelebA show distinct predictions within the causal bounds derived from labeled semantics. On MIMIC-CXR-JPG, generated images more closely match the unconfounded ground truth images on average than those from the compared baselines. Determining the full range of effects consistent with the observed data remains an open question.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs, implement methods, clean and reformat dataset. We have not used generative AI tools to generate synthetic datasets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, support qualitative and thematic data analysis, interpret results, and assist with translation are not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures or images, create or edit software code, summarize or analyse existing literature, edit a research paper to improve readability, identify relevant literature. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by 2 authors. All LLM-suggested relevant work and references were verified and confirmed for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly available or credentialed-access datasets under their stated terms: CelebA (Liu et al., 2015), a synthetic Confounded Color-MNIST variant constructed from MNIST, and MIMIC-CXR-JPG (Johnson et al., 2024), accessed via PhysioNet under its credentialed-access data use agreement. MIMIC-CXR-JPG is a pre-existing, de-identified clinical dataset; this work does not involve new human-subjects research, crowdsourcing, or IRB review. CAUVADE is a generative modeling method for partial identification under hidden confounding; as with any conditional image generator, it could in principle be misused to produce synthetic media (e.g., counterfactual demographic or disease-label edits). We do not release pretrained models with this submission.

REPRODUCIBILITY STATEMENT

Full proofs of Thm. 2.2 (density theorem), Lem. 3.1, and Thm. 3.2 and Cor. 3.3 (the results on entropy-weight variation), together with (A1)–(A3), the exogenous-domain condition of Thm. 2.2, local assumptions (A4)–(A5), and additional slack conditions, are given in Secs. D.2 to D.5. Experimental details—dataset construction for Color-MNIST, CelebA, and MIMIC-CXR-JPG, architecture and training hyperparameters, optimizer settings, cluster count K, and the per-dataset γ-grid—are reported in Sec. E and Sec. E.2, and pseudocode for both per-γ training and training across entropy weights is given in Sec. C. Code is not released with this submission but will be made available upon acceptance; all datasets used are publicly available or accessible via credentialed access as described in Sec. E.

REFERENCES

Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997. doi: 10.1080/01621459.1997.10474074.

Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On Pearl's hierarchy and the foundations of causal inference. In Hector Geffner, Rina Dechter, and Joseph Y. Halpern (eds.), *Probabilistic and Causal Inference: The Works of Judea Pearl*, ACM Books, pp. 507–556. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3501714.3501743.

Guilherme Duarte, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association*, 119(547):1778–1793, 2024. doi: 10.1080/01621459.2023.2216909.

Constantine E. Frangakis and Donald B. Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002. doi: 10.1111/j.0006-341X.2002.00021.x.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23): e215–e220, June 2000. doi: 10.1161/01.CIR.101.23.e215. PMID: 10851218.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. URL https://arxiv.org/abs/1406.2661.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1414–1423. PMLR, 2017. URL https://proceedings.mlr.press/v70/hartford17a.html.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020. URL https://arxiv.org/abs/2006.11239.

Adrián Javaloy, Pablo Sánchez-Martín, and Isabel Valera. Causal normalizing flows: From theory to practice. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, pp. 58833–58864, 2023. doi: 10.52202/075280-2566.

Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1965–1972, 2017. doi: 10.24963/ijcai.2017/273.

Alistair Johnson, Matthew Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. MIMIC-CXR-JPG — chest radiographs with structured labels. *PhysioNet*, March 2024. doi: 10.13026/jsn5-t979. URL https://doi.org/10.13026/jsn5-t979. Version 2.1.0.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs, 2019. URL https://arxiv.org/abs/1901.07042.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

Ilyes Khemakhem, Ricardo Monti, Robert Leech, and Aapo Hyvärinen. Causal autoregressive flows. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3520–3528. PMLR, 2021. URL https://proceedings.mlr.press/v130/khemakhem21a.html.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations (ICLR)*, 2014. URL https://arxiv.org/abs/1312.6114.

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 (ICLR)*, 2018. URL https://openreview.net/forum?id=BJE-4xW0W.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

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 *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Charles F. Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80 (2):319–323, 1990. URL https://www.jstor.org/stable/2006592.

Charles F. Manski. Monotone treatment response. *Econometrica*, 65(6):1311–1334, 1997. doi: 10.2307/2171738.

Wang Miao, Zhi Geng, and Eric J. Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018. doi: 10.1093/biomet/asy038.

Miguel Monteiro, Fabio De Sousa Ribeiro, Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Measuring axiomatic soundness of counterfactual image models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=lZOUQQvwI3q.

Arash Nasr-Esfahany, Mohammad Alizadeh, and Devavrat Shah. Counterfactual identifiability of bijective causal models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25733–25754. PMLR, 2023. URL https://proceedings.mlr.press/v202/nasr-esfahany23a.html.

Kirtan Padh, Jakob Zeitler, David Watson, Matt Kusner, Ricardo Silva, and Niki Kilbertus. Stochastic causal programming for bounding treatment effects. In *Proceedings of the Second Conference on Causal Learning and Reasoning (CLeaR)*, volume 213 of *Proceedings of Machine Learning Research*, pp. 142–176. PMLR, 2023. URL https://proceedings.mlr.press/v213/padh23a.html.

Yushu Pan and Elias Bareinboim. Counterfactual image editing. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39087–39101. PMLR, 2024. URL https://proceedings.mlr.press/v235/pan24a.html.

Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 857–869, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0987b8b338d6c90bbedd8631bc499221-Abstract.html.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000. ISBN 978-0-521-77362-1.

Paul R. Rosenbaum. *Observational Studies*. Springer, New York, 2 edition, 2002. doi: 10.1007/978-1-4757-3692-2.

Pedro Sanchez and Sotirios A. Tsaftaris. Diffusion causal models for counterfactual estimation. In *Proceedings of the First Conference on Causal Learning and Reasoning (CLeaR)*, volume 177 of *Proceedings of Machine Learning Research*, pp. 647–668. PMLR, 2022. URL https://proceedings.mlr.press/v177/sanchez22a.html.

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. doi: 10.1109/JPROC.2021.3058954.

Jiyuan Tan, Jose Blanchet, and Vasilis Syrgkanis. Consistency of neural causal partial identification. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7f9220f90cc85b0da693643add6618e6-Abstract-Conference.html.

Eric J. Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal inference. *Statistical Science*, 39(3):375–390, 2024. doi: 10.1214/23-STS911.

Tyler J. VanderWeele and Peng Ding. Sensitivity analysis in observational research: Introducing the E-value. *Annals of Internal Medicine*, 167(4):268–274, 2017. doi: 10.7326/M16-2607.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009. ISBN 978-3-540-71049-3. doi: 10.1007/978-3-540-71050-9.

Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/5989add1703e4b0480f75e2390739f34-Abstract.html.

Kevin Xia, Yushu Pan, and Elias Bareinboim. Neural causal models for counterfactual identification and estimation. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=vouQcZS8KfW.

Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=sy4Kg_ZQmS7.

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 (CVPR)*, pp. 9593–9602, 2021. doi: 10.1109/CVPR46437.2021.00947.

Junzhe Zhang, Jin Tian, and Elias Bareinboim. Partial counterfactual identification from observational and experimental data. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26548–26558. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhang22ab.html.

A Related Work

Causal generative models and counterfactual evaluation. Causal generative models incorporate structural assumptions into image synthesis and editing. CausalGAN combines a generator over attributes, structured by a supplied causal graph, with conditional image generation (Kocaoglu et al., 2018). CausalVAE learns causal relations among semantic factors with label supervision and a linear causal layer driven by independent Gaussian exogenous variables (Yang et al., 2021). DeepSCM and Diff-SCM enable counterfactual image generation through inference of exogenous noise and diffusion-based mechanisms, respectively (Pawlowski et al., 2020; Sanchez & Tsaftaris, 2022). Causal autoregressive and normalizing flows exploit invertible mechanisms for causal inference (Khemakhem et al., 2021; Javaloy et al., 2023), while bijective causal models establish counterfactual identifiability for specific structures that include hidden confounding (Nasr-Esfahany et al., 2023). Complementing these modeling approaches, Monteiro et al. (2023) evaluate counterfactual image models through composition, reversibility, and effectiveness. Together, these works develop mechanisms and criteria for causal generation under distinct structural and identification assumptions. Our focus is on sampling a family of candidate interventional image distributions $P_x(I)$ under hidden confounding; this distributional target does not require identifying individual-level counterfactual images.

Causal representation learning. Causal representation learning seeks to recover high-level causal variables from observations (Schölkopf et al., 2021). Auxiliary observations can make latent-variable models identifiable under suitable conditional-prior and model assumptions (Khemakhem et al., 2020), whereas unsupervised disentanglement is generally impossible without inductive biases on the model and data (Locatello et al., 2019). CauVaDE uses latent representations to model incompletely labeled semantic attributes and generate candidate intervention responses. Its approximation result concerns observational and interventional image distributions and does not assert recovery of the true semantic coordinates.

Augmented neural causal models. Augmented neural causal models (ANCMs) extend neural causal modeling to images (Pan & Bareinboim, 2024), building on methods for causal identification and estimation with neural SCMs (Xia et al., 2021; 2023). ANCMs explicitly accommodate hidden confounding and non-identifiability, targeting counterfactual consistency with respect to selected semantic features. Their training procedure uses a supplied causal graph and labels for the modeled semantic factors, which have finite discrete domains; additional image factors may remain unobserved. This detailed semantic graph specifies how an edit to one attribute can affect others, supporting counterfactual image editing conditioned on a given source image. CauVaDE builds on this augmented SCM framework, allowing semantic variables with continuous domains and incomplete

labels for non-treatment factors. Our focus is the simpler task of deconfounded interventional generation: setting $X = x$ breaks its dependence on hidden confounders while preserving their effects on other variables. We target the population-level distribution $P_x(I)$, rather than an individual counterfactual edit conditioned on a source image. We group semantic variables relative to the treatment and approximate exogenous confounding with a finite-support latent variable under the regularity conditions of Thm. 2.2. The causal bootstrap then varies the cluster-posterior entropy weight to obtain multiple candidate interventional image distributions. Empirically, ANCM behaves similarly to CAUVADE with $\gamma = 0$, with both producing interventional estimates that concentrate tightly across random seeds (see the ablation study in Sec. E.5.1).

Causal inference under hidden confounding. Sensitivity analysis characterizes how causal conclusions depend on assumptions about unmeasured confounding (Rosenbaum, 2002; VanderWeele & Ding, 2017). Instrumental-variable methods exploit external sources of treatment variation (Hartford et al., 2017; Xu et al., 2021), while proxy-variable methods use measurements related to the hidden confounder (Tchetgen Tchetgen et al., 2024; Miao et al., 2018). Identification in these settings depends on assumptions about the auxiliary variables and causal mechanisms, including rank or completeness conditions where required. Our entropy sweep explores variation in fitted causal predictions without requiring instruments or proxies. The weight γ controls cluster-posterior uncertainty and has no calibrated interpretation as the strength of hidden confounding.

Partial identification. Partial identification characterizes the causal conclusions compatible with available data and assumptions. Foundational results include nonparametric treatment-effect bounds (Manski, 1990), sharp bounds under instrumental-variable restrictions (Balke & Pearl, 1997), and bounds under monotone treatment response (Manski, 1997). Principal stratification provides a complementary framework for defining effects within strata determined by potential post-treatment values (Frangakis & Rubin, 2002). For finite observed domains, Zhang et al. (2022) show that canonical SCMs with finite exogenous support preserve counterfactual distributions and reduce sharp-bound computation to polynomial optimization; Duarte et al. (2024) develop automated bounding procedures for discrete settings. Our contribution combines a representation guarantee for observational and interventional image distributions with a procedure for generating diverse candidate models. The representation theorem allows continuous semantic mechanisms under its stated regularity conditions, and the entropy sweep explores intervention responses while retaining candidates that satisfy observational checks. This procedure does not guarantee sharp bounds or exhaustive coverage of the compatible distributions.

Optimization-based causal bounds and image generation. Neural causal models minimize and maximize specified causal queries over models constrained by the data and causal graph (Xia et al., 2021; 2023). Stochastic causal programming similarly optimizes causal effects over distributions of response functions subject to data-fitting constraints, including settings with multivariate continuous treatments (Padh et al., 2023). These approaches do not directly solve our interventional image-generation task for two reasons. First, their optimization requires a specified target functional, whereas our target is a family of image distributions $P_x(I)$. There is no canonical scalar objective whose minimum and maximum characterize this entire family: choosing an image statistic, such as an attribute score, permits bounding its expectation but does not recover the diversity of compatible image distributions. Second, neural parameterizations and response-function models can yield nonconvex objectives and constraints, so local optimization can miss feasible extrema even in structured-data settings and return overly narrow estimated interventional ranges. Padh et al. (2023) explicitly acknowledge this risk, although they do not report consistent empirical evidence of becoming trapped in poor local optima. This limitation of local solvers should be distinguished from discrete polynomial approaches (Zhang et al., 2022; Duarte et al., 2024): global bounding procedures such as that of Duarte et al. (2024) provide certified bounds, but enumerating discrete response types for image-valued variables is prohibitive. CAUVADE instead varies a cluster-posterior entropy weight to explore candidate interventional image distributions without prescribing a scalar image query to extremize. Our goal is to generate diverse interventional image samples compatible with the observed data, rather than to recover exhaustive causal bounds; entropy regularization supports this exploration by varying the fitted models while retaining those that pass observational checks.

B DERIVATION OF THE ELBO

We derive Eq. (4) for the generative model in Eq. (2). As in Sec. 3, the data consists of observed pairs $\mathcal{D} = \{(x_i, I_i)\}_{i=1}^N$, with only X labeled. The latent vector $\mathbf{Z} = (Z, \mathcal{E}_Z, \mathcal{E}_X, \mathcal{E}_Y, \mathcal{E}_I)$ collects the pre-treatment attribute and the four independent exogenous noises, and the discrete cluster C represents the canonical confounder U. The encoder has the factorization

$$q_\phi(\mathbf{z}, c \mid I) = q_\phi(\mathbf{z} \mid I)\, q_\phi(c \mid \mathbf{z}).$$

We first specify the likelihood to which the variational bound is applied. The target is the interventional log-evidence $\log P_x^\theta(I)$, evaluated at the observed treatment value $x = X$. Since the post-treatment attribute Y is unobserved, we integrate it out and write

$$p_\theta(I \mid X, \mathbf{Z}, C) \;\triangleq\; \int p_\theta(Y \mid X, \mathbf{Z}, C)\, p_\theta(I \mid X, Y, \mathbf{Z}, C)\, dY. \tag{10}$$

Here the integral is part of the definition of the image likelihood. If instead one samples Y and averages the conditional log-likelihood, Jensen's inequality gives a further lower bound on the logarithm of this integral.

Derivation of Eq. (4). Let

$$r_\theta(\mathbf{z}, c) = p_\theta(\mathbf{z} \mid c)\, P(c)$$

denote the joint latent prior. Under $\mathrm{do}(X = x)$, the latent variables are drawn from this prior and the treatment is fixed to x. Thus, at $x = X$,

$$P_x^\theta(I) = \int \sum_{c=1}^{K} p_\theta(I \mid X, \mathbf{z}, c)\, r_\theta(\mathbf{z}, c)\, d\mathbf{z}.$$

Following the variational argument of Kingma & Welling (2014), we introduce an expectation under the encoder. With the density ratio taken on its support, this gives

$$\log P_x^\theta(I) \geq \log \mathbb{E}_{q_\phi(\mathbf{z}, c \mid I)} \left[\frac{p_\theta(I \mid X, \mathbf{z}, c)\, r_\theta(\mathbf{z}, c)}{q_\phi(\mathbf{z}, c \mid I)} \right].$$

This step is an equality when the encoder covers the support of the integrand. Since $\log$ is concave, Jensen's inequality yields

$$\log P_x^\theta(I) \geq \mathbb{E}_{q_\phi(\mathbf{z}, c \mid I)} \left[\log p_\theta(I \mid X, \mathbf{z}, c) + \log \frac{r_\theta(\mathbf{z}, c)}{q_\phi(\mathbf{z}, c \mid I)} \right].$$

By definition of the KL divergence,

$$\mathbb{E}_{q_\phi} \left[\log \frac{r_\theta(\mathbf{Z}, C)}{q_\phi(\mathbf{Z}, C \mid I)} \right] = -D_{\mathrm{KL}}\big(q_\phi(\mathbf{Z}, C \mid I) \,\|\, r_\theta(\mathbf{Z}, C)\big).$$

Therefore,

$$\log P_x^\theta(I) \geq \mathbb{E}_{q_\phi} \big[\log p_\theta(I \mid X, \mathbf{Z}, C) \big] - D_{\mathrm{KL}}\big(q_\phi(\mathbf{Z}, C \mid I) \,\|\, r_\theta(\mathbf{Z}, C)\big).$$

Substituting $r_\theta(\mathbf{z}, c) = p_\theta(\mathbf{z} \mid c)\, P(c)$ gives Eq. (4). The bound is attained when the encoder equals the posterior under the intervened model,

$$q_\phi(\mathbf{z}, c \mid I) = p_\theta(\mathbf{z}, c \mid I, \mathrm{do}(X = x)).$$

This proves the stated variational bound. □

We next expand the KL term using the encoder and prior factorizations. By definition,

$$\mathcal{L}_{\mathrm{KL}} = \mathbb{E}_{q_\phi(\mathbf{z} \mid I)} \left[\sum_{c=1}^{K} q_\phi(c \mid \mathbf{z}) \log \frac{q_\phi(\mathbf{z} \mid I)\, q_\phi(c \mid \mathbf{z})}{p_\theta(\mathbf{z} \mid c)\, P(c)} \right].$$

For each $(\mathbf{z}, c)$, the logarithm separates as

$$\log \frac{q_\phi(\mathbf{z} \mid I)\, q_\phi(c \mid \mathbf{z})}{p_\theta(\mathbf{z} \mid c)\, P(c)} = \log q_\phi(\mathbf{z} \mid I) - \log p_\theta(\mathbf{z} \mid c) + \log \frac{q_\phi(c \mid \mathbf{z})}{P(c)}.$$

The sum of the last term over c is the KL divergence between the cluster posterior and the categorical prior. Consequently,

$$\mathcal{L}_{\mathrm{KL}} = \mathbb{E}_{q_\phi(\mathbf{z}|I)}\left[\sum_{c=1}^{K} q_\phi(c \mid \mathbf{z})\big(\log q_\phi(\mathbf{z} \mid I) - \log p_\theta(\mathbf{z} \mid c)\big)\right] + \mathbb{E}_{q_\phi(\mathbf{z}|I)}\big[D_{\mathrm{KL}}\big(q_\phi(C \mid \mathbf{z}) \,\|\, P(C)\big)\big]. \tag{11}$$

The Gaussian encoder density $q_\phi(\mathbf{z} \mid I) = \mathcal{N}(\mu_\phi(I), \sigma_\phi^2(I))$ and the Gaussian prior density $p_\theta(\mathbf{z} \mid c) = \mathcal{N}(\mathbf{z}; \mu_c, \sigma_c^2)$ can both be evaluated in closed form. For a given $\mathbf{z}$, the categorical KL term is also a finite sum. Thus the sum over C can be evaluated analytically, while the remaining expectation over $\mathbf{Z}$ is estimated as described in Sec. C.

The encoder takes only I as input, whereas the posterior that makes the bound tight is conditioned on both I and the intervention. This architectural restriction can therefore loosen the bound. It does not change the intervention sampling procedure: generation draws from the latent prior, so allowing the encoder to depend on X would also be compatible with that procedure.

To distinguish this derivation from observational conditioning, consider the observational joint distribution. In that model, X is generated from its mechanism, giving

$$p_\theta(x, I) = \sum_c \int r_\theta(\mathbf{z}, c)\, p_\theta(x \mid \mathbf{z}, c)\, p_\theta(I \mid x, \mathbf{z}, c)\, d\mathbf{z}. \tag{12}$$

Conditioning on the observed treatment therefore weights the latent variables by $p_\theta(\mathbf{z}, c \mid X = x)$, whereas intervention leaves their prior $r_\theta(\mathbf{z}, c)$ unchanged. A variational bound on the observational joint log-likelihood would also contain the generative treatment log-likelihood. The auxiliary classifier loss $\mathcal{L}_X$ in Sec. 3 does not establish such an interpretation without a specified relationship to the treatment mechanism and objective weights. Accordingly, Eq. (6) remains a training surrogate, and observational consistency is assessed separately using the constraints associated with Cor. 3.3.

The two auxiliary terms have distinct roles. The treatment-consistency term encourages the latent variables to predict X, although this alone does not ensure that the image decoder uses X. The entropy term encourages uncertainty in the cluster posterior. To conclude that changing its weight changes interventional predictions, the analysis additionally uses the distributional response in (A5).

Finally, suppose X and Y are absent and the dataset consists only of images $\{I_i\}_{i=1}^N$. Then the reconstruction term reduces to $\mathbb{E}_{q_\phi}\big[\log p_\theta(I \mid \mathbf{Z}, C)\big]$, and the KL term retains Eq. (11) with the same Gaussian-mixture prior $r_\theta(\mathbf{z}, c) = P(c)\, p_\theta(\mathbf{z} \mid c)$. Hence Eq. (4) reduces to the VaDE objective (Jiang et al., 2017). In CauVaDE, the treatment additionally conditions the decoder, the post-treatment attribute is marginalized as in Eq. (10), and the auxiliary terms $\mathcal{L}_X$ and H encourage treatment information and generate candidates for observational acceptance.

C Training Algorithm

This appendix gives pseudocode for training CauVaDE at a single value of the entropy weight γ (Alg. 1) and for training across entropy weights to explore candidate interventional distributions (Alg. 2). The notation matches Sec. 3: $\mathcal{D} = \{(x_i, I_i)\}_{i=1}^N$ is the confounded training set, K is the number of mixture components, and (θ, ϕ, ψ) collect the decoder, encoder, and classifier-head parameters. As in Sec. 3, $\mathbf{Z}$ is the augmented continuous latent and C is the discrete cluster. The GMM parameters $\{P(c), \mu_c, \sigma_c^2\}_{c=1}^K$ are part of θ but are initialized separately by the VaDE-style pretraining phase (Jiang et al., 2017).

Two design choices keep training tractable. First, gradients with respect to ϕ flow through $\mathbf{Z}$ via the standard Gaussian reparametrization $\mathbf{z} = \mu_\phi(I) + \sigma_\phi(I) \odot \epsilon$ with $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. Second, we marginalize the discrete cluster C analytically rather than sampling, following VaDE (Jiang et al., 2017): the categorical posterior $q_\phi(c \mid \mathbf{z})$ is bounded (K values) and computable in closed form, so every loss term that involves C is written as a finite sum

$$\mathbb{E}_{q_\phi(C|\mathbf{z})}[g(C)] \;=\; \sum_{c=1}^{K} q_\phi(c \mid \mathbf{z})\, g(c),$$

Algorithm 1 CAUVADE training at fixed γ

Require: Dataset $\mathcal{D} = \{(x_i, I_i)\}_{i=1}^{N}$; cluster count K; weights $(\alpha, \beta, \lambda, \gamma)$; learning rate η_{lr}; number of epochs T; batch size B; pretrained GMM parameters $\{\mu_c^{(0)}, \sigma_c^{2,(0)}, P^{(0)}(c)\}_{c=1}^{K}$.
Ensure: Trained parameters (θ, ϕ, ψ).
Initialize encoder ϕ, decoders $\theta \supseteq \{\mu_c, \sigma_c^2, P(c)\}_{c=1}^{K}$, classifier head ψ; load pretrained GMM parameters into θ.
for epoch $= 1, \ldots, T$ **do**
  for minibatch $\mathcal{B} \subset \mathcal{D}$ of size B **do**
    Encode $(\mu_\phi(I), \sigma_\phi(I)) \leftarrow \text{Enc}_\phi(I)$ for each $I \in \mathcal{B}$.
    Sample $\mathbf{z} = \mu_\phi(I) + \sigma_\phi(I) \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. ▷ Gaussian reparam.
    Compute cluster posterior $q_\phi(c \mid \mathbf{z}) \in \Delta^{K-1}$ for $c = 1, \ldots, K$. ▷ Closed form.
    Compute per-cluster decoder log-likelihoods $\ell_c \leftarrow \log p_\theta(I \mid X, \mathbf{z}, c)$ for $c = 1, \ldots, K$. ▷ K decoder passes.
    Compute $\mathcal{L}_{\text{rec}}, \mathcal{L}_{\text{KL}}$ from Eq. (4) and $\mathcal{L}_X, H$ as defined in Sec. 3 and Sec. 3.1, respectively, on $\mathcal{B}$, with $\mathbb{E}_{q_\phi(C|\mathbf{z})}[\,\cdot\,] = \sum_c q_\phi(c \mid \mathbf{z})\,[\,\cdot\,]$. ▷ Finite sums over C; marginalize $\mathbf{Z}$ for H.
    $\mathcal{L} \leftarrow -\alpha\,\mathcal{L}_{\text{rec}} + \beta\,\mathcal{L}_{\text{KL}} + \lambda\,\mathcal{L}_X - \gamma\,H$. ▷ Eq. (6).
    $(\theta, \phi, \psi) \leftarrow (\theta, \phi, \psi) - \eta_{\text{lr}} \nabla_{(\theta,\phi,\psi)} \mathcal{L}$.
  end for
end for
return (θ, ϕ, ψ).

Algorithm 2 CAUVADE training across entropy weights

Require: Dataset $\mathcal{D}$; cluster count K; entropy-weight grid $\Gamma = \{\gamma_1, \ldots, \gamma_M\}$; weights (α, β, λ); target intervention values $\{x^{(1)}, \ldots, x^{(L)}\} \subset \mathcal{D}_X$; sample size S.
Ensure: Family of interventional sample sets $\{\mathcal{S}_x^{(m)}\}_{m=1,\ldots,M;\, x \in \{x^{(1)},\ldots,x^{(L)}\}}$.
Pretrain GMM parameters $\{\mu_c^{(0)}, \sigma_c^{2,(0)}, P^{(0)}(c)\}_{c=1}^{K}$ on $\mathcal{D}$ as described above.
for $m = 1, \ldots, M$ **do**
  $(\theta_m, \phi_m, \psi_m) \leftarrow \text{TRAINCAUVADE}(\mathcal{D}, K, \alpha, \beta, \lambda, \gamma_m, \{\mu_c^{(0)}, \sigma_c^{2,(0)}, P^{(0)}(c)\})$. ▷ Alg. 1.
  for $x \in \{x^{(1)}, \ldots, x^{(L)}\}$ **do**
    $\mathcal{S}_x^{(m)} \leftarrow \emptyset$.
    for $s = 1, \ldots, S$ **do**
      Sample $c \sim P_{\theta_m}(C)$.
      Sample $\mathbf{z} \sim p_{\theta_m}(\mathbf{Z} \mid c) = \mathcal{N}(\mu_c, \sigma_c^2 \mathbf{I})$.
      Sample $y \sim p_{\theta_m}(Y \mid x, \mathbf{z}, c)$.
      Sample $I \sim p_{\theta_m}(I \mid x, y, \mathbf{z}, c)$.
      $\mathcal{S}_x^{(m)} \leftarrow \mathcal{S}_x^{(m)} \cup \{I\}$.
    end for
  end for
end for
return $\{\mathcal{S}_x^{(m)}\}$.

and gradients propagate through the probabilities $q_\phi(c \mid \mathbf{z})$ directly. For the reconstruction, KL, and treatment terms, the reconstruction term becomes $\sum_c q_\phi(c \mid \mathbf{z}) \log p_\theta(I \mid X, \mathbf{z}, c)$, and the KL term decomposes into expressions involving $q_\phi(c \mid \mathbf{z})$, $P(c)$, and the cluster-conditional Gaussian parameters (μ_c, σ_c^2). For the entropy H defined in Sec. 3.1, average $q_\phi(c \mid \mathbf{z})$ over reparametrized latent draws before taking the categorical entropy; averaging the entropies instead gives a different regularizer. Hard categorical samples of C are used at generation time.

Pretraining initializes (μ_c, σ_c^2) by fitting a Gaussian mixture model to the embeddings produced by a vanilla VAE pretrained for a few epochs on $\mathcal{D}$ with $-\mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}}$ alone (i.e., $\lambda = \gamma = 0$, $K = 1$ in the pretraining stage). This avoids the well-known cluster-collapse issue of jointly training a Gaussian-mixture VAE from scratch and provides a common initialization across the evaluated values of the entropy weight γ, so that differences in the learned cluster posterior reflect the regularizer rather than the initialization.

The sample sets $\mathcal{S}_x^{(m)}$ are empirical realizations of $P_x^{\theta_m}(I)$. Lem. 3.1 and Thm. 3.2 describe local surrogate cost and interventional response under (A4)–(A5); independent training runs need not follow such a branch. The returned samples are candidates: observationally accepted models are selected using the observational constraints discussed before Cor. 3.3, with the local guarantee in Cor. 3.3 requiring strict baseline slack.

D PROOFS

We first show how to reduce an acyclic semantic model to the canonical diagram in Sec. D.1. We then prove the approximation result in Sec. D.2 and the entropy-weight results in Secs. D.3 to D.5. The approximation proof uses (A1)–(A3) and the finite-dimensional exogenous domain specified in Thm. 2.2. It establishes approximation within the canonical structural class and does not assert universal approximation by a particular neural architecture. The entropy-weight proofs use (A4)–(A5); the admissibility result additionally requires strict baseline slack.

D.1 REDUCTION TO THE CANONICAL DIAGRAM

Consider an ASCM whose finitely many semantic variables form an acyclic directed graph, with a designated treatment node X, joint semantic exogenous vector $\boldsymbol{U}$, and an image renderer with independent noise $\mathcal{E}_I$. Let Y collect the semantic descendants of X and let Z collect all other semantic variables except X. Either block may be empty, in which case its domain is a singleton. Assume the original structural assignments are well-defined for the intervention values under consideration.

Proof. We first construct the grouped semantic mechanisms. No variable in Z can have a parent in $\{X\} \cup Y$, since such an edge would make that variable a descendant of X. We may therefore evaluate the equations for Z in topological order and substitute them into one another. This gives a joint assignment

$$Z = f_Z(\boldsymbol{U}).$$

Every semantic parent of X belongs to Z. Hence the original treatment assignment can be written as

$$X = f_X(Z, \boldsymbol{U}),$$

where f_X ignores any coordinates of Z that it does not use. Finally, evaluating the descendants of X in topological order gives

$$Y = f_Y(X, Z, \boldsymbol{U}).$$

We retain the original joint distribution of $\boldsymbol{U}$. The resulting assignments have the block order and shared exogenous dependence allowed by Fig. 3a. Some displayed edges may be unused, since the canonical diagram permits all dependencies needed by this construction.

We next verify preservation of the observational and interventional distributions. Reindex the original renderer as $f_I(X, Y, Z, \mathcal{E}_I)$ and fix a realization of $(\boldsymbol{U}, \mathcal{E}_I)$. By construction, evaluating the grouped equations gives the same semantic values as evaluating the original equations. Applying the unchanged renderer therefore gives the same image. Since the exogenous distribution is unchanged, the grouped model has the same $P(X, I)$.

Under $\mathrm{do}(X{=}x)$, the treatment assignment is replaced by $X \leftarrow x$. The equations for Z remain unchanged, and substituting x into the grouped descendant mechanism gives exactly the original descendant values under this intervention. The rendered image is therefore also unchanged for each exogenous realization. It follows that the grouped model preserves every $P_x(I)$ for the designated treatment.

Finally, each grouped mechanism is a finite composition of the original measurable mechanisms and is therefore measurable. The grouping preserves $Y_x(u)$ for every x, u and only reindexes the semantic inputs of the renderer, so it preserves (A2)–(A3) whenever the original model satisfies them. The exogenous vector is unchanged, so its finite-dimensionality is also preserved.

This construction groups variables and composes their mechanisms; it does not marginalize semantic nodes or discretize the exogenous variables. The canonical class may allow additional dependencies, so the reduction need not preserve the smaller feasible set imposed by a detailed graph. It also does not identify the latent blocks from data or address cyclic models or arbitrary collections of treatment

nodes. Subject to the stated domain conditions, the construction proves the claimed reduction; applying Thm. 2.2 additionally requires all the assumptions of that theorem. □

D.2 Proof of Thm. 2.2

We first represent the original model by a joint distribution of treatment assignments, pre-treatment covariates, and response functions. We then approximate this distribution by finitely many response types while keeping the image mechanism unchanged.

Let ρ_I and ρ_X denote the metrics on the image domain Ω_I and treatment domain Ω_X, respectively. To pass from the coupled images to their distributions, let ρ denote the metric on the space supporting P and Q, and let W_1 be the Wasserstein distance induced by ρ. For any jointly distributed random variables $A \sim P$ and $B \sim Q$, their joint law is a coupling of P and Q, so

$$W_1(P, Q) \leq \mathbb{E}[\rho(A, B)].$$

Here $\rho(A, B)$ measures the distance between the two coupled samples. For image distributions we use the image-domain metric ρ_I; for joint image–treatment distributions we use the sum metric $\rho_I(i, i') + \rho_X(x, x')$ (Villani, 2009).

Joint response distribution. Fix the ASCM $\mathcal{M}$ in Thm. 2.2, and write $S \subseteq \mathbb{R}^m$ for the Borel domain of its shared exogenous vector $\boldsymbol{U}$. For $u \in S$, define

$$z(u) = f_Z(u), \qquad x^\natural(u) = f_X(z(u), u), \qquad r_u(x) = Y_x(u) = f_Y(x, z(u), u), \quad x \in \Omega_X.$$

Here $x^\natural(u)$ is the naturally assigned treatment, and $r_u(x)$ is the outcome obtained by setting the treatment to x while keeping the exogenous realization fixed. Define

$$T(u) = (x^\natural(u), z(u), r_u), \qquad (X^\natural, Z, R) = T(\boldsymbol{U}), \qquad Q = T_\# P_{\boldsymbol{U}}.$$

Thus Q is the distribution of the treatment assignment, covariates, and entire response function induced by one exogenous draw. It can be written as

$$Q(dx^\natural, dz, dr) = P_{X,Z}(dx^\natural, dz)\, P_{R|X,Z}(dr \mid x^\natural, z),$$

where $P_{X,Z}$ is the original observational joint distribution of treatment and covariates. The standard Borel property needed for this conditional distribution is verified in the lemma below. For each fixed x, evaluating the response at x gives the joint distribution $P_{X,Z}(dx^\natural, dz)P(Y_x \in dy \mid X = x^\natural, Z = z)$. Keeping the entire response R ensures that the same draw specifies outcomes for every treatment. The candidate SCM supplies this joint response distribution; the construction requires neither its identification from observational data nor conditional independence among potential outcomes.

To see why this representation preserves the original model, sample $(x^\natural, z, r) \sim Q$ as a shared exogenous variable and set

$$Z \leftarrow z, \qquad X \leftarrow x^\natural, \qquad Y \leftarrow r(X).$$

By the definitions of $x^\natural$ and r_u, the observational outcome $r_{\boldsymbol{U}}(x^\natural(\boldsymbol{U}))$ equals the original Y. Under $\mathrm{do}(X{=}x)$, the same draw gives

$$(X, Y, Z) = (x, r_{\boldsymbol{U}}(x), z(\boldsymbol{U})),$$

which is the original interventional semantic vector. Keeping f_I and the independent rendering noise therefore preserves $P(X, I)$ and every $P_x(I)$ exactly.

The shared exogenous variable in this representation need not have finite support. For a finite treatment domain $\{x_1, \ldots, x_k\}$, its response component is the vector $(Y_{x_1}, \ldots, Y_{x_k})$; for continuous treatments it is a continuous function. The following lemma constructs finite approximations of this representation.

Lemma D.1 (Finite approximation of the joint response distribution)**.** *Under the assumptions of Thm. 2.2, there is a sequence of CASCMs $\mathcal{N}_n$ with finite confounders, the same image mechanism f_I, and the same image-noise distribution as $\mathcal{M}$, admitting a coupling such that*

$$\boldsymbol{V}^{\mathcal{N}_n} \longrightarrow \boldsymbol{V}^{\mathcal{M}}, \qquad \boldsymbol{V}_x^{\mathcal{N}_n} \longrightarrow \boldsymbol{V}_x^{\mathcal{M}} \quad \textit{for every } x \in \Omega_X,$$

almost surely, where $\boldsymbol{V} = (X, Y, Z)$. The same probability-one event works for all x.

Proof. We first specify a metric in which response types can be approximated. For two continuous responses, define

$$d_\infty(r, r') = \sup_{x \in \Omega_X} \rho_Y(r(x), r'(x)).$$

For response types, use the sum metric

$$d\big((x, z, r), (x', z', r')\big) = \rho_X(x, x') + \rho_Z(z, z') + d_\infty(r, r').$$

Here ρ_X, ρ_Y, ρ_Z denote the metrics on the semantic domains. For example, continuous scalar or vector attributes may use Euclidean distance, $\rho(a, b) = \|a - b\|_2$, while finite categorical attributes may use the discrete metric, $\rho(a, b) = \mathbf{1}\{a \neq b\}$. For attributes containing both continuous and categorical components, the component metrics may be summed. Since each response is continuous on the compact domain Ω_X, the supremum in d_∞ is finite.

We next verify measurability of T in this metric. By (A2), each r_u belongs to $C(\Omega_X, \Omega_Y)$, which is separable under d_∞ because Ω_X is compact metric and Ω_Y is separable metric. Choose a countable dense subset $\{t_k\}_{k \geq 1}$ of Ω_X. For each fixed $r \in C(\Omega_X, \Omega_Y)$, continuity of the responses gives

$$d_\infty(r_u, r) = \sup_{k \geq 1} \rho_Y(r_u(t_k), r(t_k)).$$

Each evaluation $u \mapsto r_u(t_k)$ is Borel measurable by measurability of the structural assignments. The displayed supremum is therefore measurable. A countable base of balls in the response space shows that $u \mapsto r_u$ is Borel measurable. The maps $z(u)$ and $x^\natural(u)$ are also Borel measurable, so T is a Borel map into the separable metric space $\Omega_X \times \Omega_Z \times C(\Omega_X, \Omega_Y)$.

We now construct the finite approximation. Since S is a Borel subset of a finite-dimensional Euclidean space, $P_{\boldsymbol{U}}$ is a Radon probability measure on S. Lusin's theorem applied to the measurable map T gives, for every n, a compact set $K_n \subseteq S$ such that $T|_{K_n}$ is continuous and

$$P(\boldsymbol{U} \notin K_n) \leq 2^{-n}.$$

Continuity of the restriction $T|_{K_n}$ implies that $L_n = T(K_n)$ is compact. Moreover,

$$Q(L_n^c) = P(T(\boldsymbol{U}) \notin L_n) \leq P(\boldsymbol{U} \notin K_n) \leq 2^{-n}.$$

It follows that Q is concentrated on $\bigcup_n L_n$. This countable union of compact metric spaces is a standard Borel space. The response coordinate has the same property after projection, which justifies the conditional distribution used above on a set of full probability. Only the compact sets L_n are needed; the entire response-function space need not be compact.

By compactness, L_n admits a finite Borel partition into nonempty cells $B_{n,1}, \ldots, B_{n,d_n}$ of diameter at most $1/n$. Choose one representative from each cell,

$$a_{n,j} = (x_{n,j}, z_{n,j}, r_{n,j}) \in B_{n,j}.$$

For the complement, set $B_{n,0} = L_n^c$ and choose $a_{n,0} = T(u_*)$ for a fixed $u_* \in S$. Define the cell index and the approximation map by

$$J_n(a) = j \quad \text{when } a \in B_{n,j}, \qquad q_n(a) = a_{n,J_n(a)}.$$

These maps are Borel measurable. Let the finite confounder $U_n \in \{0, \ldots, d_n\}$ have probabilities

$$P(U_n = j) = Q(B_{n,j}), \qquad Q_n = \sum_{j=0}^{d_n} Q(B_{n,j})\, \delta_{a_{n,j}}.$$

Thus Q_n is the distribution obtained by replacing each response type by its cell representative. Each mixture component retains the treatment assignment, covariates, and entire response together. Taking the inverse images of the cells under T also gives a finite measurable partition of the original exogenous domain.

We use these representatives to define the semantic mechanisms of $\mathcal{N}_n$:

$$f_Z^{(n)}(j, e_Z) = z_{n,j},$$
$$f_X^{(n)}(z, j, e_X) = x_{n,j},$$
$$f_Y^{(n)}(x, z, j, e_Y) = r_{n,j}(x).$$

Choose auxiliary noises that are mutually independent and independent of U_n and $\mathcal{E}_I$. The displayed mechanisms ignore these auxiliary noises and the z inputs. Since U_n has the discrete topology and every $r_{n,j}$ is continuous, all three mechanisms are continuous. The canonical diagram allows unused parents and permits the dependence induced by the shared confounder.

Set $f_I^{(n)} = f_I$ and retain the original distribution of $\mathcal{E}_I$, independently of U_n. The observed domains are unchanged, so (A1) remains satisfied. Every constructed response $r_{n,j}$ is continuous, giving (A2), and the unchanged mixing function preserves (A3). After relabeling the finite confounder states, $\mathcal{N}_n$ is therefore a CASCM. This construction is designed for interventions on X; it need not preserve interventions on Z or restrictions imposed by a smaller causal diagram.

It remains to prove convergence under a common coupling. Draw $\boldsymbol{U}$ from its original distribution and set $U_n = J_n(T(\boldsymbol{U}))$. By construction, $P(U_n = j) = Q(B_{n,j})$, so this coupling has the required marginal distributions. Write

$$(X_n^\natural, Z_n, R_n) = q_n(T(\boldsymbol{U})).$$

Since $Q(L_n^c) \le 2^{-n}$, we have $\sum_n Q(L_n^c) < \infty$. The Borel–Cantelli lemma therefore implies that $T(\boldsymbol{U}) \in L_n$ for all sufficiently large n, almost surely. On this event, the response type and its representative belong to the same cell, so

$$d\big((X_n^\natural, Z_n, R_n), (X^\natural, Z, R)\big) \le 1/n \quad \text{eventually.}$$

Each term in the sum metric is nonnegative. Hence

$$X_n^\natural \to X^\natural, \qquad Z_n \to Z, \qquad d_\infty(R_n, R) \to 0 \quad \text{almost surely.}$$

For the observational outcome, the triangle inequality gives

$$\rho_Y\big(R_n(X_n^\natural), R(X^\natural)\big) \le d_\infty(R_n, R) + \rho_Y\big(R(X_n^\natural), R(X^\natural)\big) \longrightarrow 0.$$

The first term tends to zero by uniform convergence of the responses, and the second tends to zero by continuity of the sampled response R. Together with convergence of $X_n^\natural$ and Z_n, this proves $\boldsymbol{V}^{\mathcal{N}_n} \to \boldsymbol{V}^{\mathcal{M}}$ almost surely.

Under $\mathrm{do}(X{=}x)$, the coupled semantic vectors are $(x, R_n(x), Z_n)$ and $(x, R(x), Z)$. For every $x \in \Omega_X$,

$$\rho_Y(R_n(x), R(x)) \le d_\infty(R_n, R) \longrightarrow 0.$$

This bound holds on the same probability-one event for all x. Combining it with $Z_n \to Z$ proves the interventional convergence and completes the proof. □

Proof of Thm. 2.2. We first prove convergence of the observational distributions. Use the coupling from Lem. D.1, and give both models the same draw of $\mathcal{E}_I$, independent of their semantic exogenous variables. By (A3), $f_I(\cdot, e)$ is continuous in its endogenous inputs for each fixed e. On the probability-one convergence event of the lemma, applying the unchanged mixing function with the shared rendering-noise draw gives

$$I^{\mathcal{N}_n} = f_I(\boldsymbol{V}^{\mathcal{N}_n}, \mathcal{E}_I) \longrightarrow f_I(\boldsymbol{V}^{\mathcal{M}}, \mathcal{E}_I) = I^{\mathcal{M}} \quad \text{almost surely.}$$

Equip $\Omega_I \times \Omega_X$ with the sum metric. Applying the coupling bound to the observational pairs yields

$$W_1\big(P(I, X; \mathcal{N}_n), P(I, X; \mathcal{M})\big) \le \mathbb{E}\big[\rho_I(I^{\mathcal{N}_n}, I^{\mathcal{M}}) + \rho_X(X_n^\natural, X^\natural)\big].$$

Both distances inside the expectation tend to zero almost surely. By (A1), they are bounded by the finite diameters D_I and D_X. Dominated convergence therefore gives

$$W_1\big(P(I, X; \mathcal{N}_n), P(I, X; \mathcal{M})\big) \longrightarrow 0.$$

The bound depends only on the image and treatment domains, so neither compactness of the rendering-noise domain nor continuity in the rendering noise is required.

We next prove convergence of the interventional distributions. Fix $x \in \Omega_X$. On the same probability-one event, (A3) and the semantic convergence give

$$I_x^{\mathcal{N}_n} = f_I(\boldsymbol{V}_x^{\mathcal{N}_n}, \mathcal{E}_I) \longrightarrow f_I(\boldsymbol{V}_x^{\mathcal{M}}, \mathcal{E}_I) = I_x^{\mathcal{M}} \quad \text{almost surely.}$$

Since the image distance is bounded by D_I, the coupling bound and dominated convergence imply

$$W_1\big(P_x(I;\mathcal{N}_n), P_x(I;\mathcal{M})\big) \le \mathbb{E}\big[\rho_I(I_x^{\mathcal{N}_n}, I_x^{\mathcal{M}})\big] \longrightarrow 0.$$

It remains to integrate this estimate over the treatment domain. The evaluation map $(x, r) \mapsto r(x)$ is continuous for the uniform response metric, and T, q_n, and f_I are Borel measurable. Thus the coupled images are jointly measurable in the treatment and the exogenous draws. Integrating the coupling bound gives

$$\int_{\Omega_X} W_1\big(P_x(I;\mathcal{N}_n), P_x(I;\mathcal{M})\big)\, dx \le \int_{\Omega_X} \mathbb{E}\big[\rho_I(I_x^{\mathcal{N}_n}, I_x^{\mathcal{M}})\big]\, dx.$$

The integrand on the right tends to zero and is bounded by D_I. For continuous treatments, the compact Euclidean treatment domain has finite Lebesgue measure, so this constant is integrable. For finite categorical treatments, the integral is a finite sum. Dominated convergence therefore yields

$$\int_{\Omega_X} W_1\big(P_x(I;\mathcal{N}_n), P_x(I;\mathcal{M})\big)\, dx \longrightarrow 0.$$

Finally, let $\epsilon > 0$. Observational convergence gives an index n_{obs} after which the first approximation error is below ϵ, and integrated interventional convergence gives an index n_{int} after which the second is below ϵ. Choose $n \ge \max\{n_{\text{obs}}, n_{\text{int}}\}$ and set $\mathcal{N} = \mathcal{N}_n$. Then the same CASCM satisfies both inequalities in Eq. (1). This completes the proof. □

D.3 Proof of Lem. 3.1

Throughout these proofs, we use the notation $\eta = (\theta, \phi, \psi)$ and D, H, A, b, v, κ_0 from Sec. 3.1. The losses are averaged over a fixed data distribution or training sample. We study the local branch of optima in Lem. 3.1; an arbitrary sequence of SGD outputs need not follow this branch.

Proof. We first construct the local branch of minimizers. Define

$$F(\eta, \gamma) = \nabla_\eta D(\eta) - \gamma \nabla_\eta H(\eta).$$

Thus $F(\eta, \gamma) = 0$ is the stationarity condition for $D - \gamma H$. By (A4), F is C^1 and

$$F(\eta_0, 0) = 0, \qquad \partial_\eta F(\eta_0, 0) = A \succ 0.$$

Since A is invertible, the implicit function theorem gives a unique nearby C^1 branch η_γ satisfying $F(\eta_\gamma, \gamma) = 0$ and passing through η_0.

It remains to verify that these stationary points are strict local minimizers. The Hessian of the regularized objective along the branch is

$$\mathcal{H}_\gamma = \nabla_\eta^2 D(\eta_\gamma) - \gamma \nabla_\eta^2 H(\eta_\gamma).$$

At $\gamma = 0$, this matrix equals $A \succ 0$. By continuity, there is a sufficiently small $\bar{\gamma} > 0$ such that $\mathcal{H}_\gamma \succ 0$ for $0 \le \gamma < \bar{\gamma}$. Hence every stationary point on this interval is a strict local minimizer.

We next derive the three expansions. Differentiating $F(\eta_\gamma, \gamma) = 0$ with respect to γ gives

$$\mathcal{H}_\gamma \eta_\gamma' - \nabla_\eta H(\eta_\gamma) = 0.$$

Since $\mathcal{H}_\gamma$ is invertible, it follows that

$$\eta_\gamma' = \mathcal{H}_\gamma^{-1} \nabla_\eta H(\eta_\gamma), \qquad \eta_0' = A^{-1} b = v. \tag{13}$$

The differentiability of the branch therefore yields

$$\eta_\gamma = \eta_0 + \gamma v + o(\gamma).$$

Since $\nabla_\eta D(\eta_0) = 0$, Taylor expansion of D at η_0 gives

$$D(\eta_\gamma) - D(\eta_0) = \tfrac{1}{2}(\eta_\gamma - \eta_0)^\top A (\eta_\gamma - \eta_0) + o(\|\eta_\gamma - \eta_0\|^2).$$

Substituting the parameter expansion and using $v = A^{-1}b$, we obtain

$$D(\eta_\gamma) - D(\eta_0) = \tfrac{1}{2}\gamma^2 v^\top A v + o(\gamma^2) = \tfrac{1}{2}\kappa_0\gamma^2 + o(\gamma^2).$$

Finally, H depends only on ϕ, and its gradient with respect to η at the baseline is b. Its first-order expansion is therefore

$$H(\phi_\gamma) - H(\phi_0) = b^\top(\eta_\gamma - \eta_0) + o(\|\eta_\gamma - \eta_0\|).$$

Using $b^\top v = b^\top A^{-1} b = \kappa_0$, we conclude that

$$H(\phi_\gamma) - H(\phi_0) = \kappa_0\gamma + o(\gamma).$$

All three expansions remain valid when $b = 0$, in which case $v = 0$ and $\kappa_0 = 0$. Since $A \succ 0$, we also have $\kappa_0 > 0$ whenever $b \neq 0$. The C^2 assumption gives the stated little-o remainders; a cubic big-O remainder for the cost would require additional regularity. This proves the lemma. □

D.4 Proof of Thm. 3.2

Proof. We first express the distributional response in the norm from (A5). For a finite signed measure μ of total mass zero on the compact image domain from (A1), define

$$\|\mu\|_{\mathrm{KR}} = \sup_{\mathrm{Lip}(u)\le 1} \left| \int u \, d\mu \right|.$$

By the Kantorovich–Rubinstein identity, $\|P - Q\|_{\mathrm{KR}} = W_1(P, Q)$ for probability measures P, Q. Thus it is enough to estimate the difference of the interventional distributions in this norm.

Let $\mathrm{d}\mathsf{P}_x(\eta_0)$ be the bounded linear derivative supplied by (A5). Its values are finite signed measures of total mass zero, since the derivative describes a change in a probability distribution. By differentiability,

$$\|\mathsf{P}_x(\eta_0 + h) - \mathsf{P}_x(\eta_0) - \mathrm{d}\mathsf{P}_x(\eta_0)[h]\|_{\mathrm{KR}} = o(\|h\|), \qquad h \to 0.$$

Assumption (A5) also gives $\dot{P}_x = \mathrm{d}\mathsf{P}_x(\eta_0)[v] \neq 0$. Set

$$\rho_x = \|\dot{P}_x\|_{\mathrm{KR}} > 0.$$

By Lem. 3.1, the parameter displacement satisfies $\eta_\gamma - \eta_0 = \gamma v + o(\gamma)$. Since the derivative is bounded and linear,

$$\mathrm{d}\mathsf{P}_x(\eta_0)[\eta_\gamma - \eta_0] = \gamma\dot{P}_x + o(\gamma) \quad \text{in the KR norm.}$$

Substituting $h = \eta_\gamma - \eta_0$ into the differentiability expansion therefore gives

$$\|P_x^{\theta_\gamma} - P_x^{\theta_0} - \gamma\dot{P}_x\|_{\mathrm{KR}} = o(\gamma).$$

We now convert this estimate into a Wasserstein distance. By the reverse triangle inequality,

$$\left|W_1(P_x^{\theta_\gamma}, P_x^{\theta_0}) - \gamma\rho_x\right| \le \|P_x^{\theta_\gamma} - P_x^{\theta_0} - \gamma\dot{P}_x\|_{\mathrm{KR}} = o(\gamma).$$

Hence

$$W_1(P_x^{\theta_\gamma}, P_x^{\theta_0}) = \gamma\rho_x + o(\gamma),$$

which proves the first limit. Since $\rho_x > 0$, this distance is positive for every sufficiently small $\gamma > 0$.

It remains to compare this movement with the surrogate cost. Using the cost expansion in Lem. 3.1, we obtain

$$\begin{aligned}\frac{D(\eta_\gamma) - D(\eta_0)}{W_1(P_x^{\theta_\gamma}, P_x^{\theta_0})} &= \frac{\tfrac{1}{2}\kappa_0\gamma^2 + o(\gamma^2)}{\rho_x\gamma + o(\gamma)} \\ &= \frac{\kappa_0}{2\rho_x}\gamma + o(\gamma) \longrightarrow 0.\end{aligned}$$

The argument uses differentiability of the distribution in the KR norm as required by (A5); differentiability of a single expectation alone would not give the norm remainder above. This proves the second limit and completes the proof. □

D.5 Proof of Cor. 3.3

Proof. We first show that all observational constraints remain satisfied on a common interval. For each $j \in \{1, \dots, J\}$, define the baseline margin

$$m_j = \varepsilon_j - \Delta_j(\theta_0) > 0.$$

The branch θ_γ is continuous at $\gamma = 0$, and Δ_j is continuous at θ_0. Hence there is a $\delta_j > 0$ such that

$$|\Delta_j(\theta_\gamma) - \Delta_j(\theta_0)| < m_j \qquad \text{for } 0 \le \gamma < \delta_j.$$

On this interval, it follows that

$$\Delta_j(\theta_\gamma) < \Delta_j(\theta_0) + m_j = \varepsilon_j.$$

Since there are finitely many constraints, $\delta_{\text{obs}} = \min_{1 \le j \le J} \delta_j$ is positive. Every $0 \le \gamma < \delta_{\text{obs}}$ therefore satisfies all the observational constraints.

We next ensure that the interventional distribution changes on the same interval. By Thm. 3.2, there is a $\delta_{\text{move}} > 0$ such that

$$W_1(P_x^{\theta_\gamma}, P_x^{\theta_0}) > 0 \qquad \text{for } 0 < \gamma < \delta_{\text{move}}.$$

Choose $\delta = \min\{\delta_{\text{obs}}, \delta_{\text{move}}, \bar{\gamma}\} > 0$. Then every $0 \le \gamma < \delta$ satisfies the observational constraints, and every $0 < \gamma < \delta$ gives $P_x^{\theta_\gamma} \neq P_x^{\theta_0}$.

This argument uses the strict baseline margins and continuity. If the discrepancies are also differentiable, the parameter expansion gives

$$\Delta_j(\theta_\gamma) - \Delta_j(\theta_0) = \gamma \nabla_\theta \Delta_j(\theta_0)^\top v_\theta + o(\gamma),$$

which need not be second order. The conclusion therefore concerns acceptance under the fixed observational tolerances and does not assert equality of the full observational distributions. This completes the proof. □

E Experimental Setup and Additional Results

We evaluate CauVaDE on synthetic Color-MNIST at 28×28 resolution, CelebA face attributes at 64×64, and MIMIC-CXR-JPG chest radiographs at 224×224. We train on confounded data and use an unconfounded reference only for evaluation. Following Sec. 2, only the treatment X is used as supervision; outcome Y, covariate Z, and confounder U labels are not supplied to the models during training.

We first describe dataset construction (Secs. E.1.1 to E.1.3) and training configurations (Sec. E.2), then report reconstruction diagnostics and qualitative samples. We next examine mixed discrete–continuous confounding and non-binary attributes, followed by an entropy ablation, cluster-count sensitivity, and exploration of plausible confounder distributions.

E.1 Dataset Construction

For each dataset, we construct confounded and unconfounded variants through synthetic generation or resampling existing images. In the confounded variant, treatment assignment depends on U, creating the backdoor path $X \leftarrow U \rightarrow Y$. The unconfounded variant samples X independently of U while retaining the remaining construction, providing a reference for evaluating interventional samples from $P_x(I)$. The dataset-specific mechanisms and assignment probabilities are given below.

E.1.1 Confounded Color-MNIST

Color-MNIST is built from three binary variables: treatment X, outcome Y, and unobserved confounder U, with $P(U{=}0) = 0.6$ and $P(U{=}1) = 0.4$. The outcome follows the XOR mechanism $Y = U \oplus X$, taking value 1 whenever U and X disagree. A Conditional GAN, conditioned on the (X, Y) stratum, then generates 28×28 coloured digit images. We use $N = 1{,}000$ samples. The mechanism specifies the population reference exactly; empirical frequencies remain subject to sampling error. Treatment assignment is specified as follows:

- **Confounded**: $P(X{=}1 \mid U{=}0) = 0.7$ and $P(X{=}1 \mid U{=}1) = 0.3$, inducing the spurious correlations $P(Y{=}1 \mid X{=}0) = 0.609$ and $P(Y{=}1 \mid X{=}1) = 0.778$.
- **Unconfounded**: $X \perp\!\!\!\perp U$, yielding the ground-truth interventional probabilities $P(Y{=}1 \mid \mathrm{do}(X{=}0)) = 0.40$ and $P(Y{=}1 \mid \mathrm{do}(X{=}1)) = 0.60$.

E.1.2 CONFOUNDED CELEBA

CelebA (Liu et al., 2015) is built by mapping five attributes to causal variables: Z (`Male`), U (`Attractive`, unobserved), X (`Young`, treatment), W (`Wearing_Earrings`), and Y (`Heavy_Makeup`, outcome). The joint distribution is specified by $P(Z{=}1) = 0.4$, $P(U{=}1 \mid Z{=}0) = 0.60$, $P(U{=}1 \mid Z{=}1) = 0.40$, and $P(W{=}1 \mid X, U)$ rising monotonically from 0.1 at $(X, U) = (0, 0)$ to 0.9 at $(X, U) = (1, 1)$. Images are then resampled from the original CelebA corpus to match the target joint distribution over (Z, X, W) with U marginalized out. Treatment assignment is specified as follows:

- **Confounded** ($N = 10{,}000$): $P(X{=}1 \mid U{=}0) = 0.1$ and $P(X{=}1 \mid U{=}1) = 0.9$.
- **Unconfounded** ($N = 10{,}000$): $P(X{=}1) = 0.5$ regardless of U.

E.1.3 CONFOUNDED MIMIC-CXR-JPG

MIMIC-CXR-JPG (Johnson et al., 2024; 2019; Goldberger et al., 2000) is built from three causal variables: X (`Pneumonia`, treatment), Y (subsequent outcome variables, e.g., `Lung Opacity`, unlabeled), and U (unobserved confounder), grouping the patient's age A in years and whether a radiograph was acquired in the anteroposterior (AP) view V. These attributes define the hidden grouping used in our resampling experiment; the stipulated causal mechanism is part of the experimental construction. The dataset is partitioned into four atomic pools by (X, Y) stratum, from which samples are drawn to match the target distribution. Treatment is assigned through a logistic model,

$$P(X{=}1 \mid A{=}a, V{=}v) = \sigma\left(\beta_0 + \beta_A \frac{a - 60}{10} + \beta_V\, v\right), \qquad \sigma(z) = \frac{1}{1 + e^{-z}}, \tag{14}$$

with the following coefficients:

- **Confounded** ($N = 20{,}000$): $\beta_0 = 0$, $\beta_A = \frac{1}{2} \ln 9 \approx 1.10$, and $\beta_V = \ln 9 \approx 2.20$.
- **Unconfounded** ($N = 20{,}000$): $P(X{=}1) = 0.5$ regardless of $U = (A, V)$.

E.2 TRAINING DETAILS AND HYPERPARAMETERS

Optimization. All models are trained with the AdamW optimizer (Loshchilov & Hutter, 2019) and weight decay $1{\times}10^{-5}$. Color-MNIST and CelebA experiments run on a single NVIDIA RTX 3080 Ti (16GB VRAM); MIMIC-CXR-JPG experiments run on a single NVIDIA A100 (80GB VRAM) on an internal cluster.

Architectural scaling. The architecture scales with image resolution through the base channel capacity h and latent dimension d_z. Convolutional stages use channel widths $h, 2h, 4h, \ldots$, while d_z sets the dimension of the global latent $\mathbf{z}$ and increases for higher-resolution datasets.

Dataset-specific configurations. Table 3 lists the configurations for all three datasets. The reconstruction, KL, and SCM loss weights (α, β, λ) are calibrated against the input resolution to maintain stable convergence and avoid posterior collapse; the cluster count $K = 8$ is held fixed across datasets, with extra clusters absorbed by the prior $P(C)$ rather than overfit. The entropy weight γ is varied over a per-dataset grid, producing the family of generators reported in the main text.

E.3 RECONSTRUCTION QUALITY AND ADDITIONAL SAMPLES

For Color-MNIST and CelebA, we assess observational admissibility through agreement with the labeled joint distribution $P(X, Y)$ and the stability of image reconstruction NLL and MSE across entropy weights. For MIMIC-CXR-JPG, we use $\gamma = 0$ as the baseline and assess changes in observational fit alongside the corresponding change in FID against the unconfounded reference. This comparison checks whether improved interventional image-distribution agreement is accompanied

Table 3: Dataset-specific hyperparameters for CAUVADE.

Hyperparameter	**Colored MNIST**	**CelebA**	**MIMIC-CXR**
Input Resolution	28×28	64×64	224×224
Total Epochs	400	800	800
Batch Size	128	64	128
Learning Rate	1×10^{-4}	1×10^{-4}	1×10^{-4}
Base Channel Capacity (h)	8	64	32
Latent Dimension (d_z)	64	512	512
Number of Clusters (K)	8	8	8
Recon. Weight (α)	10.0	15.0	15.0
KL Weight (β)	0.1	0.05	0.05
SCM Weight (λ)	5.0	3.0	3.0
Entropy Weight (γ)	$\{0, 1, 2, 10, 20, 50, 100\}$	$\{0, 1, 10, 100\}$	

Table 4: Reconstruction quality on Confounded Color-MNIST across the entropy weights γ evaluated for this diagnostic. Mean squared error (MSE) and reconstruction negative log-likelihood (NLL) are reported per pixel; lower is better. Label RMSE is the root-mean-square error between the model-implied observational joint probabilities $P(X, Y)$ and the empirical joint probabilities estimated from labeled data.

Metric	γ=0	γ=1	γ=2	γ=10	γ=20	γ=30	γ=40	γ=50	γ=60	γ=70	γ=80	γ=90	γ=100
MSE (per pixel)	0.0307	0.0303	0.0313	0.0315	0.0323	0.0366	0.0330	0.0336	0.0329	0.0321	0.0319	0.0312	0.0318
NLL (per pixel)	0.1161	0.1153	0.1186	0.1190	0.1225	0.1362	0.1243	0.1260	0.1238	0.1215	0.1202	0.1182	0.1199
Label RMSE	0.0517	0.0519	0.0516	0.0411	0.0441	0.0309	0.0344	0.0232	0.0102	0.0143	0.0067	0.0114	0.0079

by deterioration in observational fit. We also inspect generated images across entropy weights and interventions.

E.3.1 RECONSTRUCTION DIAGNOSTICS

Tables 4 and 5 report per-pixel reconstruction MSE and NLL, together with Label RMSE for the observational joint distribution $P(X, Y)$, across the entropy weights evaluated for this diagnostic on Confounded Color-MNIST and CelebA, respectively. These grids differ from the generation settings in Table 3. On Confounded Color-MNIST, MSE ranges from 0.0303 to 0.0366, and NLL from 0.1153 to 0.1362. On CelebA, the variation is smaller: MSE ranges from 0.0216 to 0.0237, and NLL from 0.5374 to 0.5426. Label RMSE measures agreement between the model-implied observational joint probabilities and those estimated from labeled (X, Y) data. It ranges from 0.0067 to 0.0519 on Color-MNIST and from 0.0055 to 0.0387 on CelebA. These small errors directly support fit to the labeled semantic distribution. Together with the stable reconstruction metrics, they provide strong empirical evidence for observational admissibility at the measured image and semantic levels as interventional predictions vary. They do not establish equality of the full observational distributions; the formal local guarantee remains subject to the assumptions and discrepancy tolerances of Cor. 3.3.

E.3.2 ADDITIONAL QUALITATIVE SAMPLES

Color-MNIST. Fig. 10 shows representative CAUVADE samples on Confounded Color-MNIST as the entropy weight γ varies, illustrating the visual fidelity of the model at 28×28 resolution and how the entropy regularizer reshapes the cluster posterior at each setting.

CelebA. Fig. 11 compares samples from the confounded training set, the unconfounded ground truth, two baselines (VAE and ANCM), and CAUVADE at four values of γ, illustrating the visual fidelity of the model at 64×64 resolution and the diversity of generators CAUVADE produces across entropy-weight settings.

MIMIC-CXR-JPG. Fig. 12 shows CAUVADE samples under both interventions, $P_{X=0}(I)$ and $P_{X=1}(I)$, illustrating the visual fidelity of the model at 224×224 resolution.

Table 5: Reconstruction quality on CelebA across the entropy weights γ evaluated for this diagnostic. Mean squared error (MSE) and reconstruction negative log-likelihood (NLL) are reported per pixel; lower is better. Label RMSE is the root-mean-square error between the model-implied observational joint probabilities $P(X, Y)$ and the empirical joint probabilities estimated from labeled data.

Metric	γ=0	γ=1	γ=10	γ=20	γ=30	γ=40	γ=50	γ=60	γ=70	γ=80	γ=90	γ=100
MSE (per pixel)	0.0216	0.0221	0.0224	0.0216	0.0228	0.0233	0.0228	0.0237	0.0227	0.0233	0.0237	0.0233
NLL (per pixel)	0.5374	0.5392	0.5397	0.5378	0.5409	0.5421	0.5402	0.5426	0.5399	0.5413	0.5424	0.5410
Label RMSE	0.0101	0.0252	0.0132	0.0387	0.0255	0.0383	0.0210	0.0322	0.0055	0.0343	0.0157	0.0069

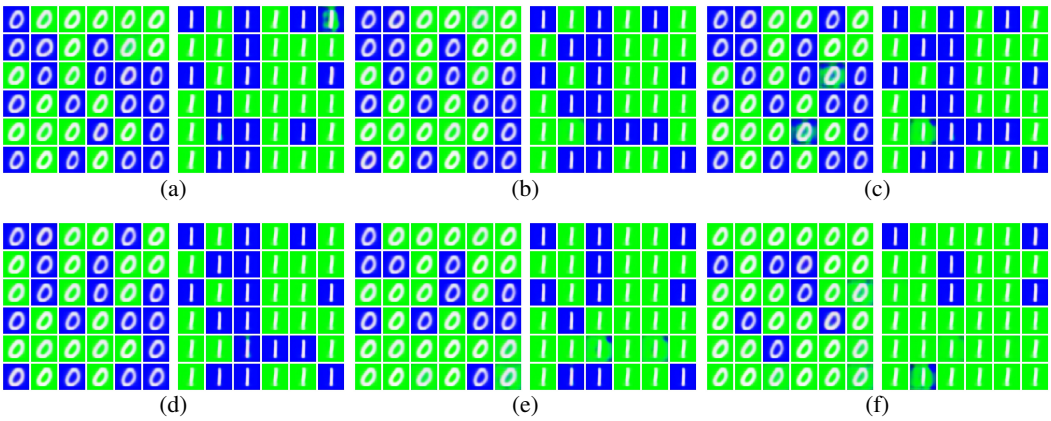


Figure 10: Generated samples from CAUVADE on Colored MNIST across varying constraint strengths: (a) $\gamma = 1$; (b) $\gamma = 2$; (c) $\gamma = 10$; (d) $\gamma = 20$; (e) $\gamma = 50$; and (f) $\gamma = 100$.

E.4 EXTENDED EXPERIMENTAL SETTINGS

We extend the binary Color-MNIST construction in two directions: mixed discrete–continuous confounding and non-binary treatment and outcome attributes.

E.4.1 MIXED DISCRETE–CONTINUOUS CONFOUNDING

We evaluate on a Confounded Color-MNIST variant whose confounder mixes discrete and continuous sources: $U = (U_0, U_1, U_2, U_3)$ with $U_0, U_1 \sim \text{Bernoulli}(0.4)$ and $U_2, U_3 \sim \text{Uniform}(0, 1)$, combined into a single continuous score $Z = \sum_i U_i$. The treatment propensity is a smooth logistic function of the full continuous Z,

$$P(X{=}1 \mid Z) = \sigma\big(k(z^* - Z)\big). \tag{15}$$

A quantile-calibrated threshold on Z keeps the marginal $P(b{=}1) = 0.4$, matching the binarized confounder of Sec. E.1.1. The resulting Manski intervals are $[0.288, 0.837]$ for $do(X{=}0)$ and $[0.423, 0.874]$ for $do(X{=}1)$. Table 6 reports the γ-sweep; all twelve swept estimates fall within these intervals.

Table 6: γ-sweep on the mixed discrete–continuous confounder. All six swept estimates on both interventions fall within the Manski intervals (12/12).

γ	$P(do\,X{=}0)$	in $[0.288, 0.837]$	$P(do\,X{=}1)$	in $[0.423, 0.874]$
0	0.619	✓	0.656	✓
1	0.663	✓	0.703	✓
2	0.666	✓	0.706	✓
10	0.682	✓	0.720	✓
20	0.660	✓	0.704	✓
50	0.707	✓	0.737	✓

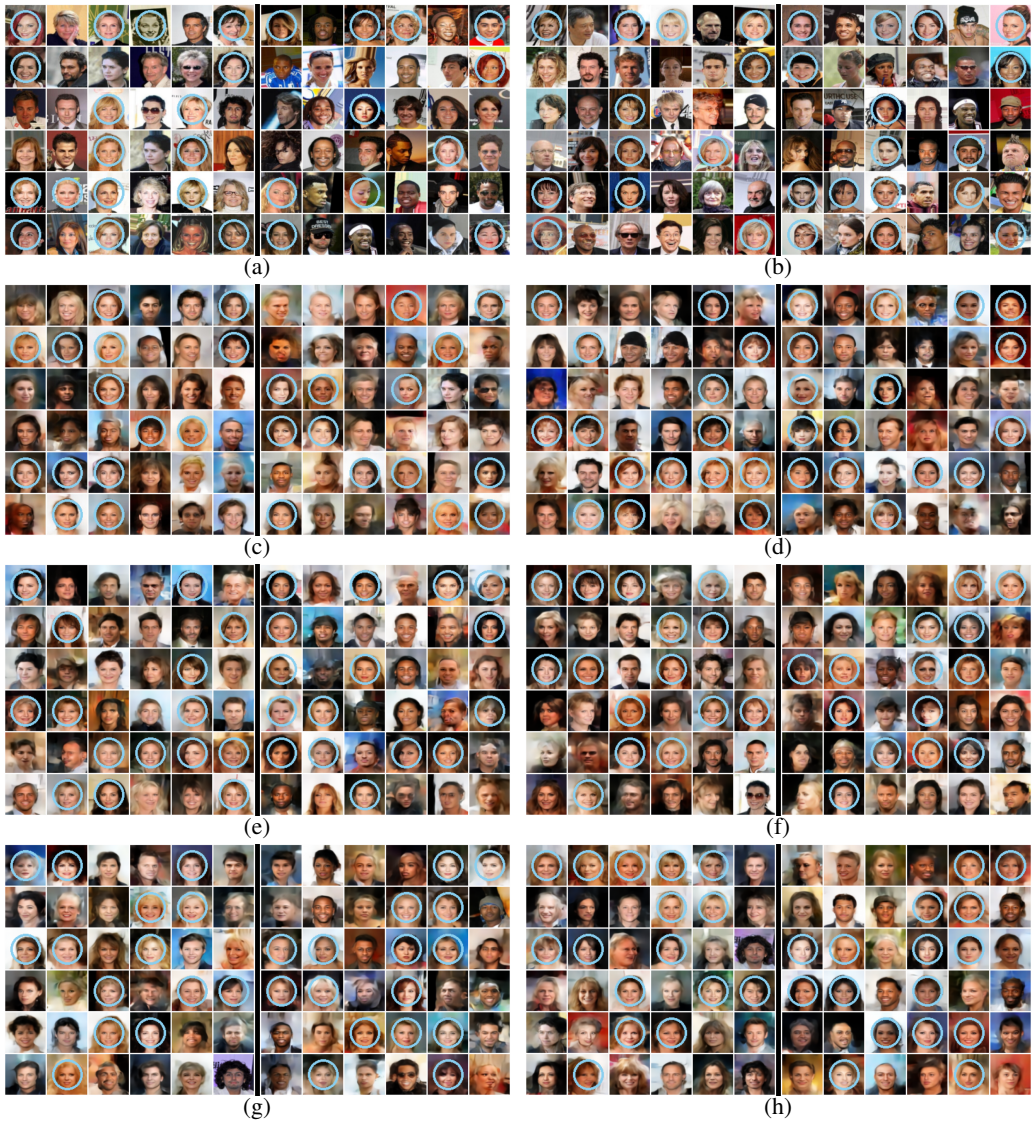


Figure 11: Visualizing data distributions on CelebA. Samples are drawn from: (a) the confounded training set; (b) the unconfounded ground truth; (c) synthetic outputs from VAE; (d) synthetic outputs from ANCM; (e) synthetic outputs from CAUVADE with γ=0; (f) synthetic outputs from CAUVADE with γ=1; (g) synthetic outputs from CAUVADE with γ=10; and (h) synthetic outputs from CAUVADE with γ=100.

E.4.2 NON-BINARY ATTRIBUTES

We evaluate on a 3-value Confounded Color-MNIST: the treatment X takes four intervention values and the outcome is a 3-way color distribution over the B/G/R channels. Table 7 reports the per-intervention B/G/R distributions across the γ-sweep, together with the ground truth. The in-bounds column counts swept predictions falling inside the Manski feasible region; all twelve do so at both γ=1 and γ=10.

E.5 ABLATIONS AND CONFOUNDER DIAGNOSTICS

We isolate the role of the entropy term, examine sensitivity to the number of latent clusters, and explore variation in fitted confounder distributions using the injected distribution as a diagnostic reference.

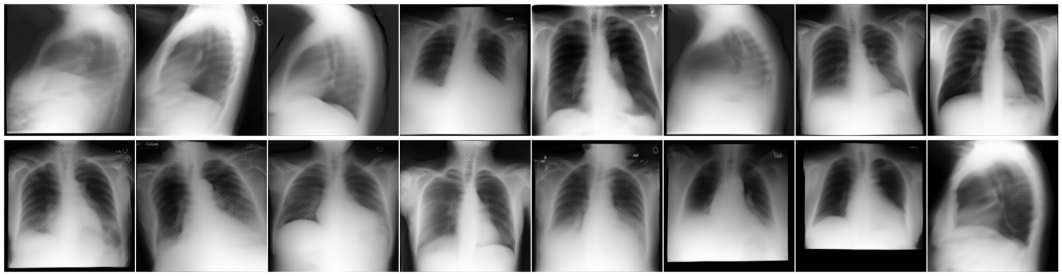


Figure 12: Generated samples from CAUVADE under interventions $P_{X=0}(I)$ (top) and $P_{X=1}(I)$ (bottom).

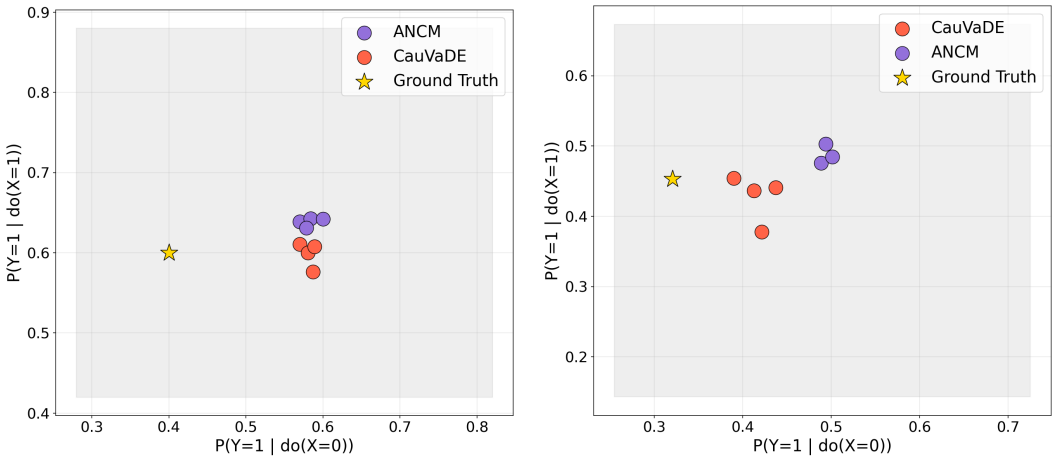
Table 7: γ-sweep on 3-value Confounded Color-MNIST. Per-intervention B/G/R outcome distributions for the four intervention values of X, with the ground truth (GT) in the first row. The in-bounds column counts swept predictions (4 interventions $\times$ 3 color channels) falling inside the Manski feasible region.

γ	$do(X{=}0)$ B/G/R	$do(X{=}1)$ B/G/R	$do(X{=}2)$ B/G/R	$do(X{=}3)$ B/G/R	in-bounds
GT	0.54/0.24/0.22	0.28/0.54/0.18	0.18/0.28/0.54	0.32/0.36/0.32	—
1	0.248/0.380/0.372	0.259/0.371/0.370	0.251/0.379/0.371	0.250/0.366/0.384	12/12
10	0.258/0.350/0.392	0.268/0.343/0.389	0.251/0.355/0.394	0.255/0.344/0.401	12/12

E.5.1 ENTROPY ABLATION

To isolate the role of the entropy term, we set $\gamma = 0$. The term $H(\phi)$ then drops out of Eq. (6), and CAUVADE fits only the data term D. In this setting, the estimate of $P_x(Y{=}1)$ produced by CAUVADE nearly coincides with that of ANCM and concentrates tightly across random seeds (Fig. 13 and Table 8). Both estimates lie inside the Manski bounds, so they are consistent with the observed data. However, each recovers only a single point of the feasible region and gives no indication of its extent.

This ablation shows that fitting the data alone does not trace the partially identified set. The entropy term moves the solution within the region, and sweeping γ is how CAUVADE produces multiple distinct candidates rather than one point inside it, without guaranteeing coverage of the region.



(a) Color-MNIST (b) Confounded CelebA

Figure 13: Estimated $P_x(Y = 1)$ on Confounded Color-MNIST and CelebA under $\gamma = 0$, which closely matches ANCM. The shaded region denotes Manski's causal bounds (Manski, 1990).

Table 8: Min and max of $P(Y = 1 \mid do(X{=}0))$ and $P(Y = 1 \mid do(X{=}1))$ across CAUVADE $\gamma = 0$ and ANCM.

Dataset	Model	$P(Y = 1 \mid do(X{=}0))$		$P(Y = 1 \mid do(X{=}1))$	
		Min	Max	Min	Max
CMNIST	ANCM	0.5700	0.6000	0.6308	0.6426
	CAUVADE	0.5700	0.5891	0.5765	0.6108
CelebA	ANCM	0.4885	0.5011	0.4755	0.5027
	CAUVADE	0.3899	0.4373	0.3779	0.4539

E.5.2 SENSITIVITY TO THE CLUSTER COUNT

We first vary K by changing γ at $K \in \{2, 4, 6\}$ on Confounded Color-MNIST, and we measure the width of the recovered interventional range for each intervention (Table 9). The in-bounds column counts how many of the 28 swept estimates per intervention fall inside the ground-truth Manski region. Going from $K{=}2$ to $K{=}4$, the width increases from 0.167 to 0.352 for $do(X{=}0)$ and from 0.171 to 0.283 for $do(X{=}1)$, and in-bounds coverage rises from $14/28$ to $26/28$. At $K{=}6$, the width is 0.216 for $do(X{=}0)$ and 0.178 for $do(X{=}1)$, and in-bounds coverage reaches $28/28$.

We next consider a large K. Table 10 reports single fitted instances at $K{=}100$ on Color-MNIST. Occupancy is the number of components with non-negligible aggregated responsibility, and $I_q(C; Z)$ is the categorical KL term of Eq. (4). Occupancy stays between 12 and 21 of 100 components, and the holdout BCE is flat across γ.

Table 9: Recovered interventional range on Confounded Color-MNIST at $K \in \{2, 4, 6\}$, aggregated over the γ-sweep. Ranges span the swept estimates inside the region; the in-bounds column counts how many of the 28 swept estimates per intervention those are.

K	in-bounds	$P(Y \mid \mathrm{do}(X{=}0))$ range (width)	$P(Y \mid \mathrm{do}(X{=}1))$ range (width)
2	14/28	$[0.500, 0.667]$ (0.167)	$[0.494, 0.665]$ (0.171)
4	26/28	$[0.392, 0.744]$ (0.352)	$[0.462, 0.745]$ (0.283)
6	28/28	$[0.593, 0.808]$ (0.216)	$[0.504, 0.682]$ (0.178)

Table 10: Single fitted instances at $K{=}100$ on Color-MNIST. These are individual model evaluations, not a recovered bound.

γ	Occupancy	$I_q(C; Z)$	Holdout BCE	$P(Y \mid \mathrm{do}(X{=}0))$	$P(Y = 1 \mid \mathrm{do}(X{=}1))$
1	21 / 100	1.290	0.762	0.574	0.596
10	20 / 100	1.289	0.760	0.583	0.610
100	12 / 100	1.274	0.760	0.625	0.696

E.5.3 EXPLORING PLAUSIBLE CONFOUNDER DISTRIBUTIONS

Under hidden confounding, neither the latent representation U nor its distribution $P(U)$ is generally identifiable from observational data. This structural ambiguity motivates our approach: drawing inspiration from partial identification, CAUVADE explores plausible confounder distributions compatible with the observations and uses them to generate diverse interventional image predictions. Recovering the true latent representation or confounder distribution is therefore not our objective.

Confounded Color-MNIST (Sec. E.1.1) provides an injected $P(U)$ that is known by construction and can serve as a diagnostic reference. Table 11 reports the average L_2 distance between the learned $P(C)$ and this reference, after matching cluster indices to confounder states. Across $\gamma \in \{1, 10, 100\}$, the distances are 0.194, 0.368, and 0.566, respectively, compared with 0.516 for ANCM. Both CAUVADE and the ANCM baseline learn confounder distributions that differ from the injected

Table 11: Average L_2 distance between the learned confounder distribution $P(C)$ and the injected $P(U)$ on Confounded Color-MNIST, used as a diagnostic reference across the γ-sweep and for ANCM. Cluster indices are matched to confounder states before the distance is computed. This comparison does not establish identification of the latent representation or its distribution.

Model	ANCM	CAUVADE		
γ	—	1	10	100
L_2	0.516	0.194	0.368	0.566

distribution. This comparison illustrates the ambiguity associated with non-identifiability, which also affects the baseline; recovering the injected confounder is not the goal of this experiment or our method. Instead, CAUVADE explores plausible exogenous distributions $P(U)$ compatible with the confounded data and uses them to generate diverse predictions of images under intervention. The varying distances across the γ-variation reflect this exploration: different values of γ select different observationally compatible confounder distributions, not converging to a specific solution.

F COMMON QUESTIONS

1. **Does varying γ compromise observational fit?**
   In Eq. (6), γ weights the additive cluster-posterior entropy term $H(\phi)$; the reconstruction, KL, and treatment-consistency terms defining D remain unchanged. Changing γ therefore preserves the definition of data fit, although it can change the fitted solution. Under their local assumptions, Lem. 3.1 and Thm. 3.2 show second-order surrogate cost and first-order interventional movement near $\gamma = 0$. With strict slack in the baseline observational checks, Cor. 3.3 guarantees that sufficiently small weights preserve admissibility. Larger weights still require checking the fitted candidates, as prescribed in Sec. 3.1.

   The Color-MNIST and CelebA experiments provide strong empirical evidence for observational admissibility at the measured image and semantic levels. Tables 4 and 5 report Label RMSE between the model-implied observational joint distribution of (X, Y) and the empirical joint distribution estimated from labeled data. It remains small across the evaluated weights: 0.0067–0.0519 on Color-MNIST and 0.0055–0.0387 on CelebA. This directly assesses agreement with labeled $P(X, Y)$, complementing the stable reconstruction MSE and NLL. Together, these diagnostics support observational fit as interventional predictions vary, without establishing equality of the full observational distributions or guaranteeing admissibility for arbitrary γ.

   For MIMIC-CXR-JPG, we compare observational fit with the $\gamma = 0$ baseline alongside the change in FID against the unconfounded reference. This checks whether gains in interventional image-distribution agreement come with a loss of observational fit; FID against that reference evaluates the interventional samples and does not itself measure observational admissibility.

2. **What happens when $\gamma = 0$?**
   The entropy term drops out, and CAUVADE optimizes only D (Eq. (6)). In the ablation of Sec. E.5.1, its interventional estimates concentrate tightly across random seeds, exhibiting behavior similar to ANCM (Fig. 13 and Table 8). These estimates lie within the scalar Manski bounds but explore little of the variation compatible with the observations. Sweeping γ supplies an additional means of obtaining diverse candidate interventional predictions; it does not guarantee exhaustive coverage or sharp bounds.

3. **How should K be chosen to obtain diverse hypotheses?**
   K controls the capacity of the latent confounder mixture. Increasing it can support more diverse causal hypotheses: in Table 9, increasing K from 2 to 4 expands the sampled interventional range from 0.167 to 0.352 for $do(X{=}0)$ and from 0.171 to 0.283 for $do(X{=}1)$. Realized diversity need not increase monotonically with K, however; the reported ranges narrow again at $K = 6$.

   Our goal is to generate diverse hypotheses that fit the observations, without requiring an exhaustive characterization of the compatible set. Thus, K need not recover the true confounder cardinality or be large enough to represent every compatible model. Choosing fewer clusters is acceptable when the fitted family remains sufficiently diverse and passes observational checks. Too little capacity can still restrict fit and diversity: at $K = 2$, only $14/28$ estimates per intervention pass the scalar

bounds check (Table 9). In practice, compare a small set of cluster counts using observational fit and variation in the resulting interventional predictions; increase K if capacity limits either. We use $K = 8$ in the main experiments (Table 3).

4. **Are U and $P(U)$ identifiable from observational data?**
In general, neither the latent confounder representation U nor its distribution $P(U)$ is identifiable from observational data under our assumptions. We explicitly acknowledge this structural uncertainty; it is the starting point of our work. Inspired by partial identification, CAUVADE explores plausible confounder distributions within models that pass observational checks and uses them to generate diverse interventional image predictions. Our objective is not to recover the true latent representation U or exogenous distribution $P(U)$. The synthetic experiment in Sec. E.5.3 illustrates this non-identifiability: even though the injected confounder distribution is known by construction, it is not identified by the observational data, and both CAUVADE and the ANCM baseline learn distributions that differ from it. The comparison highlights this shared non-identifiable nature; recovering the ground-truth confounder has never been the goal of our method. Instead, our method explores plausible exogenous distributions $P(U)$ compatible with the confounded data and uses them to generate diverse predictions of images under intervention.