GENERATING CHEST X-RAY COUNTERFACTUALS BY SPECIALISING FOUNDATION IMAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual image generation answers questions about how a subject would have looked under retrospective, hypothetical scenarios. Recent methods have improved perceptual quality, identity preservation and faithfulness to an underlying causal model, but their adoption in healthcare is limited by scarce annotated data, distribution shift between datasets, and mismatches between pretrained generative models and those required for counterfactual inference. We propose ***specialisation***, a data and parameter-efficient framework for adapting pretrained, non-causal generative models into causal mechanisms under distribution shift. Based on this framework, we train a radiology counterfactual image generation model, called RadCF, using latent flow matching. We validate our approach on three chest X-ray datasets spanning different dataset shifts, data volumes, and counterfactual questions, associated with challenging, highly-localised interventions. Our results show that RadCF and specialisation improve counterfactual soundness over existing methods while being data and parameter efficient, and that the resulting counterfactuals can detect and mitigate shortcut learning in a downstream medical classifier. Code available here.

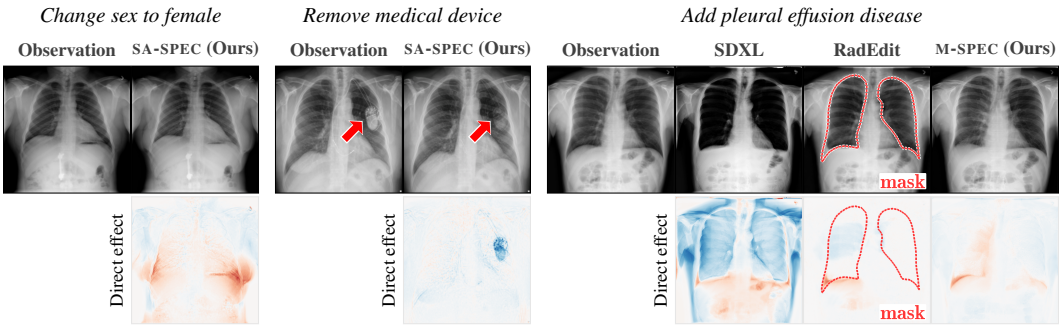
1 INTRODUCTION

Advances in generative modelling have improved faithfulness and control in medical image synthesis, particularly for chest X-rays (Bluethgen et al., 2024; Pérez-García et al., 2025a; Pinaya et al., 2023; Trang et al., 2025; Weber et al., 2023). However, these models often struggle with fine-grained, *causal* edits, as they implicitly learn broad statistical associations between dense conditioning signals and images. In the healthcare domain, models require an understanding of how genetic, environmental, or lifestyle factors interact to generate observations (Bareinboim et al., 2022; Peters et al., 2017; Schölkopf* et al., 2021). This is particularly important when reasoning about counterfactual scenarios in patients, e.g., *"What would this scan look like in different imaging conditions?"* or *"How would a disease have progressed in this individual under alternative treatments?"*. Answering such questions requires generative models that can perform isolated edits while preserving patient-specific anatomy.

Pearl (2009) formalises counterfactual reasoning using structural causal models (SCMs), in which variables are generated by a set of functional assignments, or *mechanisms*, forming a directed acyclic graph. Counterfactual inference is then performed via a simple three-step procedure: *abduction-action-prediction*. Pawlowski et al. (2020) implement SCMs with deep generative models to generate counterfactual images, and, in the absence of ground-truth counterfactuals, Monteiro et al. (2023) evaluate counterfactual soundness using metrics grounded in the axiomatic properties of mechanisms (Galles & Pearl, 1998; Halpern, 2000). Building on this, hierarchical VAEs (HVAEs, Child, 2020; Vahdat & Kautz, 2020), diffusion models (Preechakul et al., 2022; Rombach et al., 2022; Song et al., 2021a), and optimal-transport flow matching (Lipman et al., 2023; Pooladian et al., 2023; Tong et al., 2024) have been used to parameterise mechanisms for generating high-fidelity counterfactuals in mammography, 2D and 3D brain MRIs, subcortical meshes and chest X-rays (De Sousa Ribeiro et al., 2023; Peng et al., 2025; Rasal et al., 2022; 2025; Ribeiro et al., 2025; Xia et al., 2025a). However, deploying these models in clinical settings is limited by access to large, annotated datasets (Kaissis et al., 2020; Willemink et al., 2020). As such, bespoke training is impractical, motivating methods that adapt pretrained models for counterfactual inference in low-data regimes.

A natural alternative is image editing with pretrained text-to-image foundation models. This is analogous to counterfactual inference (Pan & Bareinboim, 2024; 2025; Zečević et al., 2022) and

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) Low-data SA-SPEC sex intervention (10%). **(b) BRAX device removal via specialisation.** **(c) M-SPEC specialises a pretrained text-to-image backbone without requiring masks.**

Figure 1: **Specialisation enables principled counterfactual generation across three clinical setups, adapting across distribution shift. (a)** SA-SPEC produces controlled sex interventions even with only 10% of target-domain training data. **(b)** SA-SPEC can introduce interventions not modelled during pretraining, like device. **(c)** M-SPEC adds explicit causal control to a pretrained text-to-image backbone, whereas prior work such as SDXL (Podell et al., 2023) alters the image globally and RadEdit (Pérez-García et al., 2025a) requires a spatial mask.

is increasingly applied to medical image editing (Cooke et al., 2025; Gu et al., 2023; Kumar et al., 2025; Pérez-García et al., 2025a). These methods control images through text prompts, often with spatial annotations such as masks or bounding boxes to localise edits (Couairon et al., 2023; Lin et al., 2024). However, prompt engineering requires significant expertise, and spatial annotations are often unavailable in clinical practice (Ma et al., 2024a; Wang et al., 2021; Xing et al., 2023). Moreover, masking cannot capture global morphological variations required for edits on attributes such as view, age, or anatomy. While recent work introduces causal structure into foundation models via adapters (Tong et al., 2026), the resulting approach ignores distribution shifts and biases that medical images are often subject to (Castro et al., 2020; Roschewitz et al., 2023).

We argue that widespread adoption of counterfactual reasoning in the healthcare domain requires adapting foundation models to operate under distribution shift and limited data, as in medical image generation (Bluethgen et al., 2024). In this work, we propose ***specialisation***, a framework for adapting pretrained generators to parameterise causal mechanisms under distribution shift in low-data regimes. We then train RadCF, our method for counterfactual image generation in radiology, and study three specialisation settings (Figure 1). Our contributions are as follows:

- We propose *specialisation*, a data and parameter-efficient framework that adapts pretrained, non-causal generative models into counterfactual image mechanisms (Section 3.2), by matching structural, acquisition and measurement mismatches in medical imaging settings.
- We introduce RadCF for counterfactual inference with latent flow matching (Section 3.1), improving counterfactual image fidelity over prior work whilst retaining competitive axiomatic soundness on radiology datasets.
- We show that *specialisation* extends across dataset shifts (Section 4.1) and generative model classes (Section 4.2), and that the resulting counterfactuals can detect and mitigate shortcut learning in a downstream classifier (Section 4.3).

Relation to prior work (Appendix A). We build on three areas: (1) Deep generative SCMs use VAEs, diffusion models, or flow-based mechanisms for counterfactual image generation(De Sousa Ribeiro et al., 2023; Peng et al., 2025; Rasal et al., 2022; 2025; Ribeiro et al., 2025; Xia et al., 2025a), but are typically trained from scratch for each target domain. This is impractical when large annotated clinical datasets are difficult to obtain (Kaissis et al., 2020; Willemink et al., 2020). Unlike image-space flows (Ribeiro et al., 2025), we operate in latent space for high-resolution generation. (2) Generative foundation models provide high-quality edits through prompts, masks, or bounding boxes (Cooke et al., 2025; Gu et al., 2023; Kumar et al., 2025; Pérez-García et al., 2025a), but offer associative rather than interventional control over structured parent variables. (3) Methods for adding conditioning and efficiently adapting pretrained generators, including ControlNet and LoRA (Hu et al., 2022; Zhang et al., 2023), provide the technical tools for specialisation. Causal-Adapter (Tong et al., 2026) adapts a frozen text-to-image backbone for causal inference, but does not study distribution shift.

2 BACKGROUND

Our work generates counterfactuals following the methodology in Pearl (2009) and is implemented using flow matching models (Lipman et al., 2023; Ma et al., 2024b). We briefly review these topics.

2.1 COUNTERFACTUAL INFERENCE

Pearl (2009) introduces structural causal models (SCMs) to describe causal relationships and compute causal quantities within a system. We consider acyclic Markovian SCMs $\mathfrak{S} := (\mathbf{U}, \mathbf{X}, \mathbf{F})$, with a set of exogenous variables $\mathbf{U} = \{\mathbf{u}_k\}_{k=1}^K$ which are mutually independent, a set of observed variables $\mathbf{X} = \{\mathbf{x}_k\}_{k=1}^K$, and a set of deterministic functions $\mathbf{F} = \{f_k\}_{k=1}^K$, called *mechanisms*, defining causal relationships via acyclic assignments between an $\mathbf{x}_k$ and its direct causes, or *parents*, $\mathbf{pa}_k$,

$$\mathbf{x}_k = f_k(\mathbf{pa}_k, \mathbf{u}_k), \qquad \mathbf{pa}_k \subseteq \mathbf{X} \setminus \{\mathbf{x}_k\}, \qquad k \in \{1, \dots, K\}. \tag{1}$$

Counterfactual inference follows a three-step procedure (Pearl, 2009): *abduction* infers the exogenous posterior given the factual observations; *action* replaces the structural assignment of the intervened variable, denoted $\mathrm{do}(\cdot)$; and *prediction* evaluates the modified SCM at the counterfactual parents $\widetilde{\mathbf{pa}}_k$, retaining the same abducted exogenous states. For mechanisms that are bijective in $\mathbf{u}_k$ at fixed $\mathbf{pa}_k$, abduction reduces to inversion, $\mathbf{u}_k = f_k^{-1}(\mathbf{x}_k, \mathbf{pa}_k)$. For non-invertible mechanisms, it instead requires the posterior over exogenous states compatible with the factual observation.

Counterfactuals from correct SCMs satisfy axiomatic properties (Galles & Pearl, 1998; Halpern, 2000). Monteiro et al. (2023) operationalise these as metrics for *composition*, *reversibility*, and *effectiveness*. *Composition* measures consistency under null interventions ($\mathrm{do}(\mathbf{pa})$) and *reversibility* evaluates cycle-consistency ($\mathrm{do}(\widetilde{\mathbf{pa}}) \to \mathrm{do}(\mathbf{pa})$), scored by LPIPS (Zhang et al., 2018). Together, they indicate a mechanism's ability to preserve identity in counterfactuals. Finally, *effectiveness* measures faithfulness to interventions via pseudo-oracle classifiers $p(\mathbf{pa}|\mathbf{x})$ to verify that target attributes change and non-target attributes are preserved (e.g., sex, age), scored by accuracy and MAE. This evaluation strategy is standard for counterfactuals (Rasal et al., 2025; Ribeiro et al., 2025; Roschewitz et al., 2023); all implementation details are in Appendix B. Melistas et al. (2025) further propose *realism* for counterfactual image fidelity, which we score by KID (Bińkowski et al., 2018).

2.2 FLOW MATCHING

Flow matching models (Lipman et al., 2023) learn transport dynamics from a base to a data distribution to generate samples. Ma et al. (2024b) define a linear probability path $\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\mathbf{z}_1$, $t \in [0, 1]$, in the latent space of a VAE (Esser et al., 2020; Rombach et al., 2022), interpolating between the data encoding $\mathbf{z}_0 = E_{\phi_1}(\mathbf{x}_0)$ at $t = 0$ and noise $\mathbf{z}_1 \sim \mathcal{N}(0, \mathbf{I})$ at $t = 1$, with decoder $\hat{\mathbf{x}}_0 = D_{\phi_2}(\mathbf{z}_0)$. Its time derivative is $\mathbf{z}_1 - \mathbf{z}_0$, which the conditional velocity field $\mathbf{v}_\theta(\mathbf{z}_t, t, \mathbf{c})$ is trained to approximate,

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1],\mathbf{z}_1\sim\mathcal{N}(0,\mathbf{I}),\mathbf{z}_0,\mathbf{c}} \left\| \mathbf{v}_\theta(\mathbf{z}_t, t, \mathbf{c}) - (\mathbf{z}_1 - \mathbf{z}_0) \right\|_2^2. \tag{2}$$

When the encoder and decoder are trained independently of the velocity field $\mathbf{v}_\theta(\cdot)$ (Peebles & Xie, 2022; Rombach et al., 2022), mismatches between the latent representation and the learned transport can degrade generative quality. The REPA-E framework (Leng et al., 2025) jointly optimises the velocity field θ and the VAE parameters $\phi = \{\phi_1, \phi_2\}$, aligning them to DINOv2 representations (Oquab et al., 2024) for efficient training (Yu et al., 2025). The full objective is

$$\mathcal{L}_{\text{REPA-E}}(\theta, \phi, \xi) = \mathcal{L}_{\text{FM}}(\theta) + \lambda \mathcal{L}_{\text{align}}(\theta, \phi, \xi) + \eta \mathcal{L}_{\text{reg}}(\phi), \tag{3}$$

where $\mathcal{L}_{\text{align}}(\cdot)$ is the alignment term, $\mathcal{L}_{\text{reg}}(\cdot)$ learns the VAE with perceptual losses and a KL-divergence over the prior, and coefficients λ and η are hyperparameters. Alignment matches transport features to frozen DINOv2 features (Oquab et al., 2024) via a projection head with parameters ξ.

We solve the probability-flow ODE $\frac{d\mathbf{z}_t}{dt} = \mathbf{v}_\theta(\mathbf{z}_t, t, \mathbf{c})$ backward in time for generation ($\mathbf{z}_0$) and forward in time for inversion ($\mathbf{z}_1$),

$$\mathbf{z}_0 := T_\theta(\mathbf{z}_1, \mathbf{c}) = \mathbf{z}_1 - \int_0^1 \mathbf{v}_\theta(\mathbf{z}_t, t, \mathbf{c})\, dt, \qquad \mathbf{z}_1 := T_\theta^{-1}(\mathbf{z}_0, \mathbf{c}) = \mathbf{z}_0 + \int_0^1 \mathbf{v}_\theta(\mathbf{z}_t, t, \mathbf{c})\, dt. \tag{4}$$

where $T_\theta(\cdot)$ defines a bijective flow under regularity assumptions on $\mathbf{v}_\theta(\cdot)$ (Albergo et al., 2023; Chen et al., 2018; Ribeiro et al., 2025), commonly implemented via the Euler sampler (Chen et al., 2018; Lipman et al., 2023; Ma et al., 2024b; Song et al., 2021b).

3 METHOD: SPECIALISATION FOR IMAGE MECHANISMS

In medical settings, imaging datasets are typically small and cannot easily be pooled due to privacy concerns (Kaissis et al., 2020; Willemink et al., 2020). Since many existing counterfactual generative methods are trained from scratch, these data challenges limit their use in the healthcare domain. Additionally, observed distributions vary due to acquisition shifts in the image arising from differences in scanners and imaging protocols, and measurement bias in its parents where observed labels act as imperfect proxies for underlying clinical factors (Castro et al., 2020; Roschewitz et al., 2023). This limits cross-domain generalisability of pretrained models.

To address these gaps, we introduce *specialisation*, a parameter-efficient procedure that adapts a pretrained generative model, in the source domain s, into an image mechanism, in the target domain t. To that end, we first outline our image mechanisms, parameterised by a flow-matching model in the latent space of a VAE, in Section 3.1. We then describe three forms of specialisation addressing structural, acquisition, and measurement mismatch in medical imaging in Section 3.2.

3.1 RADCF: LATENT FLOW-BASED IMAGE MECHANISM

Our image mechanism $f(\cdot)$ is parameterised by the reverse latent flow in Equation (4) for generation,

$$\mathbf{x} := f_\psi(\mathbf{u}, \mathbf{pa}) \approx D_{\phi_2}(\mathbf{z}_0), \quad \mathbf{z}_0 = T_\theta(\mathbf{u}, \mathbf{pa}) \tag{5}$$

where $\mathbf{x}$ is the image, $\mathbf{pa}$ are its parents and $p(\mathbf{u}) = p(\mathbf{z}_1) = \mathcal{N}(0, \mathbf{I})$ defines the exogenous prior as the base distribution of the flow. Abduction is performed via the forward latent flow in Equation (4),

$$\mathbf{u} := f_\psi^{-1}(\mathbf{x}, \mathbf{pa}) \approx T_\theta^{-1}(\mathbf{z}_0, \mathbf{pa}), \quad \mathbf{z}_0 = E_{\phi_1}(\mathbf{x}_0). \tag{6}$$

The parameters ψ is the set $\{\phi_1, \phi_2, \theta\}$. A counterfactual image $\widetilde{\mathbf{x}}$ is then generated as

$$\widetilde{\mathbf{x}} = f_\psi(\mathbf{u}, \widetilde{\mathbf{pa}}) \tag{7}$$

where $\widetilde{\mathbf{pa}}$ is the set of counterfactual parents. The parents $\mathbf{pa}$ are passed into $T_\theta(\cdot)$ via a learnable embedding $g_\omega(\mathbf{pa}) = \sum_{\mathbf{pa}_k \in \mathbf{pa}} g_{\omega_k}(\mathbf{pa}_k)$, detailed in Appendix D.2, which is added to the timestep embedding akin to the metadata conditioning pattern in Khanna et al. (2024). It follows that $T_\theta(\cdot)$ can be learned on observational data using the REPA-E objective in Equation (3) with $\mathbf{c} = \mathbf{pa}$. Compared to the image-space *spatial mechanisms* of Rasal et al. (2025), operating in latent space enables higher-resolution, higher-fidelity counterfactuals with amortised compute.

Identity preservation. We assess identity preservation through composition (reconstruction) and reversibility (cycle-consistency) (Monteiro et al., 2023; Rasal et al., 2025). Our mechanism is parameterised by $T_\theta(\cdot)$, a continuous-time PF-ODE which is bijective under standard assumptions (Albergo et al., 2023; Chen et al., 2018). However, this bijectivity holds only in latent space $\mathbf{z}_0$: information lost by $E_{\phi_1}(\cdot)$ cannot be recovered by $T_\theta(\cdot)$ or $D_{\phi_2}(\cdot)$. To account for this, REPA-E jointly learns the VAE (ϕ) and flow parameters (θ), reducing autoencoder reconstruction error (Appendix E.2), thereby retaining composition and reversibility (Table 1).

Effectiveness and realism. Existing flow-based mechanisms operate in image-space (Ribeiro et al., 2025). We instead perform transport in the compressed latent space $\mathbf{z}_0$. The parents $\mathbf{pa}$ condition the latent transport via $T_\theta(\cdot)$, providing the explicit control needed for effectiveness (intervention faithfulness). The decoder $D_{\phi_2}(\cdot)$ maps the transported latent back to a high-resolution image. Thus, performing transport in latent space reduces the computational burden relative to image-space flows, facilitating efficient high-resolution counterfactual generation (Rombach et al., 2022). The REPA-E objective also adapts the VAE to our radiology data, further improving realism (image fidelity).

3.2 SPECIALISATION

We now describe how pretrained generators can be specialised into image mechanisms for counterfactual inference. The required adaptations account for differences in distributions and conditioning requirements. We consider a generator pretrained on $p_s(\mathbf{x})$ or $p_s(\mathbf{x} \mid \mathbf{c}_s)$, where $\mathbf{c}_s$ denotes the conditioning variables from the source domain s. The target distribution is $p_t(\mathbf{x}, \mathbf{pa})$, where $\mathbf{pa}$ denotes the image parents specified by an SCM. We believe that source pretraining learns a reusable image prior from $p_s(\mathbf{x})$ or $p_s(\mathbf{x} \mid \text{TEXT})$, reducing what must be learned from limited target-domain

data. Specialisation adapts this prior for counterfactual inference by adding explicit dependence on $\mathbf{pa}$, tackling acquisition shift between $p_t(\mathbf{x})$ and $p_s(\mathbf{x})$, or replacing noisy text proxies of $\mathbf{pa}$ with a structured parent embedding. Our theoretical analysis in Appendix C provides further intuition for specialisation under idealised assumptions.

Structural specialisation (S-SPEC). Unconditional generative models, $\mathbf{x} = T_{\theta'}(\mathbf{u})$, implicitly encode information about $\mathbf{pa}$ over noise levels or architectural components (Burgess et al., 2018; Parmar et al., 2023; Tinaz et al., 2025; Wang & Vastola, 2023; Yu et al., 2025). However, counterfactual inference requires explicit control of $\mathbf{pa}$. S-SPEC uses flow matching to adapt a pretrained velocity field $\mathbf{v}_{\theta'}(\mathbf{z}_t, t)$ to use $\mathbf{pa}$ via Low-Rank Adaptation (LoRA, Hu et al., 2022):

$$\mathcal{L}_{\text{S-SPEC}}(\Delta\theta', \omega) = \mathbb{E}_{t,\boldsymbol{\epsilon},\mathbf{z}_0,\mathbf{pa}} \left\| \mathbf{v}_{\theta'+\Delta\theta'}(\mathbf{z}_t, t, g_\omega(\mathbf{pa})) - (\boldsymbol{\epsilon} - \mathbf{z}_0) \right\|_2^2. \tag{8}$$

This yields the conditional model $T_\theta(\mathbf{u}, \mathbf{pa})$ with $\theta = \theta' + \Delta\theta' \in \mathbb{R}^{d\times k}$ that parameterises our mechanism, where $\Delta\theta'$ applies to all linear layers in $\mathbf{v}_{\theta'}(\cdot)$. As per the definition of LoRA, $\Delta\theta' = BA$ where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$ and $r \ll \min(d, k)$, making S-SPEC data and parameter efficient. While prior works only adapt conditioning layers (Tong et al., 2026; Zhang et al., 2023), S-SPEC adapts all linear layers.

Structural, acquisition specialisation (SA-SPEC). Our mechanism decouples components responsible for composition (reconstruction) and effectiveness (intervention faithfulness); domain-specific pixel-level details are captured by a VAE and semantic control is governed by the latent flow. SA-SPEC jointly adapts the VAE parameters $\phi = \{\phi_1, \phi_2\}$ and the flow parameters θ' by modifying the REPA-E objective in Equation (3) as

$$\mathcal{L}_{\text{SA-SPEC}}(\Delta\theta', \phi, \omega) = \mathcal{L}_{\text{S-SPEC}}(\Delta\theta') + \lambda\mathcal{L}_{\text{align}}(\Delta\theta', \phi, \omega) + \eta\mathcal{L}_{\text{reg}}(\phi). \tag{9}$$

This formulation aligns our mechanism with the target domain, aiming to improve effectiveness while maintaining composition. It is specifically designed to address acquisition shift across source and target domains (Castro et al., 2020; Roschewitz et al., 2023).

Measurement specialisation (M-SPEC). Text-conditional models $T_{\theta'}(\mathbf{u}, \text{TEXT})$ in medical imaging implicitly encode $\mathbf{pa}$ via text (Cooke et al., 2025; Kumar et al., 2025; Pérez-García et al., 2025a). This introduces measurement bias (Hernán & Cole, 2009; Shahar, 2009) since text may be an inaccurate or incomplete representation of $\mathbf{pa}$ (Castro et al., 2020). As a result, these methods require prompt engineering or auxiliary information such as masks to perform localised interventions that follow the SCM. In contrast, structured metadata such as disease labels, view position, and demographics provide either observed proxies or direct measurements of the variables in $\mathbf{pa}$, and have been used to model $\mathbf{pa}$ in prior works (De Sousa Ribeiro et al., 2023; Mehta et al., 2025; Roschewitz et al., 2024; Xia et al., 2025b). To obtain explicit control over $\mathbf{pa}$ for text-conditioned models, we introduce M-SPEC, which replaces text conditioning with $g_\omega(\mathbf{pa})$ using the objective in Equation (8). Since the velocity field $\mathbf{v}_{\theta'}(\cdot)$ can be reparameterised as $\boldsymbol{\epsilon}_{\theta'}(\mathbf{z}_t, t, \mathbf{pa}) = \sigma_t\mathbf{z}_t + \alpha_t\mathbf{v}_{\theta'}(\mathbf{z}_t, t, \mathbf{pa})$, M-SPEC generalises to latent diffusion models,

$$\mathcal{L}_{\text{M-SPEC}}(\Delta\theta', \omega) = \mathbb{E}_{t,\boldsymbol{\epsilon},\mathbf{z}_0,\mathbf{pa}} \left\| \boldsymbol{\epsilon}_{\theta'+\Delta\theta'}(\mathbf{z}_t, t, g_\omega(\mathbf{pa})) - \boldsymbol{\epsilon} \right\|_2^2. \tag{10}$$

4 RESULTS

We evaluate specialisation as a data and parameter-efficient adaptation strategy (Section 4.1), M-SPEC as an extension to text-conditioned editors (Section 4.2) and demonstrate the utility of our counterfactuals on a downstream use case to complement the axiomatic evaluations (Section 4.3).

Unless otherwise stated, experiments specialise a CheXpert-pretrained model (Irvin et al., 2019) to NIH-14 (Wang et al., 2017) using SA-SPEC. Across all datasets we extract four parent variables from structured metadata: view position $v \in \{\text{AP}, \text{PA}\}$, sex $s \in \{\text{M}, \text{F}\}$, age $a \in [0, 1]$ and effusion status $e \in \{0, 1\}$. BRAX (Reis et al., 2022) additionally provides a device label $d \in \{0, 1\}$, and presents distinct acquisition and population shifts. See Appendix D for experiment details.

For demographic variables such as age and sex, $\text{do}(\cdot)$ denotes a counterfactual appearance query: "What would this radiograph look like if the patient were older / of a different sex?" rather than a physical clinical action. We acknowledge the ethical limitations of such interventions, although they are valuable for estimating counterfactual fairness (Kusner et al., 2018).

Table 1: **RadCF has the best realism and is competitive on all soundness metrics.** Each row block shows one intervention. Pseudo-oracles measure effectiveness for the intervened attribute and preservation of non-intervened attributes. KID is scaled by $\times 10^2$. Best values in each column for each intervention are bold.

		Effectiveness				Composition	Reversibility	Realism
Interv.	Method	Acc(v) ↑	Acc(e) ↑	Acc(s) ↑	MAE(a) ↓	LPIPS ↓	LPIPS ↓	KID ↓
do(v)	HVAE	0.976 (0.003)	0.706 (0.008)	0.734 (0.008)	0.085 (0.067)	0.045 (0.011)	0.091 (0.050)	7.01 (0.62)
	Diffusion	0.957 (0.004)	0.759 (0.008)	**0.953** (0.004)	0.073 (0.060)	0.032 (0.032)	0.067 (0.054)	3.02 (0.55)
	Image-space flow	0.984 (0.002)	**0.762** (0.008)	0.931 (0.005)	0.076 (0.061)	**0.011** (0.016)	**0.048** (0.039)	1.98 (0.58)
	RadCF	**0.996** (0.001)	0.726 (0.008)	0.948 (0.004)	**0.060** (0.049)	0.017 (0.029)	0.116 (0.091)	**0.53** (0.26)
do(e)	HVAE	0.971 (0.003)	**0.991** (0.002)	0.771 (0.007)	0.147 (0.104)	0.045 (0.011)	0.246 (0.029)	12.82 (1.28)
	Diffusion	**0.998** (0.001)	0.338 (0.008)	0.977 (0.003)	**0.056** (0.048)	0.032 (0.032)	0.058 (0.048)	1.61 (0.34)
	Image-space flow	0.996 (0.001)	0.497 (0.009)	**0.977** (0.003)	0.058 (0.050)	**0.011** (0.016)	**0.034** (0.029)	1.24 (0.41)
	RadCF	0.997 (0.001)	0.659 (0.009)	0.968 (0.003)	0.058 (0.049)	0.017 (0.029)	0.039 (0.067)	**0.83** (0.28)
do(s)	HVAE	0.895 (0.005)	0.665 (0.008)	**0.981** (0.002)	0.094 (0.074)	0.045 (0.011)	0.135 (0.037)	5.44 (0.70)
	Diffusion	0.986 (0.002)	0.854 (0.006)	0.743 (0.008)	0.062 (0.051)	0.032 (0.032)	0.072 (0.056)	1.71 (0.44)
	Image-space flow	**0.998** (0.001)	**0.887** (0.006)	0.818 (0.007)	0.060 (0.050)	**0.011** (0.016)	**0.037** (0.034)	0.92 (0.34)
	RadCF	0.997 (0.001)	0.807 (0.007)	0.964 (0.003)	**0.057** (0.047)	0.017 (0.029)	0.073 (0.081)	**0.48** (0.21)
do(a)	HVAE	**1.000** (0.000)	**0.958** (0.004)	**0.989** (0.002)	0.225 (0.132)	0.045 (0.011)	**0.000** (0.003)	0.74 (0.27)
	Diffusion	0.994 (0.001)	0.875 (0.006)	0.932 (0.004)	0.155 (0.095)	0.032 (0.030)	0.059 (0.049)	1.73 (0.34)
	Image-space flow	0.981 (0.002)	0.878 (0.006)	0.936 (0.004)	0.152 (0.099)	**0.011** (0.016)	0.036 (0.044)	1.01 (0.39)
	RadCF	0.986 (0.002)	0.827 (0.007)	0.908 (0.005)	**0.094** (0.073)	0.016 (0.028)	0.047 (0.097)	**0.68** (0.31)

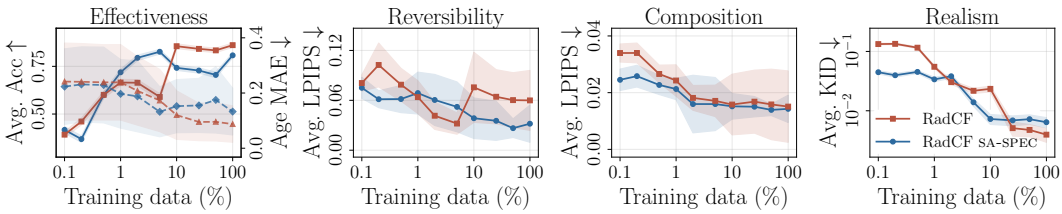


Figure 2: **Specialisation is data-efficient.** Trained on varying fractions of NIH-14 (1000 test samples). Solid lines average over the three categorical interventions; age is dashed and reported as MAE. Per-intervention results in Appendix F.

4.1 Specialisation under target-domain shift and data scarcity

Axiomatic soundness of RadCF. We first evaluate the soundness of our method against three mechanisms: an HVAE (De Sousa Ribeiro et al., 2023), a latent-diffusion implementation of the spatial mechanism of Rasal et al. (2025), and the image-space flow of Ribeiro et al. (2025). Effectiveness and identity preservation (the latter measured by composition and reversibility) often trade off in counterfactual image generation (Monteiro et al., 2023; Rasal et al., 2025). Table 1 shows that RadCF provides a strong balance across these metrics, serving as sound counterfactual image mechanism. Namely, RadCF achieves the best realism and competitive target effectiveness across all interventions. Compared with Diffusion, RadCF consistently improves composition and realism while maintaining comparable or stronger target control; the main exception is reversibility for do(v), where Diffusion is stronger. HVAE illustrates the complementary failure mode: high target success for interventions on categorical variables often reduces identity preservation, while high reversibility for do(a) corresponds to poor effectiveness on age interventions. The image-space flow of Ribeiro et al. (2025) attains lower composition and reversibility, consistent with its exact pixel-space invertibility as discussed in Section 3.1, but at a substantially higher KID (i.e. lower realism) and weaker effectiveness across interventions.

Data and parameter efficiency. Figure 2 plots counterfactual soundness as a function of target-domain data fraction (0.1%–100%, log scale), averaged across four interventions. Specialisation matches the from-scratch baseline using only 24M LoRA parameters on the flow matching component (vs. 766M from scratch); SA-SPEC additionally adapts the 84M-parameter VAE. SA-SPEC achieves the full-data view flip rate at only 5% of NIH-14 (0.993 vs. 0.998, while training from scratch at 5% reaches only 0.684; Appendix F). Empirically, specialisation improves composition and reversibility in low-data regimes. These results are consistent with Appendix C: source pretraining reduces the

Table 2: SA-SPEC **compensates for acquisition shift, improving composition and reversibility on the target domain (CheXpert→NIH-14) but not on the source domain.** For effectiveness, we report pseudo-oracle performance on the intervened attributes. Reversibility and realism are averages across all four interventions. KID is scaled by $\times 10^2$. Best per column metrics are in bold.

		Effectiveness				Composition	Reversibility	Realism
Domain	Method	do(v)↑	do(e)↑	do(s)↑	do(a)↓	LPIPS↓	LPIPS↓	KID↓
CheXpert →NIH-14	S-SPEC	0.976 (0.003)	0.359 (0.009)	0.884 (0.006)	0.170 (0.174)	0.092 (0.043)	0.163 (0.054)	3.13 (0.27)
	SA-SPEC	**0.992** (0.002)	**0.565** (0.009)	0.874 (0.006)	0.116 (0.079)	**0.015** (0.007)	**0.034** (0.021)	**0.97** (0.15)
	S-SPEC$_{\text{full}}$	0.971 (0.003)	0.549 (0.009)	**0.944** (0.004)	**0.111** (0.081)	0.106 (0.045)	0.243 (0.057)	1.83 (0.21)
CheXpert →CheXpert	S-SPEC	**0.965** (0.003)	**0.478** (0.010)	0.657 (0.007)	0.201 (0.139)	**0.024** (0.003)	0.037 (0.009)	1.34 (0.19)
	SA-SPEC	0.925 (0.004)	0.467 (0.010)	**0.688** (0.007)	**0.166** (0.102)	0.025 (0.004)	**0.027** (0.006)	**1.01** (0.19)

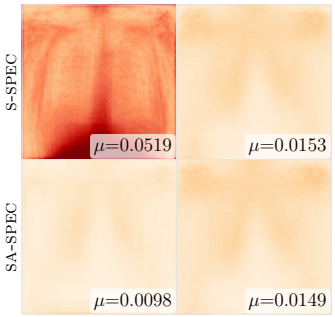


Figure 3: SA-SPEC **reduces acquisition-shift distortion.** Composition residuals for S-SPEC and SA-SPEC when a CheXpert-pretrained model is specialised to NIH-14 (out-of-domain) vs in-domain.

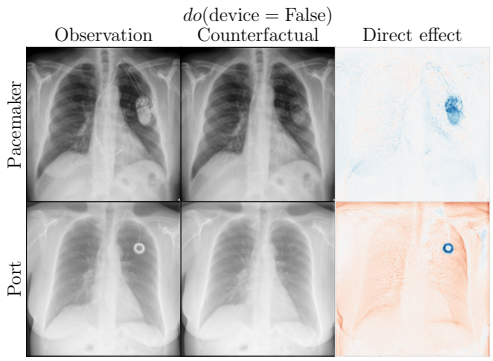


Figure 4: SA-SPEC **enables device-removal counterfactuals on BRAX.** Counterfactuals remove support devices under do(d) while preserving subject-specific anatomy.

marginal image-modelling burden, while specialisation learns the residual dependence on **pa**. At 100% data, specialisation retains a substantial composition and reversibility advantage at a modest cost to effectiveness on some interventions, a trade-off also observed by Rasal et al. (2025).

When is SA-SPEC **necessary?** Table 2 isolates the role of acquisition specialisation by comparing S-SPEC (flow-only) and SA-SPEC (flow + VAE) under CheXpert→NIH-14 acquisition shift and in-domain CheXpert→CheXpert adaptation. Under acquisition shift, SA-SPEC substantially improves composition and reversibility while preserving effectiveness for most attributes. In-domain, the same adaptation yields minimal change in composition and reversibility. Full fine-tuning of the flow alone (S-SPEC$_{\text{full}}$) does not recover the same gains, indicating that the improvement comes from acquisition specialisation rather than from additional trainable parameters. Figure 3 visualises this: under acquisition shift, S-SPEC produces large pixel residuals around the lungs even on null interventions, indicating systematic decoder distortion of target-domain anatomy. SA-SPEC largely eliminates these residuals, while in-domain both methods produce comparably small residuals.

Extending pa in a new target domain. We apply SA-SPEC to BRAX (Reis et al., 2022), adding device presence to the target-domain **pa** to enable do(d) counterfactuals. Device is a localised attribute correlated with view position and pathology, and BRAX differs distinctly from the source domain in acquisition and population. Despite this, composition and reversibility remain comparable to NIH-14 (0.018 and 0.019 LPIPS, against 0.015 and 0.034). Device interventions achieve a target accuracy of 0.729, and Figure 4 shows device removal preserving patient-specific anatomy. Per-intervention results are in Appendix G.

4.2 CAUSAL TEXT-TO-IMAGE GENERATORS WITH M-SPEC

Table 3 reports counterfactual soundness on 1,000 NIH-14 test samples. We compare our M-SPEC version of RadEdit against three methods: Causal-Adapter (Tong et al., 2026), SDXL, and mask-guided RadEdit for disease editing. Here, Causal-Adapter is applied to the same RadEdit backbone as M-SPEC, with results on the original Stable Diffusion 1.5 backbone in Appendix H.

Table 3: **Structured specialisation extends RadEdit to mask-free interventions on all variables in pa, with RadEdit M-SPEC outperforming Causal-Adapter on the same backbone.** Relative to Causal-Adapter, RadEdit M-SPEC achieves better target-control point estimates for every intervention and consistently lower composition error, reversibility error, and KID. SDXL and mask-guided RadEdit are included only for $\mathrm{do}(e)$; RadEdit additionally requires a localisation mask. KID is scaled by $\times 10^2$. Best per column metrics are in bold.

		Effectiveness				Composition	Reversibility	Realism
Intervention	Method	Acc(v) ↑	Acc(e) ↑	Acc(s) ↑	MAE(a) ↓	LPIPS ↓	LPIPS ↓	KID ↓
$\mathrm{do}(v)$	RadEdit Causal-Adapter	0.563 (0.031)	0.658 (0.029)	**0.963** (0.012)	0.073 (0.061)	0.094 (0.072)	0.278 (0.091)	4.64 (0.61)
	RadEdit M-SPEC	**0.600** (0.030)	**0.670** (0.029)	0.947 (0.014)	**0.067** (0.056)	**0.061** (0.057)	**0.202** (0.067)	**2.69** (0.50)
$\mathrm{do}(e)$	SDXL	0.748 (0.027)	**0.776** (0.027)	0.680 (0.029)	0.150 (0.115)	**0.015** (0.003)	0.488 (0.099)	3.54 (0.39)
	RadEdit	0.995 (0.004)	0.566 (0.031)	0.972 (0.010)	0.070 (0.056)	0.027 (0.015)	**0.107** (0.050)	**1.08** (0.24)
	RadEdit Causal-Adapter	0.979 (0.009)	0.531 (0.032)	0.974 (0.010)	0.065 (0.058)	0.094 (0.072)	0.223 (0.108)	1.85 (0.42)
	RadEdit M-SPEC	**1.000** (0.000)	0.682 (0.030)	**0.976** (0.009)	**0.053** (0.043)	0.061 (0.057)	0.174 (0.070)	1.40 (0.39)
$\mathrm{do}(s)$	RadEdit Causal-Adapter	0.986 (0.007)	0.768 (0.026)	0.160 (0.023)	0.063 (0.054)	0.094 (0.072)	0.186 (0.106)	1.64 (0.45)
	RadEdit M-SPEC	**0.997** (0.003)	**0.839** (0.023)	**0.547** (0.031)	**0.048** (0.039)	**0.061** (0.057)	**0.150** (0.070)	**0.75** (0.26)
$\mathrm{do}(a)$	RadEdit Causal-Adapter	0.981 (0.008)	0.790 (0.025)	**0.976** (0.009)	0.200 (0.133)	0.094 (0.071)	0.213 (0.107)	1.67 (0.39)
	RadEdit M-SPEC	**0.993** (0.005)	**0.848** (0.022)	0.962 (0.012)	**0.191** (0.126)	**0.062** (0.058)	**0.143** (0.077)	**0.57** (0.22)

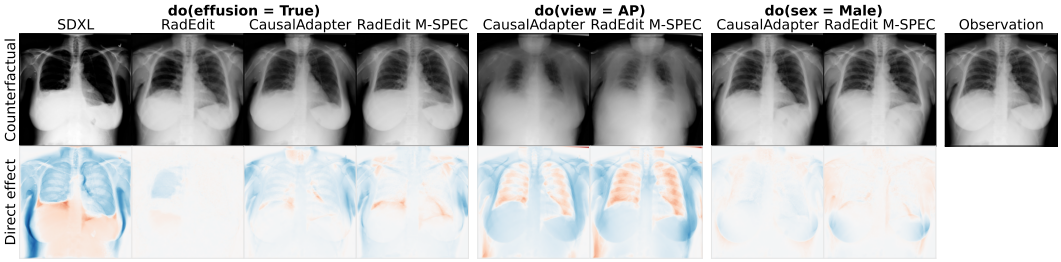


Figure 5: **Both structured specialisation methods generate sound counterfactuals without the need for masks.** Counterfactuals (top) and direct-effect maps (bottom) are shown for representative disease, view, and sex interventions. Compared with Causal-Adapter, RadEdit M-SPEC more reliably changes the intended attribute while limiting collateral changes to patient-specific anatomy. For disease editing, mask-guided RadEdit localises the intervention but requires a mask, whereas text-only SDXL introduces broader anatomical changes.

Our results demonstrate the benefit of structured specialisation. Under our taxonomy, both RadEdit M-SPEC and Causal-Adapter (Tong et al., 2026) are instances of M-SPEC: they both turn text-conditioned models into intervened causal variable conditions in $\mathbf{pa}$. For $\mathrm{do}(e)$, both M-SPEC instances preserve non-target attributes better than text-only SDXL without requiring a localisation mask, indicating more targeted counterfactual changes. RadEdit M-SPEC also improves effectiveness over RadEdit. This broader intervention interface comes with a trade-off: mask-guided RadEdit retains better composition, reversibility, and realism, as shown in Figure 5.

We next isolate the choice of specialisation mechanism. On the same RadEdit backbone, M-SPEC achieves better target-control point estimates for all four interventions and consistently lower composition error, reversibility error, and KID than Causal-Adapter. We hypothesise that this advantage arises from the more direct integration of structured variables into RadEdit's pretrained conditioning pathway: M-SPEC replaces text embeddings with metadata embeddings at the native conditioning interface, whereas Causal-Adapter first aligns causal attributes with textual tokens and injects their interactions through an additional adapter. This simpler, backbone-aligned parameterisation may make the structured conditions easier to optimise while better preserving the pretrained image distribution.

4.3 Practicality and Utility

While our approach improves on current deep generative SCM methods under data scarcity and domain shift, counterfactual inference requires strong assumptions. In particular, causal sufficiency is assumed: no unobserved confounders are taken into account, and only observed parents can be controlled for or intervened upon. This assumption underpins many prior works on counterfactual image generation (De Sousa Ribeiro et al., 2023; Monteiro et al., 2023; Pawlowski et al., 2020;

Rasal et al., 2025). Despite a likely violation in real data, counterfactual images may still be practically useful. We support this claim with additional analyses of their downstream utility, since the assumption cannot be tested directly from observational data (Pan & Bareinboim, 2024).

Stress-testing shortcut learning in classifiers. Counterfactuals can reveal whether a diagnostic model relies on shortcuts in the data to predict the disease status of a patient (Geirhos et al., 2020). If the test set has the same spurious correlation as the training set, performance on that test set may obscure shortcut learning; an unbiased test set can help reveal it Brown et al. (2023). We construct datasets at ten different levels of spurious correlation between the view and effusion variables, and train a classifier on each dataset. For each test image, we compare the classifier's prediction with its prediction on the corresponding view counterfactual. The fraction of predictions that change under $\mathrm{do}(v)$ rises from 11% at the weakest bias level to 47% at the strongest (Spearman's $\rho = 0.94$, $p < 0.05$). AUC on the original images ranges from 0.72 to 0.84 across the ten levels (Appendix I). This increasing sensitivity to view interventions is consistent with greater reliance on view features.

Debiasing classifiers via counterfactual augmentation. We test whether view counterfactuals reduce the classifiers' sensitivity to view when added to training data. At three correlation levels, we compare counterfactual augmentation with S-SPEC against minority upsampling and standard augmentations, matching training set size across strategies. At the strongest bias, the fraction of predictions that change under $\mathrm{do}(v)$ falls from 0.593 without augmentation to 0.115 with counterfactuals, compared to 0.468 for minority upsampling and 0.655 for standard augmentation (Appendix I).

Using our mechanisms within an SCM. Previous sections assess the effect of intervening on one parent at a time while the other parents remain fixed. A specialised mechanism is, however, not restricted to that setting and accepts parent values supplied by a causal graph. To show this, we incorporate our image mechanism into a deep SCM defined as $a \to e \to v$, $\{s, a, e, v\} \to \mathbf{x}$. The edge $e \to v$ encodes a plausible acquisition mechanism: patients with more severe effusion are less mobile and are therefore more often imaged supine (AP) rather than standing (PA). We compare two $\mathrm{do}(e)$ edits: one holds the view fixed, while the other updates the view according to the SCM. In the latter case, our method achieves the same effusion effectiveness (0.827) as when we keep the view fixed, while changing the view reduces realism and reversibility (Appendix J).

5 Discussion and limitations

We show that pretrained generative models can be *specialised* into image mechanisms for counterfactual generation, accounting for common distribution shifts in medical imaging. Our central contribution is an examination of specialisation under limited data, acquisition shift, and measurement bias, addressing practical clinical challenges. RadCF implements this framework through latent flow matching, providing competitive counterfactual soundness and image fidelity on chest X-rays. Specialisation also applies beyond flow-based mechanisms. In particular, M-SPEC adapts a text-conditional diffusion model to enable structured interventions, without needing text prompts or auxiliary spatial annotations. In doing so, we take a step toward using foundation models for counterfactual reasoning, with applications in stress-testing and data augmentation.

Limitations and future work. While our experiments target measurement bias and acquisition shift, specialised mechanisms may still be affected by manifestation shift, where the same clinical factor induces different phenotypes across populations. We assume that our image mechanisms' exogenous noise is independent of its parents. This assumption may be violated by unobserved confounding in real-world clinical data. Appendices K and L discuss this limitation and provide stress tests under deliberately misspecified, hidden-confounding scenarios. Counterfactual identifiability has been established for non-Markovian SCMs under additional assumptions, including bijective image mechanisms (Nasr-Esfahany et al., 2023; Ribeiro et al., 2025). To scale to high-resolution counterfactual images, RadCF parameterises the mechanism in latent space and is not exactly invertible in image space. These identifiability guarantees therefore do not directly extend to our latent parameterisation. Future work could investigate conditions for counterfactual identifiability in latent image mechanisms, or explore high-resolution, image-space generative models (Li & He, 2026) as a basis for mechanisms that satisfy the assumptions of Ribeiro et al. (2025).

AI USE STATEMENT

Among the tasks for which disclosure is required, we used generative AI tools to assist with implementing analysis and plotting code, and to edit sections of the manuscript for readability. We did not use generative AI tools to propose or refine the research hypotheses, to design the method or the experimental protocol, or to draft the manuscript.

ETHICS STATEMENT

This work generates counterfactual chest radiographs under interventions on patient attributes, including sex and age. These are appearance queries rather than physical clinical actions. Counterfactuals that alter a diagnostic finding could be mistaken for real radiographs if separated from their provenance, so the mechanism is intended for model analysis rather than for clinical use. All experiments use the public, de-identified CheXpert, NIH-14 and BRAX datasets under their respective licences, and no new patient data were collected.

REPRODUCIBILITY STATEMENT

The specialisation objectives are defined in Section 3.2, and the mechanism they adapt in Section 3.1. Architecture, LoRA configuration, optimiser settings and training schedules are given in Appendix D.2, baseline implementations in Appendix D.3, and dataset construction and preprocessing in Appendix D.1. Evaluation is specified in Appendix B: pseudo-oracle training in Appendix B.1, the LPIPS and KID protocols in Appendix B.3 and Appendix B.4, and uncertainty reporting in Appendix B.5. Because effectiveness depends on the classifier used to measure it, Appendix B.2 repeats the evaluation under six independently trained evaluators. The biased-dataset construction behind the downstream experiments is tabulated in Appendix I, and the metadata SCM and its fitted parent mechanisms in Appendix J. Code is available at `https://anonymous.4open.science/r/RadCF-D7B5/`.

REFERENCES

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions, 2023. URL https://arxiv.org/abs/2303.08797.

Shahab Aslani, Watjana Lilaonitkul, Vaishnavi Gnananthan, Divya Raj, Bojidar Rangelov, Alexandra L. Young, Yipeng Hu, Paul Taylor, Daniel C. Alexander, and Joseph Jacob. Optimising chest x-rays for image analysis by identifying and removing confounding factors. In Ruidan Su, Yudong Zhang, Han Liu, and Alejandro F Frangi (eds.), *Medical Imaging and Computer-Aided Diagnosis*, pp. 245–254, Singapore, 2023. Springer Nature Singapore. ISBN 978-981-16-6775-6.

Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. *On Pearl's Hierarchy and the Foundations of Causal Inference*, pp. 507–556. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022. ISBN 9781450395861. URL https://doi.org/10.1145/3501714.3501743.

Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1lUOzWCW.

Christian Bluethgen, Pierre Chambon, Jean-Benoit Delbrouck, Rogier van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P. Langlotz, and Akshay S. Chaudhari. A vision–language foundation model for the generation of realistic chest x-ray images. *Nature Biomedical Engineering*, Aug 2024. ISSN 2157-846X. doi: 10.1038/s41551-024-01246-y. URL https://doi.org/10.1038/s41551-024-01246-y.

Alexander Brown, Nenad Tomasev, Jan Freyberg, Yuan Liu, Alan Karthikesalingam, and Jessica Schrouff. Detecting shortcut learning for fair medical ai using shortcut testing. *Nature Communications*, 14(1):4314, 2023. doi: 10.1038/s41467-023-39902-7. URL https://doi.org/10.1038/s41467-023-39902-7.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*, 2018.

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

Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nat. Commun.*, 11(1):3673, July 2020.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 2018.

Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*, 2020.

Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. Torchxrayvision: A library of chest x-ray datasets and models. In Ender Konukoglu, Bjoern Menze, Archana Venkataraman, Christian Baumgartner, Qi Dou, and Shadi Albarqouni (eds.), *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, volume 172 of *Proceedings of Machine Learning Research*, pp. 231–249. PMLR, 06–08 Jul 2022. URL https://proceedings.mlr.press/v172/cohen22a.html.

Lauren H. Cooke, Matthias Jung, Jan M. Brendel, Nora M. Kerkovits, Borek Foldyna, Michael T Lu, and Vineet K Raghu. Roentmod: A synthetic chest x-ray modification model to identify image interpretation model shortcuts. In *Medical Imaging with Deep Learning - Short Papers*, 2025. URL https://openreview.net/forum?id=jrMCsWiAYq.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *International Conference on Learning Representations*, 2023.

Fabio De Sousa Ribeiro, Tian Xia, Miguel Monteiro, Nick Pawlowski, and Ben Glocker. High fidelity image counterfactuals with probabilistic causal models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020.

David Galles and Judea Pearl. An axiomatic characterization of causal counterfactuals. *Foundations of Science*, 3:151–182, 1998.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Yu Gu, Jianwei Yang, Naoto Usuyama, Chunyuan Li, Sheng Zhang, Matthew P Lungren, Jianfeng Gao, and Hoifung Poon. Biomedjourney: Counterfactual biomedical image generation by instruction-learning from multimodal patient journeys. *arXiv preprint arXiv:2310.10765*, 2023.

Joseph Y Halpern. Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research*, 12: 317–337, 2000.

Miguel A Hernán and Stephen R Cole. Invited commentary: Causal diagrams and measurement bias. *American Journal of Epidemiology*, 170(8):959–962, 2009. doi: 10.1093/aje/kwp293.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI Conference on Artificial Intelligence*, volume 33, pp. 590–597, 2019.

Georgios A. Kaissis, Marcus R. Makowski, Daniel Rückert, and Rickmer F. Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6): 305–311, 2020. doi: 10.1038/s42256-020-0186-1.

Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David B. Lobell, and Stefano Ermon. Diffusionsat: A generative foundation model for satellite imagery. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=I5webNFDgQ.

Amar Kumar, Anita Kriz, Mohammad Havaei, and Tal Arbel. PRISM: High-resolution & precise counterfactual medical image generation using language-guided stable diffusion. In *Medical Imaging with Deep Learning*, 2025. URL https://openreview.net/forum?id=UpJMAlZNuo.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness, 2018. URL https://arxiv.org/abs/1703.06856.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. *arXiv preprint arXiv:2504.10483*, 2025.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise, 2026. URL https://arxiv.org/abs/2511.13720.

Yuanze Lin, Yi-Wen Chen, Yi-Hsuan Tsai, Lu Jiang, and Ming-Hsuan Yang. Text-driven image editing via learnable regions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7059–7068, 2024.

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

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024a. doi: 10.1038/s41467-024-44824-z.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers, 2024b. URL https://arxiv.org/abs/2401.08740.

Raghav Mehta, Fabio De Sousa Ribeiro, Tian Xia, Mélanie Roschewitz, Ainkaran Santhirasekaram, Dominic C. Marshall, and Ben Glocker. CF-Seg: Counterfactuals meet Segmentation . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15967. Springer Nature Switzerland, September 2025.

Thomas Melistas, Nikos Spyrou, Nefeli Gkouti, Pedro Sanchez, Athanasios Vlontzos, Yannis Panagakis, Giorgos Papanastasiou, and Sotirios A. Tsaftaris. Benchmarking counterfactual image generation, 2025. URL https://arxiv.org/abs/2403.20287.

Miguel Monteiro, Fabio De Sousa Ribeiro, Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Measuring axiomatic soundness of counterfactual image models, 2023. URL https://arxiv.org/abs/2303.01274.

Arash Nasr-Esfahany, Mohammad Alizadeh, and Devavrat Shah. Counterfactual identifiability of bijective causal models, 2023. URL https://arxiv.org/abs/2302.02228.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models, 2019. URL https://arxiv.org/abs/1905.05824.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Yushu Pan and Elias Bareinboim. Counterfactual image editing. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39087–39101. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/pan24a.html.

Yushu Pan and Elias Bareinboim. Counterfactual image editing with disentangled causal latent space. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=u2Lgi4NIe7.

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH '23, New York, NY, USA, 2023. Association for Computing Machinery. doi: 10.1145/3588432.3591513. URL https://doi.org/10.1145/3588432.3591513.

Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In *Advances in Neural Information Processing Systems*, volume 33, pp. 857–869, 2020.

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

Wei Peng, Tian Xia, Fabio De Sousa Ribeiro, Tomas Bosschieter, Ehsan Adeli, Qingyu Zhao, Ben Glocker, and Kilian M. Pohl. Latent causal modeling for 3d brain mri counterfactuals. In *Deep Generative Models: 5th MICCAI Workshop, DGM4MICCAI 2025, Held in Conjunction with MICCAI 2025, Daejeon, South Korea, September 23, 2025, Proceedings*, pp. 192–201, Berlin, Heidelberg, 2025. Springer-Verlag. ISBN 978-3-032-05471-5. doi: 10.1007/978-3-032-05472-2_19. URL https://doi.org/10.1007/978-3-032-05472-2_19.

Fernando Pérez-García, Sam Bond-Taylor, Pedro P. Sanchez, Boris van Breugel, Daniel C. Castro, Harshita Sharma, Valentina Salvatelli, Maria T. A. Wetscherek, Hannah Richardson, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, Ozan Oktay, and Maximilian Ilse. RadEdit: Stress-testing biomedical vision models via diffusion image editing. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 358–376, Cham, 2025a. Springer Nature Switzerland. ISBN 978-3-031-73254-6.

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Teodora Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 2025b. doi: 10.1038/s42256-024-00965-w.

Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319.

Walter HL Pinaya, Mark S Graham, Eric Kerfoot, Petru-Daniel Tudosiu, Jessica Dafflon, Virginia Fernandez, Pedro Sanchez, Julia Wolleb, Pedro F da Costa, Ashay Patel, et al. Generative ai for medical imaging: extending the monai framework. *arXiv preprint arXiv:2307.15208*, 2023. URL https://arxiv.org/abs/2307.15208.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. URL https://arxiv.org/abs/2307.01952.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28100–28127. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/pooladian23a.html.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10619–10629, 2022.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017. URL https://arxiv.org/abs/1711.05225.

Suhail Raoof, David Feigin, Arthur Sung, Sabiha Raoof, Lavanya Irugulpati, and Edward C. Rosenow. Interpretation of plain chest roentgenogram. *Chest*, 141(2):545–558, February 2012. ISSN 1931-3543. doi: 10.1378/chest.10-1302.

Rajat Rasal, Daniel C Castro, Nick Pawlowski, and Ben Glocker. Deep structural causal shape models. In *European Conference on Computer Vision*, pp. 400–432. Springer, 2022.

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

Eduardo Pontes Reis, Joselisa Paiva, Maria Carolina Bueno da Silva, Guilherme Alberto Sousa Ribeiro, Victor Fornasiero Paiva, Lucas Bulgarelli, Henrique Lee, Paulo Victor dos Santos, vanessa brito, Lucas Amaral, Gabriel Beraldo, Jorge Nebhan Haidar Filho, Gustavo Teles, Gilberto Szarf, Tom Pollard, Alistair Johnson, Leo Anthony Celi, and Edson Amaro. BRAX, a Brazilian labeled chest X-ray dataset. *PhysioNet*, June 2022. doi: 10.13026/grwk-yh18. URL https://doi.org/10.13026/grwk-yh18. Version 1.1.0.

Fabio De Sousa Ribeiro, Ainkaran Santhirasekaram, and Ben Glocker. Counterfactual identifiability via dynamic optimal transport. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=h2ttG6HkID.

Jonathan G. Richens, Ciarán M. Lee, and Saurabh Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications*, 11(1):3923, 2020. doi: 10.1038/s41467-020-17419-7.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Mélanie Roschewitz, Galvin Khara, Joe Yearsley, Nisha Sharma, Jonathan J James, Éva Ambrózay, Adam Heroux, Peter Kecskemethy, Tobias Rijken, and Ben Glocker. Automatic correction of performance drift under acquisition shift in medical image classification. *Nature Communications*, 14(1):6584, 2023. doi: 10.1038/s41467-023-42396-y. URL https://doi.org/10.1038/s41467-023-42396-y.

Mélanie Roschewitz, Fabio de Sousa Ribeiro, Tian Xia, Galvin Khara, and Ben Glocker. Counterfactual contrastive learning: robust representations via causal image synthesis. In *MICCAI Workshop on Data Engineering in Medical Imaging*, pp. 22–32. Springer, 2024.

B. Schölkopf*, F. Locatello*, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9363924. *equal contribution.

Eyal Shahar. Causal diagrams for encoding and evaluation of information bias. *Journal of Evaluation in Clinical Practice*, 15(3):436–440, 2009. doi: 10.1111/j.1365-2753.2008.01031.x.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Berk Tinaz, Zalan Fabian, and Mahdi Soltanolkotabi. Emergence and evolution of interpretable concepts in diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=C7LxkebvVW.

Ekin Tiu, Ellie Talius, Pujan Patel, Curt P. Langlotz, Andrew Y. Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, 2022. doi: 10.1038/s41551-022-00936-9.

Alexander Tong, Kilian FATRAS, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=CD9Snc73AW. Expert Certification.

Lei Tong, Zhihua Liu, Chaochao Lu, Dino Oglic, Tom Diethe, Philip Teare, Sotirios A. Tsaftaris, and Chen Jin. Causal-adapter: Taming text-to-image diffusion for faithful counterfactual generation, 2026. URL https://arxiv.org/abs/2509.24798.

Bailey Trang, Parham Saremi, Alan Q. Wang, Fangrui Huang, Zahra TehraniNasab, Amar Kumar, Tal Arbel, Li Fei-Fei, and Ehsan Adeli. Discovering latent graphs with GFlownets for diverse conditional image generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=PhHrlDKcx1.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

Binxu Wang and John J Vastola. Diffusion models generate images like painters: an analytical theory of outline first, details later. *arXiv preprint arXiv:2303.02490*, 2023.

Shanshan Wang, Cheng Li, Rongpin Wang, Zaiyi Liu, Meiyun Wang, Hongna Tan, Yaping Wu, Xinfeng Liu, Hui Sun, Rui Yang, Xin Liu, Jie Chen, Huihui Zhou, Ismail Ben Ayed, and Hairong Zheng. Annotation-efficient deep learning for automatic medical image segmentation. *Nature Communications*, 12(1):5915, 2021. doi: 10.1038/s41467-021-26216-9.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2097–2106, 2017.

Tobias Weber, Michael Ingrisch, Bernd Bischl, and David Rügamer. Cascaded latent diffusion models for high-resolution chest x-ray synthesis. In *Advances in Knowledge Discovery and Data Mining: 27th Pacific-Asia Conference, PAKDD 2023*. Springer, 2023.

Martin J. Willemink, Wojciech A. Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R. Folio, Ronald M. Summers, Daniel L. Rubin, and Matthew P. Lungren. Preparing medical imaging data for machine learning. *Radiology*, 295(1):4–15, 2020. doi: 10.1148/radiol.2020192224.

Tian Xia, Fabio De Sousa Ribeiro, Rajat R Rasal, Avinash Kori, Raghav Mehta, and Ben Glocker. Decoupled classifier-free guidance for counterfactual diffusion models. *arXiv preprint arXiv:2506.14399*, 2025a. doi: 10.48550/arXiv.2506.14399. URL https://arxiv.org/abs/2506.14399.

Tian Xia, Matthew Sinclair, Andreas Schuh, Fabio De Sousa Ribeiro, Raghav Mehta, Rajat Rasal, Esther Puyol-Antón, Samuel Gerber, Kersten Petersen, Michiel Schaap, and Ben Glocker. Segmentor-guided Counterfactual Fine-Tuning for Locally Coherent and Targeted Image Synthesis . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15961. Springer Nature Switzerland, September 2025b.

Xiaodan Xing, Giorgos Papanastasiou, Simon Walsh, and Guang Yang. Less is more: Unsupervised mask-guided annotated ct image synthesis with minimum manual segmentations. *IEEE Transactions on Medical Imaging*, 42(9):2566–2576, 2023. doi: 10.1109/TMI.2023.3260169.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think, 2025. URL https://arxiv.org/abs/2410.06940.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Pearl causal hierarchy on image data: Intricacies & challenges. *arXiv e-prints*, pp. arXiv–2212, 2022.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824, 2023. doi: 10.1109/ICCV51070.2023.00355.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, Los Alamitos, CA, USA, June 2018. IEEE Computer Society. doi: 10.1109/CVPR.2018.00068. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00068.

A Related work

Counterfactual image generation. Counterfactual image generation aims to answer retrospective questions about how an observed instance would have appeared under an alternative intervention, while preserving all non-descendant, instance-specific factors. This is naturally formalised using structural causal models (SCMs), where counterfactuals are computed through abduction, action, and prediction (Pearl, 2009; Peters et al., 2017; Schölkopf* et al., 2021). Early deep SCMs parameterised image mechanisms with deep generative models, enabling counterfactual generation in settings where the image is treated as a high-dimensional endogenous variable (Pawlowski et al., 2020). Since ground-truth counterfactual images are generally unavailable, subsequent work has evaluated learned mechanisms through axiomatic criteria such as composition, reversibility, and effectiveness (Galles & Pearl, 1998; Halpern, 2000; Monteiro et al., 2023), with realism later incorporated as an additional fidelity criterion (Melistas et al., 2025). These metrics are now widely used to quantify whether generated counterfactuals preserve identity, obey interventions, and remain on the image manifold.

Building on these, hierarchical VAEs (HVAEs, Child, 2020; Vahdat & Kautz, 2020), diffusion models (Preechakul et al., 2022; Rombach et al., 2022; Song et al., 2021a), and optimal-transport flow matching (Lipman et al., 2023; Pooladian et al., 2023; Tong et al., 2024) have been used to parameterise mechanisms for generating high-fidelity counterfactuals in mammography, 2D and 3D brain MRIs, subcortical meshes and chest X-rays (De Sousa Ribeiro et al., 2023; Peng et al., 2025; Rasal et al., 2022; 2025; Ribeiro et al., 2025; Xia et al., 2025a). However, most existing counterfactual image generation methods train mechanisms from scratch on the target domain. This limits their applicability in clinical settings, where annotated data are scarce, costly to obtain, and difficult to pool across institutions because of privacy and governance constraints (Kaissis et al., 2020; Willemink et al., 2020). In contrast, our work adapts pretrained generative models into causal mechanisms, reducing the need for large target-domain datasets.

Medical image generation and editing with foundation models. General-purpose image editing methods such as DiffEdit and text-guided editing localise changes through masks or textual instructions (Couairon et al., 2023; Lin et al., 2024), and analogous approaches have been adopted in medical imaging. In radiology, recent models have used text, reports, masks, or other conditioning signals to synthesise or modify chest X-rays with improved visual fidelity and controllability (Bluethgen et al., 2024; Cooke et al., 2025; Gu et al., 2023; Kumar et al., 2025; Pérez-García et al., 2025a; Pinaya et al., 2023; Trang et al., 2025; Weber et al., 2023). These models provide strong image priors and can perform clinically meaningful edits, but their control is usually associative rather than counterfactual: prompts, masks, or bounding boxes specify desired visual changes without explicitly modelling the causal mechanisms that generated the observation. However, mask-based methods require spatial annotations that are often unavailable in clinical datasets (Ma et al., 2024a; Wang et al., 2021; Xing et al., 2023), and local masks may be insufficient for global interventions involving acquisition view, demographic attributes, or anatomical morphology. Prompt-based control also introduces ambiguity, since clinically relevant causal variables may be only indirectly represented in free text.

Our work differs from medical image editing in that we specialise pretrained image models to parameterise explicit SCM mechanisms. Rather than asking a model to edit an image through a prompt or mask, we abduct an exogenous code from the observed image and regenerate the image under intervened parent variables. This connects the flexibility of foundation image models with the formal semantics of counterfactual inference.

Domain shift and causal modelling in medical imaging. Medical imaging datasets often differ across hospitals, scanners, protocols, patient populations, and labelling pipelines. Such acquisition shifts and measurement biases can degrade model performance and lead to unreliable conclusions when models are transferred across datasets (Castro et al., 2020; Roschewitz et al., 2023). Causal approaches have been proposed as a way to reason about these shifts, since SCMs separate stable mechanisms from changing marginal distributions and provide a language for interventions and counterfactuals (Bareinboim et al., 2022; Richens et al., 2020; Schölkopf* et al., 2021). In medical imaging, causal structure has been used to study confounding, improve robustness, and define clinically meaningful image interventions (Mehta et al., 2025; Roschewitz et al., 2024; Xia et al., 2025b). However, most work either assumes that the relevant mechanisms can be learned directly from the target domain, or focuses on associative and interventional robustness rather than full

counterfactual image generation. Our setting is more demanding: the target domain may contain limited labelled data, may differ from the source distribution used for pretraining, and may require interventions on variables that were not explicitly modelled during pretraining.

Parameter-efficient adaptation of generative models. Parameter-efficient fine-tuning has emerged as a practical way to adapt large pretrained models without updating all weights. Low-Rank Adaptation (LoRA) learns low-rank weight updates while keeping the pretrained backbone fixed, reducing trainable parameters and improving fine-tuning efficiency (Hu et al., 2022). Adapter-based methods have also been used to add new forms of conditioning to pretrained image models (Zhang et al., 2023). In the causal setting, Causal-Adapter (Tong et al., 2026) is closest to our measurement-specialisation setting: it adapts a frozen text-to-image diffusion backbone by learning causal mechanisms over semantic attributes and injecting them through adapter and token-level conditioning. In contrast, our specialisation framework targets radiology counterfactual mechanisms under dataset shift: S-SPEC adds explicit parent conditioning to pretrained generators, SA-SPEC additionally adapts the latent representation and generator to acquisition-shifted target domains, and M-SPEC replaces text conditioning with structured clinical parent variables.

B Evaluation implementation details

B.1 Pseudo-oracle classifiers.

All classifiers implement $\mathbf{x} \to a$ and are trained exclusively on the training split of the corresponding evaluation dataset, ensuring no patient-level overlap with the generative evaluation data. Training uses the Adam optimiser with an 80/20 train/validation split, checkpointing the epoch with best validation AUC for discrete attributes (view, sex, effusion, device) and lowest validation MAE for continuous attributes (age). Binary decision thresholds are chosen on the validation set to maximise the sum of sensitivity and specificity, then applied to the test set. Classifier performance is summarised in Table 4.

View and Sex. Both attributes are classified using a ResNet-18 (ImageNet pretrained) with a binary head. Training ran for up to 3 epochs with $\text{lr} = 10^{-4}$ and class-balanced minibatches, identically for NIH14, CheXpert, and BRAX.

Effusion. Based on Aslani et al. (2023), classifiers are trained separately on PA and AP subsets to remove the well-established systematic view-disease confound in chest radiographs (Raoof et al., 2012). All three datasets use label smoothing ($\varepsilon = 0.1$) and ReduceLROnPlateau (factor 0.2, patience 3 epochs). On NIH14, ablation showed that domain-specific initialisation improves Effusion sensitivity: a DenseNet-121 is initialised from CheXpert-pretrained weights via TorchXRayVision (Cohen et al., 2022) and fine-tuned at $\text{lr} = 10^{-5}$ for up to 20 epochs to preserve these representations. On CheXpert, alternative domain-pretrained weights and the DenseNet-121 architecture were evaluated in ablation but offered no improvement, so an ImageNet-pretrained ResNet-18 is trained at $\text{lr} = 10^{-4}$ for up to 15 epochs. NIH14 and CheXpert classifiers are benchmarked against published Effusion classifiers in Table 4, supporting their use as sufficiently performant pseudo-oracles; no published benchmark exists for BRAX.

At evaluation time, counterfactuals are routed to the PA or AP disease classifier by predicted view rather than original or target view. This handles two edge cases: the model may accidentally flip view when intervening on another attribute, or a targeted view flip may fail. In both cases, predicted-view routing ensures the disease classifier matches the image's actual acquisition geometry. For AUC-based classifier calibration, selection uses ground truth view. When view is the target intervention, counterfactual images are routed to the target-view disease classifier, measuring whether $\widetilde{\mathbf{x}}$ convinces that classifier.

Age. A ResNet-18 with a sigmoid scalar output predicts normalised age in $[0, 1]$, trained with Huber loss ($\delta = 0.1$) for robustness to label noise. Training runs for up to 10 epochs at $\text{lr} = 10^{-4}$ with ReduceLROnPlateau (factor = 0.2, patience = 2 epochs), identically for NIH14, CheXpert, and BRAX. Direction success for age is computed only over samples where the requested change satisfies $|\widetilde{a} - a| > 10^{-3}$. Smaller changes are excluded as trivial.

Table 4: Pseudo-oracle performance. Val. and Test report AUC for binary classifiers and MAE for the age regressor (normalised age $\in [0, 1]$).

Dataset	Pseudo-oracle	Architecture	Best ep.	Val.	Test
NIH14	View	ResNet-18	2	0.999	1.000
	Sex	ResNet-18	3	0.999	0.995
	Effusion (PA)[†]	DenseNet-121	9	0.881	0.863
	Effusion (AP)[†]	DenseNet-121	17	0.827	0.812
	Age	ResNet-18	9	0.041	0.051
CheXpert	View	ResNet-18	3	0.997	0.995
	Sex	ResNet-18	2	0.993	0.992
	Effusion (PA)[‡]	ResNet-18	13	0.935	0.911
	Effusion (AP)[‡]	ResNet-18	12	0.919	0.928
	Age	ResNet-18	9	0.067	0.072
BRAX	View	ResNet-18	2	0.999	0.999
	Sex	ResNet-18	2	0.984	0.982
	Effusion (PA)[§]	ResNet-18	6	0.908	0.853
	Effusion (AP)[§]	ResNet-18	11	0.807	0.663
	Age	ResNet-18	8	0.047	0.051
	Device (PA)	ResNet-18	5	0.771	0.743
	Device (AP)	ResNet-18	7	0.965	0.939

[†] Wang et al. (2017) report Effusion AUC = 0.736 (8-class) and Rajpurkar et al. (2017) report 0.864 (14-class), both training jointly across all labels on pooled frontal views without view stratification.

[‡] Irvin et al. (2019) report Pleural Effusion AUC = 0.97 (multi-view, uncertainty-aware training, radiologist-annotated ground truth). Note that direct comparison is limited by multi-view input and the stronger reference standard.

[§] No published Effusion AUC benchmark exists for BRAX; values are reported for reference only.

Device (BRAX only). Two separate ResNet-18 classifiers (ImageNet-pretrained) with binary heads are trained, one per view (PA and AP), to account for view-dependent device appearance. Training follows the same protocol as View and Sex: Adam optimiser at $\text{lr} = 10^{-4}$, class-balanced minibatches, CrossEntropyLoss (no label smoothing), and ReduceLROnPlateau (factor 0.2, patience 3 epochs). Decision thresholds are selected by Youden's J on the validation set. At evaluation time, counterfactuals are routed to the PA or AP device classifier by predicted view, analogously to the Effusion routing described above.

B.2 Evaluator sensitivity

Effectiveness in the main text is measured with the pseudo-oracles of Appendix B.1. While pseudo-oracles are widely adopted as a measure of effectiveness, we acknowledge the potential for evaluator bias to confound conclusions. To investigate its impact on our results, we additionally re-scored every counterfactual with six independently trained public evaluators spanning different backbones, training paradigms and resolutions: four TorchXRayVision models (Cohen et al., 2022), CheXzero (Tiu et al., 2022), and a linear classifier fitted on frozen RAD-DINO features (Pérez-García et al., 2025b). Each evaluator is thresholded at its own Youden-optimal operating point on the NIH-14 train split rather than at 0.5, so that flip rates are not confounded by differences in classifier calibration. The evaluators themselves differ substantially in quality, spanning effusion AUC 0.756 to 0.864, allowing us to compare oracle-based effectiveness across performances. Table 5 reports the resulting thresholds and discriminative performance.

Importantly, the relative ordering of methods is stable across evaluators, with Kendall's $\tau = 0.949 \pm 0.048$ (all $p < 0.003$), as in Table 6. Method rankings show strong agreement across evaluators. RadCF achieves the highest effusion effectiveness among the listed models without classifier-guided counterfactual fine-tuning, while the fine-tuned HVAE ranks first under all six evaluators. It should be noted that HVAE were trained with an additional counterfactual finetuning step, which use

effectiveness as a guiding signal for optimization. Without the additional counterfactual finetuning, HVAE's effectiveness drops significantly. Absolute effectiveness, by contrast, varies widely on identical images: RadCF flips effusion for between 32.0% and 47.2% of cases depending only on which evaluator is asked, a 15-point spread that broadly tracks evaluator AUC. The comparative conclusions drawn from the main tables are therefore robust to the choice of evaluator, whereas any single absolute effectiveness number should be read as evaluator-relative. We note one limitation: six evaluators are available only for effusion, and view, sex and age are re-scored with the RAD-DINO probe alone, so rank stability is established for effusion and assumed rather than verified for the remaining attributes.

Table 5: The quality of effusion evaluators varies. Effusion evaluators: Youden-optimal operating point and discriminative performance on the NIH-14 test split.

Evaluator	Threshold	AUC ↑	Sensitivity	Specificity
TorchXRayVision (all)	0.317	0.852	0.867	0.686
TorchXRayVision (CheXpert)	0.757	0.756	0.682	0.697
TorchXRayVision (NIH)	0.507	0.823	0.793	0.706
TorchXRayVision (ResNet-50)	0.103	0.864	0.803	0.776
CheXzero (Tiu et al., 2022)	0.496	0.864	0.878	0.702
RAD-DINO linear probe	0.101	0.853	0.806	0.761

Table 6: **Method rankings are consistent under all six evaluators.** Values are the effusion flip rate on NIH-14 under each evaluator, thresholded at its Youden-optimal operating point (Table 5). Rankings agree at Kendall's $\tau = 0.949 \pm 0.048$ (all $p < 0.003$), while absolute rates differ by as much as 0.177 on identical images. TXV abbreviates TorchXRayVision. Best per column in bold.

Method	CheXzero	RAD-DINO	TXV (all)	TXV (CheX)	TXV (NIH)	TXV (R50)
HVAE (De Sousa Ribeiro et al., 2023) (40 steps)	0.009	0.012	0.010	0.015	0.019	0.016
Diffusion (Rasal et al., 2025)	0.116	0.102	0.101	0.105	0.123	0.137
S-SPEC	0.198	0.178	0.255	0.227	0.241	0.253
Image-space flow (Ribeiro et al., 2025)	0.266	0.228	0.273	0.226	0.235	0.302
HVAE (De Sousa Ribeiro et al., 2023) (20 steps)	0.438	0.261	0.342	0.284	0.288	0.303
RadCF	0.454	0.472	0.450	0.320	0.416	0.472
HVAE (De Sousa Ribeiro et al., 2023) + counterfactual finetuning	**0.627**	**0.691**	**0.673**	**0.666**	**0.696**	**0.766**

B.3 LPIPS.

We use the `lpips` package released alongside Zhang et al. (2018), which computes the ℓ_2 distance between unit-normalised intermediate activations. We use the linearly calibrated AlexNet backbone.

B.4 KID.

Diverging from Melistas et al. (2025), we use the target-domain training set rather than test-split factuals as the reference distribution for KID. This ensures realism reflects proximity to the data manifold independently of the evaluation population. KID is the squared MMD between Inception-v3 representations under a degree-3 polynomial kernel, computed with subset size 100 and reported as mean (std) across subsets.

B.5 Uncertainty reporting

Binary effectiveness (flip rate). For discrete interventions (view, effusion, sex), effectiveness is the fraction of test images whose target attribute is successfully flipped. We report 95% Wald binomial confidence intervals: $\mathrm{CI} = z\sqrt{p(1-p)/n}$, where p is the observed flip rate, n is the number of test images, and $z = 1.96$. Effusion is reported as a view-balanced mean of the PA and AP flip rates, $p = \frac{1}{2}(p_{\mathrm{PA}} + p_{\mathrm{AP}})$, so that neither view dominates the score. The two subgroups partition the test set, so their estimates are independent and

$$\mathrm{CI} = \tfrac{1}{2}\sqrt{\mathrm{CI}_{\mathrm{PA}}^2 + \mathrm{CI}_{\mathrm{AP}}^2}, \qquad \mathrm{CI}_g = z\sqrt{p_g(1-p_g)/n_g},$$

with each subgroup interval evaluated at its own size n_g. When averaging flip rates across k interventions (e.g. in Figure 2), we propagate per-intervention CIs as $\mathrm{CI}_{\mathrm{mean}} = \sqrt{\sum_{i=1}^{k} \mathrm{CI}_i^2} \,/\, k$.

Continuous per-image metrics (LPIPS, age MAE). Composition LPIPS and reversibility LPIPS are computed per image; we report the median over the test set with one standard deviation of the per-image values in parentheses. Age MAE is computed per image and averaged (mean) over the test set, reported with one standard deviation in parentheses. When averaging across k interventions, we propagate standard deviations as $\sigma_{\mathrm{mean}} = \sqrt{\sum_{i=1}^{k} \sigma_i^2} \,/\, k$, assuming independence across intervention families.

KID. Kernel Inception Distance is estimated from random sub-samples of real and generated feature vectors. We report one standard deviation over sub-samples in parentheses, as returned by the `torchmetrics` KID implementation. Cross-intervention KID in Figure 2 propagates per-intervention standard deviations using the same formula as above.

C Information-theoretic motivation for specialisation

We consider the setting where a generative model is trained on a source domain $p_s(\mathbf{x})$, or on a source text-conditional distribution $p_s(\mathbf{x} \mid \mathrm{TEXT})$, and then specialised on a target domain $p_t(\mathbf{x}, \mathbf{pa})$. In our experiments, the source domain corresponds to CheXpert or a pretrained radiology foundation model, while the target domain corresponds to NIH-14, BRAX, or a structured-metadata counterfactual task.

Throughout this appendix we analyse an idealised adaptation gap. We assume that, in the target domain, the specified image mechanism is causally sufficient for $\mathbf{x}$, i.e.

$$\mathbf{x} = f_t(\mathbf{u}, \mathbf{pa}), \qquad \mathbf{u} \perp\!\!\!\perp \mathbf{pa}.$$

Under this assumption,

$$p_t(\mathbf{x} \mid \mathrm{do}(\mathbf{pa})) = p_t(\mathbf{x} \mid \mathbf{pa}).$$

The following results should be interpreted as distributional motivation for specialisation, not as finite-sample generalisation guarantees.

Proposition 1: Source pretraining decomposes target adaptation into structure and acquisition mismatch. Assume the source-pretrained unconditional generator represents the source image marginal $p_s(\mathbf{x})$. Define the target adaptation gap before introducing explicit parent conditioning as

$$\mathcal{G}_{\mathrm{S}}(p_s) := \mathbb{E}_{p_t(\mathbf{pa})} \left[\mathrm{KL}\left(p_t(\mathbf{x} \mid \mathrm{do}(\mathbf{pa})) \,\|\, p_s(\mathbf{x}) \right) \right].$$

Then

$$\mathcal{G}_{\mathrm{S}}(p_s) = I_t(\mathbf{x}; \mathbf{pa}) + \mathrm{KL}\left(p_t(\mathbf{x}) \,\|\, p_s(\mathbf{x}) \right),$$

where $I_t(\mathbf{x}; \mathbf{pa})$ denotes mutual information under the target distribution $p_t(\mathbf{x}, \mathbf{pa})$.

Proof. Using $p_t(\mathbf{x} \mid \mathrm{do}(\mathbf{pa})) = p_t(\mathbf{x} \mid \mathbf{pa})$, we have

$$\begin{aligned}
\mathcal{G}_{\mathrm{S}}(p_s) &= \mathbb{E}_{p_t(\mathbf{pa})} \mathbb{E}_{p_t(\mathbf{x}|\mathbf{pa})} \left[\log p_t(\mathbf{x} \mid \mathbf{pa}) - \log p_s(\mathbf{x}) \right] \\
&= \mathbb{E}_{p_t(\mathbf{pa})} \mathbb{E}_{p_t(\mathbf{x}|\mathbf{pa})} \left[\log p_t(\mathbf{x} \mid \mathbf{pa}) - \log p_t(\mathbf{x}) \right] \\
&\quad + \mathbb{E}_{p_t(\mathbf{pa})} \mathbb{E}_{p_t(\mathbf{x}|\mathbf{pa})} \left[\log p_t(\mathbf{x}) - \log p_s(\mathbf{x}) \right] \\
&= I_t(\mathbf{x}; \mathbf{pa}) + \mathrm{KL}\left(p_t(\mathbf{x}) \,\|\, p_s(\mathbf{x}) \right).
\end{aligned}$$

The last equality follows from the definition of mutual information and the tower property $\mathbb{E}_{p_t(\mathbf{pa})} \mathbb{E}_{p_t(\mathbf{x}|\mathbf{pa})}[\cdot] = \mathbb{E}_{p_t(\mathbf{x})}[\cdot]$.

Interpretation. Proposition 1 separates the target adaptation gap into two terms. The first term, $I_t(\mathbf{x}; \mathbf{pa})$, is the parent-dependent structure needed for counterfactual control. This is the term targeted by S-SPEC, which introduces explicit conditioning on $\mathbf{pa}$ through low-rank updates to the pretrained mechanism.

The second term,

$$\mathrm{KL}\left(p_t(\mathbf{x}) \,\|\, p_s(\mathbf{x}) \right),$$

is the source-target marginal mismatch. In our setting, this corresponds to acquisition shift, scanner differences, image preprocessing differences, and population differences between source and target domains. This term motivates SA-SPEC: adapting the autoencoder helps align the image marginal of the source-pretrained model with the target domain, while the flow learns the parent-dependent structure. Thus, when $p_s(\mathbf{x})$ is already close to $p_t(\mathbf{x})$, S-SPEC may be sufficient; when the marginal mismatch is large, SA-SPEC is expected to help.

Proposition 2: Text-conditional pretraining reduces the residual structural gap. We now consider a source model pretrained on $p_s(\mathbf{x} \mid \mathrm{TEXT})$. Since the target domain is specified by structured metadata $p_t(\mathbf{x}, \mathbf{pa})$, we introduce an idealised measurement model $r(\mathrm{TEXT} \mid \mathbf{pa})$, which represents the text proxy associated with $\mathbf{pa}$. This induces the joint distribution

$$p_t^r(\mathbf{x}, \mathbf{pa}, \mathrm{TEXT}) := p_t(\mathbf{x}, \mathbf{pa})\, r(\mathrm{TEXT} \mid \mathbf{pa}).$$

By construction, this idealised proxy satisfies

$$\mathbf{x} \perp\!\!\!\perp \mathrm{TEXT} \mid \mathbf{pa}.$$

Define the target adaptation gap for a source text-conditional model as

$$\mathcal{G}_{\mathrm{M}}(p_s) := \mathbb{E}_{p_t^r(\mathbf{pa},\mathrm{TEXT})}\left[\mathrm{KL}\left(p_t(\mathbf{x} \mid \mathrm{do}(\mathbf{pa})) \,\|\, p_s(\mathbf{x} \mid \mathrm{TEXT})\right)\right].$$

Then

$$\mathcal{G}_{\mathrm{M}}(p_s) = I_{t,r}(\mathbf{x}; \mathbf{pa} \mid \mathrm{TEXT}) + \mathbb{E}_{p_t^r(\mathrm{TEXT})}\,\mathrm{KL}\left(p_t^r(\mathbf{x} \mid \mathrm{TEXT}) \,\|\, p_s(\mathbf{x} \mid \mathrm{TEXT})\right),$$

where $I_{t,r}$ denotes mutual information under $p_t^r(\mathbf{x}, \mathbf{pa}, \mathrm{TEXT})$.

Proof. Using $p_t(\mathbf{x} \mid \mathrm{do}(\mathbf{pa})) = p_t(\mathbf{x} \mid \mathbf{pa})$, we expand

$$\begin{aligned}
\mathcal{G}_{\mathrm{M}}(p_s) &= \mathbb{E}_{p_t^r(\mathbf{pa},\mathrm{TEXT})}\mathbb{E}_{p_t(\mathbf{x}|\mathbf{pa})}\left[\log p_t(\mathbf{x} \mid \mathbf{pa}) - \log p_s(\mathbf{x} \mid \mathrm{TEXT})\right] \\
&= \mathbb{E}_{p_t^r(\mathbf{pa},\mathrm{TEXT})}\mathbb{E}_{p_t(\mathbf{x}|\mathbf{pa})}\left[\log p_t(\mathbf{x} \mid \mathbf{pa}) - \log p_t^r(\mathbf{x} \mid \mathrm{TEXT})\right] \\
&\quad + \mathbb{E}_{p_t^r(\mathbf{pa},\mathrm{TEXT})}\mathbb{E}_{p_t(\mathbf{x}|\mathbf{pa})}\left[\log p_t^r(\mathbf{x} \mid \mathrm{TEXT}) - \log p_s(\mathbf{x} \mid \mathrm{TEXT})\right].
\end{aligned}$$

Because $p_t^r(\mathbf{x}, \mathbf{pa}, \mathrm{TEXT}) = p_t(\mathbf{x}, \mathbf{pa}) r(\mathrm{TEXT} \mid \mathbf{pa})$, we have $\mathbf{x} \perp\!\!\!\perp \mathrm{TEXT} \mid \mathbf{pa}$. Therefore, the first term is

$$H_{t,r}(\mathbf{x} \mid \mathrm{TEXT}) - H_t(\mathbf{x} \mid \mathbf{pa}) = I_{t,r}(\mathbf{x}; \mathbf{pa} \mid \mathrm{TEXT}).$$

For the second term, marginalising over $\mathbf{pa}$ gives

$$\mathbb{E}_{p_t^r(\mathrm{TEXT})}\,\mathrm{KL}\left(p_t^r(\mathbf{x} \mid \mathrm{TEXT}) \,\|\, p_s(\mathbf{x} \mid \mathrm{TEXT})\right).$$

Combining these terms proves the result.

Comparison with unconditional source pretraining. Let

$$\delta_{\mathrm{S}} = \mathrm{KL}\left(p_t(\mathbf{x}) \,\|\, p_s(\mathbf{x})\right)$$

denote the unconditional source-target marginal mismatch, and let

$$\delta_{\mathrm{M}} = \mathbb{E}_{p_t^r(\mathrm{TEXT})}\,\mathrm{KL}\left(p_t^r(\mathbf{x} \mid \mathrm{TEXT}) \,\|\, p_s(\mathbf{x} \mid \mathrm{TEXT})\right)$$

denote the text-conditional source-target mismatch. Combining Propositions 1 and 2 gives

$$\mathcal{G}_{\mathrm{S}}(p_s) - \mathcal{G}_{\mathrm{M}}(p_s) = I_{t,r}(\mathbf{x}; \mathrm{TEXT}) + \delta_{\mathrm{S}} - \delta_{\mathrm{M}}.$$

Thus, text-conditional source pretraining reduces the idealised adaptation gap whenever the information about the target image contained in the text proxy, $I_{t,r}(\mathbf{x}; \mathrm{TEXT})$, together with the unconditional source-target alignment, outweighs the conditional source-target mismatch.

Interpretation. Proposition 2 motivates M-SPEC. Text-conditional radiology models have already learned image variations associated with textual descriptions. If TEXT is informative about $\mathbf{pa}$, the remaining structural gap is reduced from $I_t(\mathbf{x}; \mathbf{pa})$ to $I_{t,r}(\mathbf{x}; \mathbf{pa} \mid \mathrm{TEXT})$, up to conditional source-target mismatch. M-SPEC reuses this text-conditional image prior while replacing imprecise text prompts with structured metadata embeddings $g_\omega(\mathbf{pa})$. This gives explicit intervention control over $\mathbf{pa}$ while preserving semantic knowledge from the pretrained text-conditional backbone.

The assumption $\mathbf{x} \perp\!\!\!\perp \mathrm{TEXT} \mid \mathbf{pa}$ is idealised. Radiology text can contain information beyond the selected variables in $\mathbf{pa}$, and structured metadata can itself be noisy. Therefore, Proposition 2 should be interpreted as an adaptation-gap motivation for measurement specialisation, rather than a guarantee that text conditioning is sufficient for counterfactual control.

Table 7: Dataset summary statistics.

	CheXpert	CheXpert Frontal	NIH-14	BRAX
Total images	223,414	33,454	112,104	14,636
Frontal / Lateral	191,027 / 32,387	33,454 / —	112,104 / —	14,636 / —
Patients	64,540	13,020	30,802	11,530
Finetune / Eval split	223,414 / —	16,499 / 16,955	100,000 / 12,104	11,636 / 3,000
AP / PA (%)	84.6 / 15.4	82.6 / 17.4	40.0 / 60.0	30.3 / 69.7
Male (%)	59.4	59.2	56.5	61.6
Age (mean $\pm$ std)	60.4 ± 17.8	61.6 ± 17.5	46.9 ± 16.6	45.7 ± 26.5
Effusion prevalence (%)	43.8	71.4	11.9	5.4
Device prevalence (%)	—	—	—	37.6

D Experimental details

D.1 Datasets and preprocessing

We use three publicly available chest radiograph datasets summarised in Table 7. All three are released under data use agreements for research purposes; no additional institutional review board approval was required.

CheXpert. The base generator is trained on 223,414 CheXpert radiographs (191,027 frontal and 32,387 lateral images) from 64,540 patients at Stanford Hospital (Irvin et al., 2019). Patient age ranges from 0 to 90 years (mean 60.4 ± 17.8); 59.4% are male; effusion prevalence is 43.8%. For the CheXpert→CheXpert adaptation experiment, we filter to 33,454 frontal images from 13,020 patients (16,499 training, 16,955 evaluation). Although these images were seen during base-model pretraining, the base model was not conditioned on the intervention variables; this setting tests whether S-SPEC alone suffices when no acquisition shift is present.

NIH-14. NIH ChestX-ray14 contains 112,120 frontal chest radiographs from 30,805 patients (Wang et al., 2017). We discard 16 images with implausible ages (>100 years), yielding 112,104 images from 30,802 patients. Pathology labels for 14 conditions, including effusion, were extracted by the original dataset authors using NLP applied to associated radiology reports; these are *not* derived from the CheXpert labeller. We use 100,000 images for specialisation training and hold out the remaining 12,104 images for evaluation, with no patient-level overlap between the two sets. NIH-14 differs from CheXpert in institution, scanner hardware, image resolution, and acquisition protocol, introducing the acquisition shift that motivates the VAE component of SA-SPEC. Patient age ranges from 1 to 100 years (mean 46.9 ± 16.6); 56.5% are male; 60.0% are PA views; effusion prevalence is 11.9%.

BRAX. BRAX comprises 40,967 radiographs (AP, PA, and lateral) from 19,351 patients at a Brazilian institution (Reis et al., 2022). We retain only frontal views for consistency with the other datasets, yielding 19,309 images from 15,113 patients. We observed substantial label noise in the BRAX device annotations, so we used a device classifier on Chexpert, and removed images whose device-classifier prediction confidence falls below 0.85, leaving 14,636 frontal images from 11,530 patients (11,636 training, 3,000 evaluation). BRAX introduces a more extreme domain shift than NIH-14: patient demographics, disease prevalence, and imaging equipment differ substantially from CheXpert. Patient age ranges from 0 to 85+ years (mean 45.7 ± 26.5); 61.6% are male; 69.7% are PA views; effusion prevalence is 5.4%; device prevalence is 37.6%.

Intervention variables. Across all datasets, we extract four parent variables from structured metadata: view position $v \in \{\text{AP}, \text{PA}\}$, sex $s \in \{\text{M}, \text{F}\}$, age $a \in [0, 1]$ (divided by 100), and effusion status $e \in \{0, 1\}$. BRAX additionally provides a support device label $d \in \{0, 1\}$. Binary labels are used as-is from the dataset metadata.

Preprocessing. All images are resized to 256×256 pixels using direct resize with bilinear interpolation. Images are converted to 3-channel RGB and pixel intensities are normalised to $[-1, 1]$. Age is normalised by dividing by 100, mapping to approximately $[0, 1]$. No histogram equalisation, windowing, or data augmentation is applied during specialisation. For evaluation, counterfactual generation is performed on held-out patients with no overlap with the generative training set.

D.2 Model architecture and training

The base transport network is `SiT-XL/2` (Ma et al., 2024b). Source-domain pretraining is performed end-to-end with the autoencoder and REPA-E alignment (Leng et al., 2025). The base model is trained for 400k steps on 4 GPUs with batch size 16. Experiments were run on an internal HPC cluster using NVIDIA H100 (80GB) GPUs, scheduled via SLURM.

We represent the parent variables with a schema-driven embedding: categorical variables are mapped through learned embeddings, continuous variables are projected linearly. The resulting embeddings are summed into a single conditioning vector and injected through the transformer's adaLN-Zero conditioning pathway.

During S-SPEC, LoRA is applied to attention, MLP, and conditioning projections, with rank $r \in \{4, 8, 16, 32\}$ and corresponding scaling factors $\alpha \in \{8, 16, 32, 64\}$, where $r = 32$ is chosen for the main results. During SA-SPEC, the autoencoder and transport are updated in alternating phases exactly as in Leng et al. (2025). Counterfactual inference uses a 200-step Euler ODE solver for both inversion and generation; for the diffusion baseline, 1000 ODE steps are used; for HVAE, we use 20 latent layers. During domain specialisation, LoRA modules are trained with AdamW (β_1=0.9, β_2=0.999, lr $= 10^{-4}$) and batch size 16. A class dropout probability of 0.1 is used during training; classifier-free guidance scale w=1 is applied at inference.

D.3 Baseline implementations

HVAE (De Sousa Ribeiro et al., 2023) is a hierarchical VAE operating at 256×256 with 8 resolution stages (35 encoder blocks, 42 decoder blocks of which 38 are stochastic at 4×4–128×128, z_{dim}=16; 8.1M parameters). Parent attributes are spatially broadcast and concatenated at every decoder block, giving a conditional prior $p(z_i \mid z_{<i}, \mathbf{pa})$. The model is trained from scratch on NIH-14 with AdamW (lr=10^{-3}, wd=0.05), optimising the negative ELBO with a discretised Gaussian likelihood. The model is then fine-tuned with a Lagrangian objective combining a judge loss from five frozen ResNet-50 classifiers, an ℓ_1 reconstruction penalty, and an ELBO constraint (AdamW, lr=10^{-4}, batch size 8).

Diffusion (Rasal et al., 2025) adapts the spatial counterfactual mechanism of (Rasal et al., 2025) to a latent diffusion architecture. We use a latent diffusion model based on the U-Net architecture of Rombach et al. (2022), operating at 256×256 resolution with the same KL-regularised autoencoder as in RadCF. Parent conditioning is injected via cross-attention. The model uses the same training data splits as RadCF for fair comparison. Counterfactual inference uses DDIM inversion (Song et al., 2021a) with 1000 steps for both the forward (abduction) and reverse (generation) passes.

E Ablation studies

E.1 LoRA rank

Figure 6 shows the effect of LoRA rank on the trade-off between edit magnitude, reversibility, and composition across all four interventions (frozen VAE, 100% NIH14 training data).

E.2 Latent-space decomposition of the SA-SPEC gain

Table 2 shows that adapting the autoencoder improves composition and reversibility under acquisition shift, but image-space LPIPS conflates three error sources: autoencoder reconstruction, vector-field error, and Euler integration error. We separate them by measuring the autoencoder round trip $\text{LPIPS}(\mathbf{x}, D_{\phi_2}(E_{\phi_1}(\mathbf{x})))$ and the latent ODE round-trip drift at two solver resolutions (Table 8).

In domain the two variants are indistinguishable on every quantity, as expected when there is no acquisition mismatch to correct. Under shift the autoencoder reconstruction error halves (0.0968 to 0.0474, $p < 0.0001$), and this gap does not close with finer integration. The apparent latent-drift difference behaves differently: it is present at 200 solver steps ($p = 0.005$), but at 1000 steps both models fall by roughly an order of magnitude and become statistically indistinguishable ($p = 0.49$). The drift gap is therefore integration error rather than a property of the learned field, whereas the reconstruction gap persists. This supports attributing the SA-SPEC improvement to target-domain

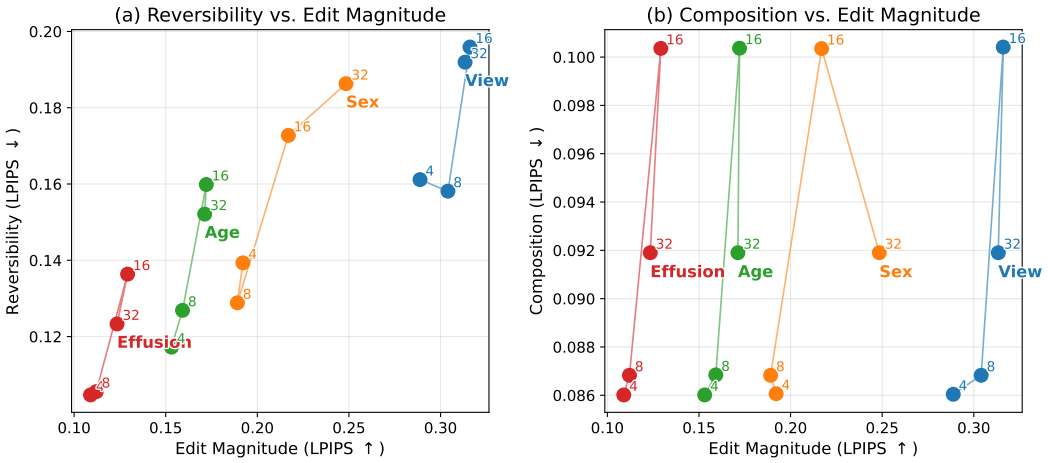


Figure 6: LoRA rank ablation ($r \in \{4, 8, 16, 32\}$, frozen VAE, 100% NIH14). Each panel shows counterfactual soundness metrics as a function of edit magnitude (CF–Orig LPIPS, higher = larger edit). **(a)** Reversibility vs. edit magnitude. **(b)** Composition vs. edit magnitude. Points of the same colour correspond to the same intervention; rank labels indicate r=4, 8, 16, 32 along each curve. Higher rank enables larger edits with better reversibility and composition; r=32 is used in all main experiments.

representation error rather than to transport capacity, consistent with the finding in Table 2 that fully fine-tuning the flow alone does not recover the gain.

Table 8: Latent-space decomposition of the round-trip error highlights SA-SPEC's impact on reducing target-domain representation error over S-SPEC. Brackets give 95% bootstrap confidence intervals; p-values are paired tests over patients. z-drift is the latent round-trip error at 200 and 1000 Euler steps.

Metric	S-SPEC	SA-SPEC	p
In-domain (CheXpert → CheXpert)			
VAE LPIPS ↓	0.0734 [.0730, .0740]	0.0732 [.0730, .0740]	0.681
z-drift @200 ↓	0.00065 [.00047, .00099]	0.00060 [.00046, .00082]	0.776
z-drift @1000 ↓	0.000025 [.000018, .000037]	0.000020 [.000017, .000025]	0.700
Acquisition shift (CheXpert → NIH-14)			
VAE LPIPS ↓	0.0968 [.0920, .1020]	**0.0474** [.0470, .0480]	<0.0001
z-drift @200 ↓	0.00622 [.0040, .0090]	**0.00291** [.0020, .0040]	0.005
z-drift @1000 ↓	0.00061 [.00005, .0017]	0.00031 [.00015, .00053]	0.490

E.3 Adapter spectrum and rank sufficiency

Appendix E.1 selects $r = 32$ by downstream performance. To check that this budget is not exhausted, we take the singular value decomposition of each trained LoRA update $\Delta\theta' = BA$ and summarise the spectrum per layer. The learned updates concentrate 95% of their energy in 18.6 directions on average, indicating a low effective rank within the chosen rank-32 parameterisation.

F Data-efficiency tables

G Extending pa on BRAX

H Cross-backbone comparison with Causal-Adapter

Table 3 in the main text compares M-SPEC against Causal-Adapter (Tong et al., 2026) under a matched backbone: every row is built on RadEdit, so the rows differ only in how causal structure is injected.

Table 9: **Comprehensive data efficiency results across all interventions.** Effectiveness is the classifier flip rate for view, effusion, and sex (↑) and judge MAE for age (↓). Reversibility (Rev.) is round-trip LPIPS (↓). Composition (Comp.) is null-intervention LPIPS (↓). Realism is KID $\times 10^2$ (↓). Binary rates report 95% binomial CI; LPIPS and age MAE report std in parentheses.

		Effectiveness				Reversibility (LPIPS ↓)				Comp.	Realism
Method	Data	do(v)↑	do(e)↑	do(s)↑	do(a)↓	do(v)	do(e)	do(s)	do(a)	LPIPS↓	KID↓
RadCF SA-SPEC	0.1%	0.561 (0.031)	0.350 (0.030)	0.340 (0.029)	0.223 (0.137)	0.090 (0.030)	0.066 (0.020)	0.091 (0.030)	0.053 (0.019)	0.025 (0.008)	4.46 (0.28)
	0.2%	0.561 (0.031)	0.315 (0.030)	0.230 (0.026)	0.231 (0.136)	0.069 (0.021)	0.060 (0.017)	0.068 (0.025)	0.048 (0.015)	0.026 (0.008)	3.97 (0.28)
	0.5%	0.812 (0.024)	0.548 (0.032)	0.440 (0.031)	0.229 (0.137)	0.077 (0.039)	0.057 (0.031)	0.070 (0.038)	0.042 (0.018)	0.023 (0.007)	4.52 (0.27)
	1%	0.922 (0.017)	0.558 (0.032)	0.676 (0.029)	0.197 (0.124)	0.087 (0.057)	0.066 (0.052)	0.077 (0.054)	0.044 (0.034)	0.021 (0.007)	3.38 (0.27)
	2%	0.932 (0.016)	0.657 (0.030)	0.795 (0.025)	0.188 (0.129)	0.091 (0.069)	0.047 (0.056)	0.064 (0.066)	0.040 (0.053)	0.016 (0.020)	3.82 (0.28)
	5%	0.993 (0.005)	0.574 (0.032)	0.913 (0.017)	0.132 (0.090)	0.077 (0.072)	0.040 (0.049)	0.053 (0.062)	0.038 (0.055)	0.016 (0.019)	1.39 (0.20)
	10%	0.992 (0.006)	0.429 (0.032)	0.805 (0.025)	0.152 (0.096)	0.068 (0.062)	0.026 (0.030)	0.034 (0.046)	0.026 (0.051)	0.015 (0.009)	0.73 (0.14)
	25%	0.986 (0.007)	0.405 (0.031)	0.793 (0.025)	0.155 (0.098)	0.059 (0.060)	0.024 (0.025)	0.032 (0.049)	0.026 (0.046)	0.015 (0.007)	0.69 (0.13)
	50%	0.957 (0.013)	0.435 (0.032)	0.723 (0.028)	0.175 (0.120)	0.039 (0.035)	0.021 (0.008)	0.024 (0.017)	0.023 (0.021)	0.014 (0.004)	0.72 (0.13)
	100%	0.995 (0.004)	0.561 (0.032)	0.867 (0.021)	0.133 (0.085)	0.051 (0.058)	0.023 (0.029)	0.030 (0.044)	0.024 (0.042)	0.014 (0.010)	0.64 (0.12)
RadCF	0.1%	0.621 (0.030)	0.228 (0.027)	0.327 (0.029)	0.241 (0.140)	0.103 (0.023)	0.055 (0.010)	0.112 (0.028)	0.053 (0.009)	0.034 (0.006)	13.30 (0.37)
	0.2%	0.637 (0.030)	0.308 (0.030)	0.437 (0.031)	0.241 (0.139)	0.141 (0.065)	0.082 (0.041)	0.134 (0.071)	0.053 (0.010)	0.034 (0.007)	13.43 (0.40)
	0.5%	0.741 (0.027)	0.497 (0.032)	0.564 (0.031)	0.239 (0.136)	0.102 (0.043)	0.070 (0.031)	0.101 (0.057)	0.042 (0.009)	0.027 (0.006)	11.57 (0.40)
	1%	0.848 (0.022)	0.523 (0.032)	0.624 (0.030)	0.239 (0.137)	0.085 (0.047)	0.056 (0.033)	0.076 (0.051)	0.039 (0.018)	0.024 (0.011)	5.51 (0.32)
	2%	0.810 (0.024)	0.521 (0.032)	0.661 (0.029)	0.208 (0.124)	0.050 (0.036)	0.039 (0.030)	0.047 (0.039)	0.029 (0.025)	0.018 (0.009)	3.03 (0.24)
	5%	0.684 (0.029)	0.523 (0.032)	0.562 (0.031)	0.173 (0.113)	0.036 (0.030)	0.032 (0.025)	0.033 (0.024)	0.028 (0.023)	0.017 (0.011)	2.17 (0.21)
	10%	0.964 (0.012)	0.735 (0.028)	0.868 (0.021)	0.120 (0.088)	0.112 (0.089)	0.057 (0.078)	0.077 (0.081)	0.056 (0.088)	0.016 (0.021)	2.34 (0.26)
	25%	0.993 (0.005)	0.592 (0.031)	0.940 (0.015)	0.095 (0.068)	0.099 (0.073)	0.039 (0.049)	0.067 (0.063)	0.053 (0.072)	0.017 (0.022)	0.51 (0.13)
	50%	0.992 (0.006)	0.573 (0.032)	0.936 (0.015)	0.096 (0.069)	0.109 (0.073)	0.034 (0.048)	0.056 (0.062)	0.042 (0.077)	0.016 (0.025)	0.48 (0.11)
	100%	0.998 (0.003)	0.633 (0.031)	0.952 (0.013)	0.087 (0.066)	0.104 (0.081)	0.034 (0.051)	0.061 (0.068)	0.039 (0.083)	0.015 (0.025)	0.40 (0.10)

Table 10: SA-SPEC **maintains identity preservation when extending** pa **in a new target domain (BRAX (Reis et al., 2022)).** Each row reports one intervention. Effectiveness is measured by pseudo-oracles. For the intervened attribute, the corresponding pseudo-oracle metric measures target success. For non-intervened attributes, the metric measures preservation of the original. KID is scaled by $\times 10^2$.

	Effectiveness					Composition	Reversibility	Realism
Interv.	Acc(v) ↑	Acc(e) ↑	Acc(s) ↑	MAE(a) ↓	Acc(d) ↑	LPIPS↓	LPIPS↓	KID↓
do(v)	0.870 (0.012)	0.613 (0.018)	0.785 (0.015)	0.094 (0.078)	0.668 (0.017)	0.018 (0.021)	0.019 (0.034)	2.67 (0.73)
do(e)	0.988 (0.004)	0.509 (0.013)	0.913 (0.010)	0.061 (0.054)	0.786 (0.015)	0.018 (0.021)	0.020 (0.044)	1.14 (0.35)
do(s)	0.989 (0.004)	0.787 (0.015)	0.867 (0.012)	0.062 (0.056)	0.732 (0.016)	0.018 (0.021)	0.019 (0.040)	0.49 (0.24)
do(a)	0.974 (0.006)	0.761 (0.015)	0.907 (0.011)	0.100 (0.081)	0.776 (0.015)	0.018 (0.021)	0.019 (0.036)	0.75 (0.23)
do(d)	0.985 (0.004)	0.774 (0.015)	0.899 (0.011)	0.074 (0.070)	0.729 (0.016)	0.018 (0.021)	0.019 (0.042)	0.78 (0.25)

Table 14 additionally reports the authors' original implementation, which uses Stable Diffusion 1.5 as its backbone.

The SD1.5 rows are not backbone-matched, so effectiveness differences confound the pretrained backbone with the adaptation method. SD1.5 Causal-Adapter attains higher effectiveness on do(v), do(s) and do(a), which we attribute mainly to its natural-image pretraining: the attributes it recovers best are global image properties rather than localised pathology. Its composition (0.113) and reversibility (0.197 mean across interventions) are nonetheless worse than M-SPEC (0.061 / 0.167), and its realism is markedly worse (KID 3.13 vs. 1.35). Under the matched RadEdit backbone, M-SPEC improves on Causal-Adapter for every target attribute and for composition, reversibility and realism; the two exceptions are non-target preservation of sex under do(v) (0.947 vs. 0.963) and under do(a) (0.962 vs. 0.976).

I DOWNSTREAM UTILITY EXPERIMENTS

Biased dataset construction. We construct biased training and test sets from NIH-14 by subsampling from the patient pool. Each bias level sets the probability that an effusion-positive case is assigned the PA view, so the effusion rate per view reflects how strongly view predicts disease (Table 11). The audit experiment spans ten levels of spurious view-effusion correlation from 0.50 to 0.95; the debiasing experiment uses three representative levels (0.50, 0.75, 0.95).

Shortcut learning audit experiment. At each bias level, we train a DenseNet-121 classifier to predict effusion with five-fold patient-level cross validation, reporting all results as the mean and standard deviation across folds. To probe each classifier, we generate view counterfactuals do(v) for every test image with SA-SPEC RadCF (CheXpert→NIH-14). As a random control, we additionally re-score

Table 11: **Training dataset bias construction.** Each level is subsampled from the same patient pool to create a different level of spurious view-effusion correlation. $P(e{=}1)$ is the effusion rate per view, and a larger gap between views means view predicts effusion more strongly.

Bias	v	$e=0$	$e=1$	$P(e{=}1)$
0.50 ($N=4{,}945$)	AP	2,671	294	9.9%
	PA	1,626	354	17.9%
0.55 ($N=5{,}061$)	AP	2,949	272	8.4%
	PA	1,450	390	21.2%
0.60 ($N=5{,}168$)	AP	3,213	239	6.9%
	PA	1,290	426	24.8%
0.65 ($N=5{,}265$)	AP	3,468	215	5.8%
	PA	1,126	456	28.8%
0.70 ($N=5{,}401$)	AP	3,739	181	4.6%
	PA	986	495	33.4%
0.75 ($N=5{,}501$)	AP	4,004	156	3.8%
	PA	808	533	39.7%
0.80 ($N=5{,}580$)	AP	4,250	121	2.8%
	PA	644	565	46.7%
0.85 ($N=5{,}704$)	AP	4,541	88	1.9%
	PA	475	600	55.8%
0.90 ($N=5{,}812$)	AP	4,796	57	1.2%
	PA	326	633	66.0%
0.95 ($N=5{,}904$)	AP	5,053	31	0.6%
	PA	156	664	81.0%

each original under label-preserving, non-generative augmentations (horizontal flip, small affine, colour jitter), which perturb the image without altering the view. We then compare the classifier's predictions across the original, counterfactual, and control images.

Counterfactual augmentation experiment. To test whether counterfactual augmentation can reduce shortcut learning during training, we compare a baseline classifier against three augmentation strategies: random upsampling, conventional random augmentation, and counterfactual augmentation. Each strategy is size-matched to twice the biased set for a fair comparison. Augmentation and auditing deliberately use different RadCF generation mechanisms: the view-flipped counterfactuals added to training come from S-SPEC trained on NIH-14, while the auditing is carried out by the independent SA-SPEC auditor. This avoids circularity: were the same generator used for both, a classifier could appear robust simply by learning to agree with that generator's edits. Auditing with an independent generator ensures that reduced reliance reflects robustness to the view shortcut rather than consistency with the augmentation generator.

J MULTI-PARENT INTERVENTIONS

J.1 SCM-GUIDED PROPAGATION

To illustrate composition with an upstream SCM, we manually specify a metadata SCM over $\mathbf{m} = (s, a, e, v)$:

$$a \to e \to v, \qquad \{s, a, e, v\} \to \mathbf{x}, \tag{11}$$

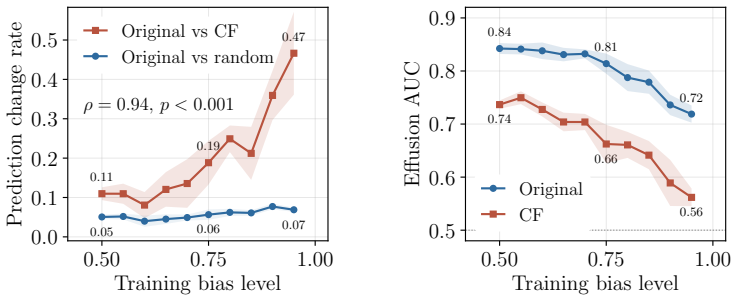
where s and a are observed root variables, e denotes effusion status, and v denotes acquisition view. The leaf mechanism $\{s, a, e, v\} \to \mathbf{x}$ is the specialised image mechanism and is kept fixed throughout this experiment. Here the graph decides which parents change; Appendix J.2 instead intervenes on a fixed number of parents chosen at random, isolating the capacity of the mechanism from the structure of the graph.

We fit Bernoulli mechanisms for the two non-root metadata variables (Oberst & Sontag, 2019):

$$e = \mathbb{1}[u_e < \hat{p}_e(a)], \qquad v = \mathbb{1}[u_v < \hat{p}_v(e)], \qquad u_e, u_v \sim \mathrm{Unif}(0, 1). \tag{12}$$

Table 12: **Augmenting training with counterfactuals reduces view-shortcut reliance without sacrificing discrimination.** Effusion classifiers are trained under three bias levels with each mitigation strategy (size-matched to 2N samples) and evaluated on original and view-counterfactual images. Prediction change rate (decision threshold=0.5) measures reliance under $\mathrm{do}(v)$, and $\mathrm{AUC}_{\mathrm{orig}}/\mathrm{AUC}_{\mathrm{cf}}$ measure discrimination on real and counterfactual images. View gap is the stratified subgroup difference $|\mathrm{AUC}_{\mathrm{AP}} - \mathrm{AUC}_{\mathrm{PA}}|$ on originals. The mean and standard deviation of five-fold cross validation are reported. Within each bias level, the best value per column is in bold.

Bias	Mitigation	$\mathrm{AUC}_{\mathrm{orig}} \uparrow$	$\mathrm{AUC}_{\mathrm{cf}} \uparrow$	Change rate $\downarrow$	View gap $\downarrow$
0.50	Baseline	0.841 (0.018)	0.748 (0.021)	**0.070** (0.022)	0.081 (0.012)
	Random upsample	0.831 (0.025)	0.721 (0.021)	0.214 (0.055)	0.078 (0.012)
	Random aug.	0.848 (0.014)	0.748 (0.022)	0.088 (0.017)	**0.076** (0.013)
	Counterf. aug.	**0.852** (0.009)	**0.805** (0.007)	0.085 (0.024)	0.081 (0.006)
0.75	Baseline	0.816 (0.014)	0.663 (0.021)	0.253 (0.042)	**0.062** (0.023)
	Random upsample	**0.821** (0.013)	0.703 (0.024)	0.307 (0.079)	0.078 (0.029)
	Random aug.	0.820 (0.015)	0.645 (0.041)	0.263 (0.079)	0.077 (0.022)
	Counterf. aug.	0.819 (0.017)	**0.760** (0.011)	**0.101** (0.014)	0.078 (0.034)
0.95	Baseline	0.703 (0.036)	0.531 (0.039)	0.593 (0.197)	0.084 (0.054)
	Random upsample	**0.734** (0.024)	0.581 (0.024)	0.468 (0.092)	**0.030** (0.012)
	Random aug.	0.714 (0.021)	0.528 (0.034)	0.655 (0.062)	0.036 (0.027)
	Counterf. aug.	0.712 (0.042)	**0.683** (0.027)	**0.115** (0.020)	0.071 (0.032)



(a) Prediction change rate.

(b) Prediction AUC.

Figure 7: **Paired original-counterfactual probing exposes shortcut learning that accuracy metrics miss.** An effusion classifier is trained at ten view-effusion bias levels and probed with counterfactuals $\mathrm{do}(v)$ and a random, non-generative control. **(a)** The change rate rises with the imposed bias while a random control (non-generative augmentation of the original) stays flat, indicating the signal reflects the view intervention rather than generic sensitivity to image perturbation. **(b)** Counterfactual AUC falls toward chance while original AUC declines more gradually, indicating that biased models increasingly rely on view-correlated features.

Here, $\hat{p}_e(a)$ approximates $P(e = 1 \mid a)$ and $\hat{p}_v(e)$ approximates $P(v = \mathrm{AP} \mid e)$. Each mechanism is parameterised by a separate two-layer perceptron and trained with binary cross-entropy on the observational NIH-14 training split. No image-model parameters are updated.

The edge $a \to e$ represents an assumed dependence of effusion prevalence on age. The edge $e \to v$ represents an assumed acquisition pathway in which the presence of effusion may be associated with reduced patient mobility and therefore a greater probability of AP acquisition. These edges are manually specified, rather than identified from the observational metadata. Counterfactual inference follows abduction–action–prediction (Pawlowski et al., 2020; Pearl, 2009).

Section 4.3 evaluates propagation under $\mathrm{do}(e)$. In this query, the intervention replaces the effusion mechanism, severing the incoming edge $a \to e$, while allowing the intervention to propagate through $e \to v$. Although the SCM also supports propagation under $\mathrm{do}(a)$, the fitted $a \to e$ mechanism produces too few realised effusion changes under our age interventions to support stable image-level evaluation.

Table 13: **SCM-guided propagation realises counterfactual image generation under a specified SCM.** Both query types are evaluated on the same randomly selected $1{,}500$ samples, stratified by whether the SCM-guided query flips the view. Effectiveness is measured by the pseudo-oracles of Table 1. KID is scaled by $\times 10^2$. The better value per column within each block is in bold.

		Effectiveness				Composition	Reversibility	Realism
Cases	Query	Acc(v) ↑	Acc(e) ↑	Acc(s) ↑	MAE(a) ↓	LPIPS ↓	LPIPS ↓	KID ↓
View changed	Direct	1.000 $_{(0.000)}$	0.827 $_{(0.042)}$	0.961 $_{(0.022)}$	**0.042** $_{(0.051)}$	0.014 $_{(0.009)}$	**0.021** $_{(0.025)}$	**0.30** $_{(0.20)}$
$n = 307$	SCM-guided	1.000 $_{(0.000)}$	0.827 $_{(0.042)}$	**0.964** $_{(0.021)}$	0.067 $_{(0.068)}$	0.014 $_{(0.009)}$	0.038 $_{(0.050)}$	4.80 $_{(0.70)}$
View unchanged	Direct	0.995 $_{(0.004)}$	0.873 $_{(0.019)}$	0.966 $_{(0.010)}$	0.041 $_{(0.045)}$	0.014 $_{(0.012)}$	0.021 $_{(0.025)}$	0.40 $_{(0.20)}$
$n = 1{,}193$	SCM-guided	0.995 $_{(0.004)}$	0.873 $_{(0.019)}$	0.966 $_{(0.010)}$	0.041 $_{(0.045)}$	0.014 $_{(0.012)}$	0.021 $_{(0.025)}$	0.40 $_{(0.20)}$

Table 14: **Full comparison with Causal-Adapter (Tong et al., 2026), including the authors' original SD1.5-backbone implementation.** Extends Table 3 with the SD1.5 Causal-Adapter rows. SD1.5 Causal-Adapter is pretrained on a different backbone (Stable Diffusion 1.5) from every other row (RadEdit), so differences in effectiveness confound backbone and adaptation method; RadEdit Causal-Adapter is the controlled comparison. KID is scaled by $\times 10^2$. Best per column metrics are in bold.

		Effectiveness				Composition	Reversibility	Realism
Intervention	Method	Acc(v) ↑	Acc(e) ↑	Acc(s) ↑	MAE(a) ↓	LPIPS ↓	LPIPS ↓	KID ↓
do(v)	RadEdit Causal-Adapter	0.563 $_{(0.031)}$	0.658 $_{(0.029)}$	0.963 $_{(0.012)}$	0.073 $_{(0.061)}$	0.094 $_{(0.072)}$	0.278 $_{(0.091)}$	4.64 $_{(0.61)}$
	SD1.5 Causal-Adapter (Tong et al., 2026)	**0.939** $_{(0.015)}$	**0.776** $_{(0.026)}$	**0.968** $_{(0.011)}$	**0.060** $_{(0.050)}$	0.113 $_{(0.038)}$	0.231 $_{(0.059)}$	4.71 $_{(0.55)}$
	RadEdit M-SPEC	0.600 $_{(0.030)}$	0.670 $_{(0.029)}$	0.947 $_{(0.014)}$	0.067 $_{(0.056)}$	**0.061** $_{(0.057)}$	**0.202** $_{(0.067)}$	**2.69** $_{(0.50)}$
do(e)	SDXL	0.748 $_{(0.027)}$	**0.776** $_{(0.027)}$	0.680 $_{(0.029)}$	0.150 $_{(0.115)}$	**0.015** $_{(0.003)}$	0.488 $_{(0.099)}$	3.54 $_{(0.39)}$
	RadEdit	0.995 $_{(0.004)}$	0.566 $_{(0.031)}$	0.972 $_{(0.010)}$	0.070 $_{(0.056)}$	0.027 $_{(0.015)}$	**0.107** $_{(0.050)}$	**1.08** $_{(0.24)}$
	RadEdit Causal-Adapter	0.979 $_{(0.009)}$	0.531 $_{(0.032)}$	0.974 $_{(0.010)}$	0.065 $_{(0.058)}$	0.094 $_{(0.072)}$	0.223 $_{(0.108)}$	1.85 $_{(0.42)}$
	SD1.5 Causal-Adapter (Tong et al., 2026)	0.996 $_{(0.004)}$	0.493 $_{(0.032)}$	0.970 $_{(0.011)}$	**0.048** $_{(0.040)}$	0.113 $_{(0.038)}$	0.181 $_{(0.061)}$	2.79 $_{(0.39)}$
	RadEdit M-SPEC	**1.000** $_{(0.000)}$	0.682 $_{(0.030)}$	**0.976** $_{(0.009)}$	0.053 $_{(0.043)}$	0.061 $_{(0.057)}$	0.174 $_{(0.070)}$	1.40 $_{(0.39)}$
do(s)	RadEdit Causal-Adapter	0.986 $_{(0.007)}$	0.768 $_{(0.026)}$	0.160 $_{(0.023)}$	0.063 $_{(0.054)}$	0.094 $_{(0.072)}$	0.186 $_{(0.106)}$	1.64 $_{(0.45)}$
	SD1.5 Causal-Adapter (Tong et al., 2026)	0.996 $_{(0.004)}$	**0.871** $_{(0.021)}$	**0.729** $_{(0.028)}$	**0.047** $_{(0.039)}$	0.113 $_{(0.038)}$	0.190 $_{(0.063)}$	2.46 $_{(0.42)}$
	RadEdit M-SPEC	**0.997** $_{(0.003)}$	0.839 $_{(0.023)}$	0.547 $_{(0.031)}$	0.048 $_{(0.039)}$	**0.061** $_{(0.057)}$	**0.150** $_{(0.070)}$	**0.75** $_{(0.26)}$
do(a)	RadEdit Causal-Adapter	0.981 $_{(0.008)}$	0.790 $_{(0.025)}$	**0.976** $_{(0.009)}$	0.200 $_{(0.133)}$	0.094 $_{(0.071)}$	0.213 $_{(0.107)}$	1.67 $_{(0.39)}$
	SD1.5 Causal-Adapter (Tong et al., 2026)	**0.997** $_{(0.003)}$	**0.874** $_{(0.021)}$	0.975 $_{(0.010)}$	**0.172** $_{(0.108)}$	0.113 $_{(0.038)}$	0.185 $_{(0.059)}$	2.57 $_{(0.37)}$
	RadEdit M-SPEC	0.993 $_{(0.005)}$	0.848 $_{(0.022)}$	0.962 $_{(0.012)}$	0.191 $_{(0.126)}$	**0.062** $_{(0.058)}$	**0.143** $_{(0.077)}$	**0.57** $_{(0.22)}$

J.2 Random joint interventions

Our main results intervene on one parent at a time. Table 15 instead reports *random joint* interventions, in which k of the four parents are sampled uniformly per image and intervened on simultaneously. This is a capacity stress test independent of any graph.

Per-attribute effectiveness is essentially invariant to k: view accuracy stays at 0.998 to 0.993, and effusion and sex accuracy vary by under two points across $k = 2, 3, 4$ rather than degrading. Reversibility rises from 0.055 to 0.069 and KID from 0.01 to 0.03, both monotonically, as the counterfactual departs from the factual image along more axes.

Table 15: **Random joint interventions on k simultaneously edited parents.** Effectiveness is reported per attribute over the subset of samples in which that attribute was among the k intervened on. Composition and reversibility are means across interventions; unlike the other tables, KID is reported unscaled here. Subscripts give standard deviations (Appendix B.5).

	Acc(v) ↑	Acc(e) ↑	Acc(s) ↑	MAE(a) ↓	Comp. ↓	Rev. ↓	KID ↓
$k = 2$	0.998 $_{(0.044)}$	0.852 $_{(0.355)}$	0.870 $_{(0.336)}$	0.096 $_{(0.070)}$	0.016 $_{(0.010)}$	0.055 $_{(0.050)}$	0.01 $_{(0.00)}$
$k = 3$	0.993 $_{(0.082)}$	0.862 $_{(0.344)}$	0.856 $_{(0.351)}$	0.108 $_{(0.079)}$	0.016 $_{(0.010)}$	0.064 $_{(0.056)}$	0.02 $_{(0.00)}$
$k = 4$	0.993 $_{(0.083)}$	0.858 $_{(0.349)}$	0.887 $_{(0.317)}$	0.109 $_{(0.082)}$	0.016 $_{(0.010)}$	0.069 $_{(0.055)}$	0.03 $_{(0.01)}$

K Discussion on the break of Markovian assumption

The mechanism in Equation (1) assumes $\mathbf{u} \perp \mathbf{pa}$, a standard assumption in deep structural causal modelling (De Sousa Ribeiro et al., 2023; Monteiro et al., 2023; Pawlowski et al., 2020; Pearl, 2009;

Rasal et al., 2025; Ribeiro et al., 2025). Two questions it raises are worth separating: whether our estimator induces a violation, and whether the data satisfy the assumption in the first place.

The model does not induce a violation. A batch-wise optimal-transport coupling (Tong et al., 2024), for instance, pairs noise with data within each minibatch and thereby ties the inferred terminal noise to the sampled batch, which can reintroduce dependence between the abducted noise and the parents. RadCF interpolates from a parent-independent Gaussian prior under an independent coupling: in Equation (2) the terminal noise is drawn as $\mathbf{z}_1 \sim \mathcal{N}(0, \mathbf{I})$ independently of both the data encoding $\mathbf{z}_0$ and the conditioning $\mathbf{c} = \mathbf{pa}$, so prior and training coupling are parent-independent by construction.

We cannot rule out confounding in the data. Causal sufficiency is not testable from i.i.d. images and labels (Pan & Bareinboim, 2024). It underpins essentially all deep-SCM counterfactual image generation (De Sousa Ribeiro et al., 2023; Monteiro et al., 2023; Pawlowski et al., 2020; Rasal et al., 2025; Ribeiro et al., 2025), and as in that literature we target counterfactuals relative to the specified mechanism rather than uniquely identified real-world counterfactuals. Under unobserved confounding RadCF remains a counterfactual estimator with respect to the fitted mechanism, but is not guaranteed to recover the true clinical counterfactual.

Measuring the residual dependence. To measure the extent to which the abducted noise $\mathbf{u}$ retains parent information, we fit linear probes to predict each parent from $\mathbf{u}$ and report held-out probe AUC (Table 16). Independence by construction during training does not guarantee independence of the *abducted* noise at inference time, so we measure it directly. Here lower is better, and 0.5 would indicate no linearly decodable parent information.

The unconditional mechanism serves as a reference point: it has no parent conditioning, so parent-related variation has nowhere to go except $\mathbf{u}$, and its probe AUCs are correspondingly high. Specialisation lowers the predictability of every parent relative to it, indicating that conditioning moves parent-related variation out of the exogenous latent. We still acknowledge that the abducted latent is not perfectly independent of the parents, and that the Markovian assumption may be violated to some degree in practice. Thus, we rely on the axiomatic criteria to measure the quality of the counterfactuals relative to the fitted mechanism.

Table 16: **Conditioning moves parent information out of the exogenous latent, but not to chance.** Held-out linear-probe AUC for predicting each parent from the abducted noise $\mathbf{u}$; lower is closer to $\mathbf{u} \perp \mathbf{pa}$, 0.5 is chance. The unconditional mechanism is a reference point, not an upper bound. Best per column in bold.

Mechanism	View	Sex	Effusion
Unconditional	0.995	0.817	0.636
Image-space flow (Ribeiro et al., 2025)	0.972	0.832	0.833
S-SPEC	**0.967**	**0.761**	**0.571**

L LATENT CONFOUNDING AND WHAT THE AXIOMATIC CRITERIA DETECT

Appendix K shows that the abducted noise retains parent information, so causal sufficiency cannot be assumed for our clinical data. Here we introduce a confounder deliberately and ask what the axiomatic criteria report when it is left out of $\mathbf{pa}$. Morpho-MNIST admits exact ground-truth counterfactuals and lets us score the criteria against true counterfactual error; NIH ChestX-ray admits none and lets us ask whether the same criteria register a latent confounder in a realistic imaging setting.

L.1 MORPHO-MNIST: AXIOMATIC SOUNDNESS VS. COUNTERFACTUAL GROUND TRUTH

Morpho-MNIST (Castro et al., 2019) renders MNIST digits under controlled morphological transformations and, unlike medical imaging, admits exact ground-truth counterfactuals, so the criteria can be scored against true counterfactual error rather than used as proxies for it.

The SCM has three variables: digit identity $Y \in \{0, \dots, 9\}$, a binary thickness attribute $A \in \{0, 1\}$, and the rendered image $\mathbf{x}$. Y confounds A and $\mathbf{x}$: it fixes the digit shape and also sets the thickness assignment. We report the strongest association we train under, in which thickness agrees with digit group in 95% of training images and the confounder is therefore hardest to detect.

Both models use a lightweight SiT-S/2 (384 hidden dimensions, 12 layers, 6 heads) applied directly to single-channel 32×32 images with no VAE and no REPA alignment, trained for 50,000 steps with flow matching (linear interpolation, v-prediction, batch size 128, learning rate 10^{-4}). They differ only in $\mathbf{pa}$:

- **Sufficient ($Y \in \mathbf{pa}$).** The mechanism conditions on both A and Y, so the graph is causally sufficient and the Markovian assumption holds by construction.
- **Confounded (Y latent).** The mechanism conditions on A alone. Y stays visible in the image but absent from $\mathbf{pa}$, so all Y-dependent variation must be absorbed by the exogenous noise $\mathbf{u}$.

Table 17: **A latent confounder degrades ground-truth fidelity and the axiomatic criteria.** Lower is better for CF MSE, higher for every other column; best per column in bold.

Setting	CF MSE ↓	CF SSIM ↑	Comp. SSIM ↑	Rev. SSIM ↑	Effectiveness ↑	Digit pres. ↑
Confounded (Y latent)	5845	0.417	0.939	0.505	1.000	0.435
Sufficient ($Y \in \mathbf{pa}$)	**5568**	**0.456**	**0.967**	**0.672**	0.999	**0.809**

Table 17 shows that the axiomatic measurements used for evaluation align with ground truth measurements, and track genuine counterfactual failure. When a causally relevant variable is omitted, ground-truth counterfactual fidelity, preservation, and reversibility degrade.

L.2 Specialisation under hidden bias

We repeat the construction on NIH ChestX-ray8 (Wang et al., 2017) with $\mathbf{pa} = (v, e, s, a)$: view position $v \in \{\text{PA}, \text{AP}\}$, effusion $e \in \{0, 1\}$, sex $s \in \{\text{M}, \text{F}\}$, and age $a \in [0, 1]$. View and effusion are already dependent in clinical data, since AP acquisition accompanies higher effusion prevalence through patient acuity, and we amplify that dependence to a controlled strength b, the proportion of training images whose view and effusion agree. Evaluation uses one held-out set of $5{,}108$ images with equal counts in all four view $\times$ effusion cells.

Similarly, we also studied two scenarios: the sufficient model conditions on all of (v, e, s, a), and the confounded model drops v from $\mathbf{pa}$, leaving it latent despite its association with e in the training distribution. Both models run S-SPEC for 10,000 steps at each bias level and flip effusion at inference, holding the remaining modelled parents fixed.

Table 18: **A latent view confounder makes the effusion intervention sensitive to the induced association.** All rows use the same $5{,}108$ test images, balanced across the four view $\times$ effusion cells. Effectiveness and effusion score change measure the size of the apparent intervention rather than its correctness, so no preferred direction is marked; lower LPIPS is better.

Setting	Effectiveness	Effusion score change	Composition LPIPS ↓	Reversibility LPIPS ↓
Sufficient ($v \in \mathbf{pa}$), $b = 0.50$	0.345	0.055	0.064	0.086
Sufficient ($v \in \mathbf{pa}$), $b = 0.75$	0.323	0.054	0.063	0.083
Sufficient ($v \in \mathbf{pa}$), $b = 0.95$	0.303	0.051	0.063	0.087
Confounded (v latent), $b = 0.50$	0.440	0.082	0.063	0.085
Confounded (v latent), $b = 0.75$	0.424	0.074	0.064	0.091
Confounded (v latent), $b = 0.95$	0.522	0.121	0.063	0.104

The axiomatic evaluation provides a practical diagnostic when ground-truth clinical counterfactuals are unavailable. With v in $\mathbf{pa}$, raising the association from $b = 0.50$ to $b = 0.95$ leaves the intervention largely unchanged: effectiveness stays within 0.303–0.345, the effusion score change within 0.051–0.055, and reversibility LPIPS within 0.083–0.087. Across the same range, the confounded model's effusion score change rises from 0.082 to 0.121 and its reversibility LPIPS worsens from 0.085 to 0.104, while the corresponding quantities in the sufficient model each move by less than 0.005. Effectiveness follows the same asymmetry less cleanly, rising by 0.082 in the confounded model and drifting down by 0.042 in the sufficient model, though not monotonically. Composition LPIPS holds at 0.063–0.064 in every row, so none of this follows from null reconstruction quality. **Thus, for specialisation, bias is not the problem; unmodelled bias is.**