TECHNICAL REPORT
R-103
First version: Dec, 2023
Last version: Feb, 2024

# Counterfactual Image Editing

**Yushu Pan** [1] **Elias Bareinboim** [1]

## Abstract

Counterfactual image editing is an important task in generative AI, which asks how an image would look if certain features were different. The current literature on the topic focuses primarily on changing individual features while remaining silent about the causal relationships between these features, as present in the real world. In this paper, we formalize the counterfactual image editing task using formal language, modeling the causal relationships between latent generative factors and images through a special type of model called augmented structural causal models (ASCMs). Second, we show two fundamental impossibility results: (1) counterfactual editing is impossible from i.i.d. image samples and their corresponding labels alone; (2) even when the causal relationships between the latent generative factors and images are available, no guarantees regarding the output of the model can be provided. Third, we propose a relaxation for this challenging problem by approximating non-identifiable counterfactual distributions with a new family of **counterfactual-consistent estimators**. This family exhibits the desirable property of preserving features that the user cares about across both factual and counterfactual worlds. Finally, we develop an efficient algorithm to generate counterfactual images by leveraging neural causal models.

## 1. Introduction

Counterfactual reasoning is a critical component of our cognitive system. It is essential for solving various tasks, including assigning credit, determining blame and responsibility, understanding why events occurred in a particular way and articulating explanations, and generalizing across changing conditions and environments (Pearl & Mackenzie, 2018; Bareinboim et al., 2022; Correa et al., 2021a). More recently, there has been a growing interest in counterfactual questions regarding image generation and editing. For instance, one might ask "how would the image change had the dog been a cat?" or "What would the image look like had the person been smiling?". Addressing these prototypical counterfactual questions is challenging and requires the understanding of the causal relationships between the features, with practical applications in various downstream tasks, including data augmentation, fairness analysis, generalizability, and transportability (Bareinboim et al., 2015; Schölkopf et al., 2021; Lee et al., 2020; Mao et al., 2022).

Some initial methods for counterfactual image editing tasks typically involve searching for adversarial samples (Goyal et al., 2019b; Wang & Vasconcelos, 2020; Dhurandhar et al., 2018). For example, (Dhurandhar et al., 2018) proposed a minimum-edit counterfactual method that aims to identify the minimum and most effective perturbations needed to change the classifier's prediction. With the ability to generate high-quality synthetic images from a latent space through GANs (Brock et al., 2019; Karras et al., 2019), VAEs (Child, 2021; Vahdat & Kautz, 2020), and Diffusion Models (Ho et al., 2020; Song et al., 2021), recent approaches edit images by manipulating vectors in the latent space (Shen et al., 2020; Härkönen et al., 2020; Khorram & Fuxin, 2022; Chai et al., 2021).

More recently, text information has also been leveraged in image editing tasks. The image description in text is beneficial to the encoding process and guiding manipulations in the latent space (Radford et al., 2021; Avrahami et al., 2022; Crowson et al., 2022; Gal et al., 2022; Patashnik et al., 2021) and the natural editing instruction text can be directly used to prompt the transition from the original to the counterfactual images (Brooks et al., 2023). However, such approaches focus primarily on changing a single categorical label of a given image, and more fundamentally, do not take the causal relationships among the underlying generative factors into account. The next example illustrates the challenge when multiple features are involved in the generation.

**Example 1.1.** Consider an image dataset of human faces. Based on our understanding of human anatomy and facial expressions, we know that both $Gender$ and $Age$ do not causally affect each other while age does affect hair color. Meanwhile, the dataset collected has older males and younger females, i.e., there exists a strong correlation

[1]Department of Computer Science, Columbia University, New York, USA. Correspondence to: Yushu Pan <yushupan@cs.columbia.edu>, Elias Bareinboim <eb@cs.columbia.edu>.

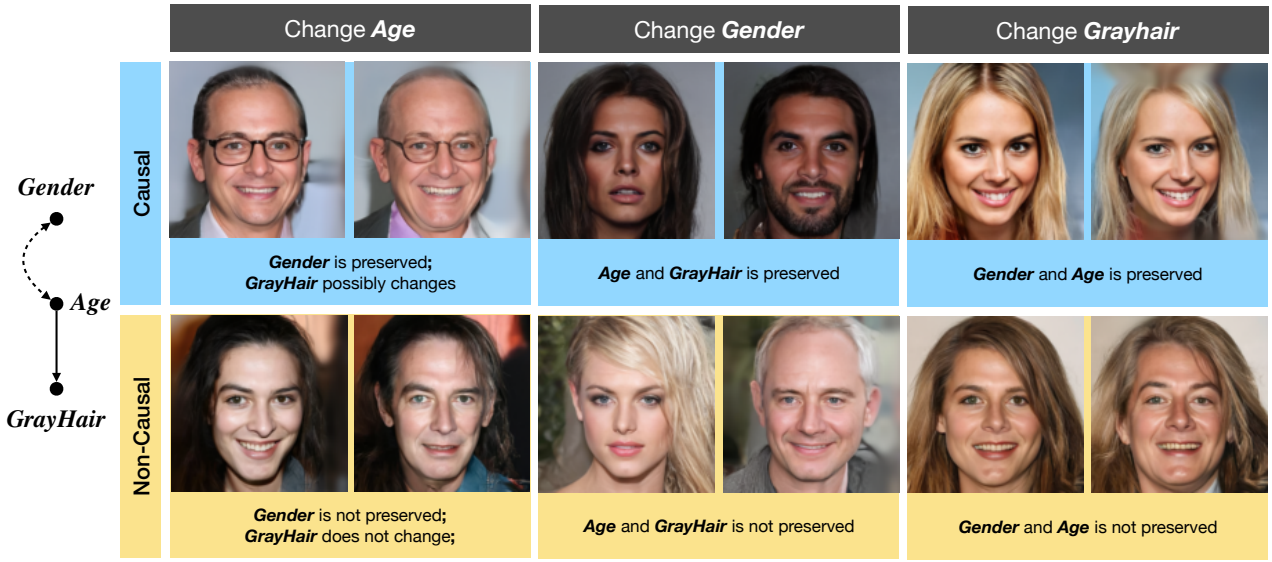



Figure 1: (Left) A causal graph depicting the causal relationships among features. (Right) Image editing results are displayed, with the first row showing edits incorporating causal relations, and the second row without them. Each column represents a unique counterfactual query, altering the age, gender, and gray hair of the individuals. These instances provide preliminary evidence that the causal approach introduced in this paper ensures the preservation of the relevant causal invariances for the query across both factual and counterfactual images.

between age and gender. Formally, the causal relationships between these three generative factors are shown in Fig. 1.

Existing methods focus on the editing of a single concept while the effects of the intervened concepts on others are not taken into account. Suppose we are evaluating the counterfactual query: "Given a certain image, what would the face look like had she been older?". If the age of the person is changed naively, gender and hair color may also change due to the correlation between these features found in the data. For example, when making an image of a woman older, it may inadvertently also change her gender to male; see the yellow row in Figure 1. However, it would be expected that changes in age should not affect gender when performing causal editing, as shown in the figure's first row (in blue).

More importantly, existing methods are unable to answer to **what extent hair color should change after an intervention on age.** Even though some recent methods may be able to enforce consistency in terms of gender, the causal effect from the age to the hair color may not be reflected correctly in the counterfactual images. For instance, gray hair may never appear after the editing by non-causal approaches. In contrast, causal image editing ensures the effects of target interventions on other features are carried over properly from factual to the proper counterfactual world. To illustrate, edits in Fig. 1 (blue) are more closely aligned with the reality in which these causal invariances are presented. ■

To capture the causal relationships among generative factors, we build on a class of generative models known as Structural Causal Models (SCMs) (Pearl, 2000). A fully instantiated SCM induces what is known as the Pearl Causal Hierarchy (PCH; also called *ladder of causation*) (Pearl & Mackenzie, 2018; Bareinboim et al., 2022). The PCH consists of families of distributions in increasing levels of refinement: layer 1 ($\mathcal{L}_1$) corresponds to passive observations and typical correlations, layer 2 ($\mathcal{L}_1$) to interventions (e.g., changing a variable to see the effect), and layer 3 ($\mathcal{L}_3$) to counterfactuals (e.g., considering what would happen under hypothetical scenarios). A result known as the causal hierarchy theorem states that higher-layer distributions cannot be answered only from the lower-layer ones (Bareinboim et al., 2022).

Recently, researchers have connected SCMs with deep generative models by implicitly finding surrogate models of the true generative model relating images and its generative factors. Despite the progress made so far, many of these works have limitations in different dimensions important in our context. First, they assume Markovianity, which implies the absence of unobserved confounding among generative factors. While this assumption may hold in specific settings, the same is certainly strong and does not hold in many others (Kocaoglu et al., 2018; Pawlowski et al., 2020; Sanchez & Tsaftaris, 2022; Sauer & Geiger, 2021).

Second, many of these works estimate counterfactual queries for images and generate samples without considering whether the target query is identifiable. In particular, samples are generated even though the query is non-identifiable, which implies that no guarantee can be pro-

vided in terms of the quality and causal consistency of the image. In particular, it is unclear whether the causal invariances present in the real systems are preserved across the original and generated images.

Third, other works focus on parametric SCMs over generative factors, such as linear mechanisms, while we study a more general class of non-parametric models (Yang et al., 2021; Shen et al., 2022). Recently, a new class of generative models has been developed, the Neural Causal Model (NCM), which encodes causal constraints into deep generative models. These models are capable of both identifying and then estimating counterfactual quantities in non-parametric settings (Xia et al., 2021; 2022). Despite the soundness of this approach to handling general, non-parametric variables in theory, it remains challenging to estimate counterfactual images, as the structure between generative factors and images is not taken into account and it's hard to scale these models to higher dimensions. Further discussion on related works is provided in Appendix C.

In this paper, we study the principles underpinning counterfactual image editing tasks and develop a practical, causally-grounded framework for these critical generative capabilities for high-dimensional settings. To achieve this goal, we formalize counterfactual image tasks according to augmented SCMs (ASCMs), a special class of SCMs taking the image generation step into account. This formulation allows for the formal encoding of causal relationships between generative factors and the image. It also enables modeling of the image editing tasks as querying counterfactual distributions induced by the true yet unknown ASCMs. More specifically, our contributions are as follows:

1. We formally show that image counterfactual distributions are almost never identifiable from only observational i.i.d image samples. Further, even when the causal relationships between generative factors and images are given, the target counterfactual distribution is still non-identifiable (Sec. 3).

2. We relax these settings and develop a new family of **counterfactual (Ctf-) consistent estimators** to approximate non-identifiable distributions. This provides the first procedure with formal guarantees of causal consistency w.r.t. the true generative model. With a sufficient condition to obtain Ctf-consistent estimators, we then develop an efficient algorithm (ANCMs) to sample counterfactual images in practice (Sec. 4). Extensive experiments are conducted to demonstrate the effectiveness of ANCMs (Sec. 5).

### 1.1. Preliminary

In this section, we provide the necessary background to understand this work. An uppercase letter $X$ indicates a random variable and a lowercase letter $x$ indicates its corresponding value; bold uppercase $\mathbf{X}$ denotes a set of random variables, and lowercase letter $\mathbf{x}$ is its corresponding values. We use $\mathcal{X}_X$ to denote the domain of $X$ and $\mathcal{X}_{\mathbf{X}} = \mathcal{X}_{X_1} \times \cdots \times \mathcal{X}_{X_d}$ for $\mathbf{X} = \{X_1, \ldots, X_d\}$. We denote $P(\mathbf{X})$ as a probability distribution over a set of random variables $\mathbf{X}$ and $P(\mathbf{X} = \mathbf{x})$ as the probability of $\mathbf{X}$ being equal to the value of $\mathbf{x}$ under the distribution $P(\mathbf{X})$.

Our work relies on the basic semantical framework structural causal models (SCMs) (Pearl, 2000, Ch. 7); we follow the presentation in (Bareinboim et al., 2022).

**Definition 1.2** (Structure Causal Model(SCM))**.** A Structure Causal Model (for short, SCM) is a 4-tuple $< \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) >$, where
(1) $\mathbf{U}$ is a set of background variables, also called exogenous variables, that are determined by factors outside the model;
(2) $\mathbf{V} = \{V_1, V_2, \ldots, V_d\}$ is the set of endogenous variables that are determined by other variables in the model;
(3) $\mathcal{F}$ is the set of functions $\{f_{V_1}, f_{V_2} \ldots, f_{V_d}\}$ mapping $\mathbf{U}_{V_j} \cup \mathbf{Pa}_{V_j}$ to $V_j$, where $\mathbf{U}_{V_j} \subseteq \mathbf{U}$ and $\mathbf{Pa}_{V_j} \subseteq \mathbf{V} \backslash V_j$;
(4) $P(\mathbf{U})$ is a probability function over the domain of $\mathbf{U}$. ■

Each SCM $\mathcal{M}$ induces a causal diagram $\mathcal{G}$, which is a directed acyclic graph where every $V_j$ is a vertex. There is a directed arrow from $V_j$ to $V_k$ if $V_j \in \mathbf{Pa}_{V_k}$. And there is a bidirected arrow between $V_j$ and $V_k$ if $\mathbf{U}_{V_j}$ and $\mathbf{U}_{V_k}$ are not independent with each other (Bareinboim et al., 2022, Def. 11).

An intervention on a subset of $\mathbf{X} \subseteq \mathbf{V}$, denoted by $do(\mathbf{x})$, is an operation where $\mathbf{X}$ takes value $\mathbf{x}$, regardless how $\mathbf{X}$ are originally defined. For an SCM $\mathcal{M}$, let $\mathcal{M}_{\mathbf{x}}$ be the submodel of $\mathcal{M}$ induced by $do(\mathbf{x})$. For any subset $\mathbf{Y} \subseteq \mathbf{V}$, the potential outcome $\mathbf{Y}_{\mathbf{x}}(\mathbf{u})$ is defined as the solution of $\mathbf{Y}$ after feeding $\mathbf{U} = \mathbf{u}$ into the submodel $\mathcal{M}_{\mathbf{x}}$. Then $\mathbf{Y}_{\mathbf{x}}$ is called a counterfactual variable induced by $\mathcal{M}$. Specifically, the event $\mathbf{Y}_{\mathbf{x}} = \mathbf{y}$ represent "$\mathbf{Y}$ would be $\mathbf{y}$ had $\mathbf{X}$ been $\mathbf{x}$". The counterfactual quantities induced by an SCM $\mathcal{M}$ are defined as (Bareinboim et al., 2022, Def. 7):

$$P^{\mathcal{M}}(\mathbf{y}_{\mathbf{x}}, \ldots, \mathbf{z}_{\mathbf{w}}) = \int_{\mathcal{X}_{\mathbf{U}}} \mathbb{1}_{\mathbf{Y}_{\mathbf{x}}(\mathbf{u})=\mathbf{y}, \ldots, \mathbf{Z}_{\mathbf{w}}(\mathbf{u})=\mathbf{z}} dP(\mathbf{u}), \tag{1}$$

where $\mathbf{Y}, \ldots, \mathbf{Z}, \mathbf{X}, \ldots, \mathbf{W} \subseteq \mathbf{V}$. Specifically, $P(\mathbf{Y}_{\mathbf{x}})$ reduces to an observational distribution $P(\mathbf{Y})$ taking $\mathbf{X}$ as an empty set.

Given the observed distribution $P(\mathbf{V})$ and causal diagram $\mathcal{G}$, the optimal counterfactual bounds are closed intervals based on the following optimization problem (Zhang et al., 2022).

**Definition 1.3** (Optimal Counterfactual Bounds)**.** For a causal diagram $\mathcal{G}$ and observed distributions $P(\mathbf{V})$, the *optimal bound* $[l, r]$ over a counterfactual probability $P^{\mathcal{M}}(\mathbf{y}_{\mathbf{x}}, \ldots, \mathbf{z}_{\mathbf{w}})$ is defined as, respectively, the minimum

and maximum of the following optimization problem:

$$\max_{\mathcal{M}\in\Omega(\mathcal{G})} / \min_{\mathcal{M}\in\Omega(\mathcal{G})} \; P^{\mathcal{M}}(\mathbf{y_x},\ldots,\mathbf{z_w}) \;\text{ s.t.} P^{\mathcal{M}}(\mathbf{V}) = P(\mathbf{V}) \quad (2)$$

where $\Omega(\mathcal{G})$ is the space of all SCMs that agree with the diagram $\mathcal{G}$, i.e., $\Omega(\mathcal{G}) = \{\forall\mathcal{M}|\mathcal{G}_{\mathcal{M}} = \mathcal{G}\}$. ■

By the formulation of Equation (1), all possible values of counterfactual query induced by SCMs that agree with the observational distributions and causal diagram are contained in the closed interval $[l, r]$.

We use neural causal models (NCMs) for estimating counterfactual distributions, which are defined as follows (Xia et al., 2021):

**Definition 1.4** ($\mathcal{G}$-Constrained Neural Causal Model ($\mathcal{G}$-NCM)**.** Given a causal diagram $\mathcal{G}$, a $\mathcal{G}$-constrained Neural Causal Model (for short, $\mathcal{G}$-NCM) $\widehat{M}(\boldsymbol{\theta})$ over variables $\mathbf{V}$ with parameters $\boldsymbol{\theta} = \{\theta_{V_i} : V_i \in \mathbf{V}\}$ is an SCM $\langle\widehat{\mathbf{U}}, \mathbf{V}, \widehat{\mathcal{F}}, \widehat{P}(\widehat{\mathbf{U}})\rangle$ such that $\widehat{\mathbf{U}} = \{\widehat{U}_{\mathbf{C}} : \mathbf{C} \subseteq \mathbf{V}\}$, where
(1) each $\widehat{U}$ is associated with some subset of variables $\mathbf{C} \subseteq \mathbf{V}$, and $\mathcal{D}_{\widehat{U}} = [0, 1]$ for all $\widehat{U} \in \widehat{\mathbf{U}}$;
(2) $\widehat{\mathcal{F}} = \{\hat{f}_{V_i} : V_i \in \mathbf{V}\}$, where each $\hat{f}_{V_i}$ is a feed forward neural network parameterized by $\theta_{V_i} \in \boldsymbol{\theta}$ mapping values of $\mathbf{U}_{V_i} \cup \mathbf{Pa}_{V_i}$ to values of $V_i$ for $\mathbf{U}_{V_i} = \{\widehat{U}_{\mathbf{C}} : \widehat{U}_{\mathbf{C}} \in \widehat{\mathbf{U}}$ s.t. $V_i \in \mathbf{C}\}$ and $\mathbf{Pa}_{V_i} = Pa_{\mathcal{G}}(V_i)$;
(3) $\widehat{P}(\widehat{\mathbf{U}})$ is defined s.t. $\widehat{U} \sim \text{Unif}(0, 1)$ for each $\widehat{U} \in \widehat{\mathbf{U}}$. ■

## 2. Augmented SCMs and Image Counterfactual Distributions

In this section, we model the image counterfactual editing problems in causal language. We first define a special type of SCM to model the generation process of an image variable $\mathbf{I}$, which is called the Augmented SCM (ASCM).

**Definition 2.1** (Augmented Structure Causal Model)**.** An Augmented Structure Causal Model (for short, ASCM) over a generative level SCM $\mathcal{M}_0 = \langle\{\mathbf{U}_0, \mathbf{V}_0, \mathcal{F}_0, P^0(\mathbf{U}_0)\}\rangle$ is a tuple $\mathcal{M} = \langle\mathbf{U}, \{\mathbf{V}, \mathbf{I}\}, \mathcal{F}, P(\mathbf{U})\rangle$ such that
(1) exogenous variables $\mathbf{U} = \{\mathbf{U}_0, \mathbf{U_I}\}$;
(2) $\mathbf{V} = \mathbf{V}_0$ are labeled observed endogenous variables; $\mathbf{I}$ is an $m$ dimensional image variable;
(3) $\mathcal{F} = \{\mathcal{F}_0, f_{\mathbf{I}}\}$, where $f_{\mathbf{I}}$ maps from (the respective domains of) $\mathbf{V} \cup \mathbf{U_I}$ to $\mathbf{I}$, which is an invertible function regarding $\mathbf{V}$. Namely, there exists a function $h$ such that $\mathbf{V} = h(\mathbf{I})$.
(4) $P(\mathbf{U}_0) = P^0(\mathbf{U}_0)$. ■

The ASCM $\mathcal{M}$ is in fact a "larger" SCM describing a two-stage generative process, where first the low-dimensional generative factors are produced and second these generative factors are mapped to a high-dimensional image. More specifically, the $\mathbf{U_I}$ interact with labeled $\mathbf{V}$ to produce other unlabeled features $\tilde{\mathbf{U}}$ through part of $f_{\mathbf{I}}$ in the first stage. In the second stage, the remaining part of $f_{\mathbf{I}}$ mixes the observed $\mathbf{V}$ and unobserved generative factors $\tilde{\mathbf{U}}$ to create the image's set of pixels. Notice that $\tilde{U}$ is not a part of $\mathbf{U_I}$ since $\tilde{U}$ can be generated from $\mathbf{V}$ plus $\mathbf{U_I}$. Throughout this paper, we assume that domains of observed generative factors $\mathbf{V}$ are discrete and finite. An important aspect of $f_{\mathbf{I}}$ is that it is invertible regarding $\mathbf{V}$ since these generative factors $\mathbf{V}$ are present directly in a given image $\mathbf{i}$. This assumption is commonly used in non-linear ICA and representation learning literature (Locatello et al., 2019b; Lachapelle et al., 2021; Hyvärinen & Pajunen, 1999; Khemakhem et al., 2020). The inverse $h$ represents a labeling process that assigns the correct labels of $\mathbf{V}$ to $\mathbf{i}$. Then, for any $\mathbf{W} \subseteq \mathbf{V}$:

$$P(\mathbf{w} \mid \mathbf{i}) = \begin{cases} 1 & \mathbf{w} = h_{\mathbf{W}}(\mathbf{i}) \\ 0 & otherwise, \end{cases} \quad (3)$$

where $h_{\mathbf{W}}(\cdot)$ is the subfunction of $h$ mapping from $\mathbf{I}$ to $\mathbf{W}$. The next example illustrates the modeling of face images discussed earlier.

| $F$ | $Y$ | $H$ | $P(F,Y,H)$ |
|---|---|---|---|
| 0 | 0 | 0 | 0.216 |
| 0 | 0 | 1 | 0.144 |
| 0 | 1 | 0 | 0.128 |
| 0 | 1 | 1 | 0.032 |
| 1 | 0 | 0 | 0.144 |
| 1 | 0 | 1 | 0.096 |
| 1 | 1 | 0 | 0.192 |
| 1 | 1 | 1 | 0.048 |

Figure 2: $P(\mathbf{V})$ induced by the ASCM in Example. 2.2.

**Example 2.2.** (Example 1.1 continued). Now we consider the augmented generative process, ASCM $\mathcal{M}^* = \langle\mathbf{U} = \{U_F, U_Y, U_{H_1}, U_{H_2}, \mathbf{U_I}\}, \{\{F, H, Y\}, \mathbf{I}\}, \mathcal{F}^*, P^*(\mathbf{U})\rangle$, where the mechanisms

$$\mathcal{F}^* = \begin{cases} F \leftarrow U_F \oplus U_Y \\ Y \leftarrow U_Y \\ H \leftarrow (\neg Y \wedge U_{H_1}) \oplus (Y \wedge U_{H_2}) \\ \mathbf{I} \leftarrow f^{\text{face}}_{\mathbf{I}}(F, Y, H, \mathbf{U_I}) \end{cases} \quad (4)$$

and the exogenous variables $U_F, U_Y, U_{H_1}, U_{H_2}$ are independent binary variables, and $P(U_F = 1) = 0.4, P(U_Y = 1) = 0.4, P(U_{H_1} = 1) = 0.4, P(U_{H_2} = 1) = 0.2$. $\mathbf{U_I}$ can be correlated with $U_F, U_Y, U_{H_1}, U_{H_2}$.

The variable $F$ represents gender (male $F = 0$; female $F = 1$), $Y$ represents age (young $Y = 0$; old $Y = 1$), and $H$ represents whether the person has gray hair (gray $H = 1$; non-gray $H = 0$). The observational distribution $P(F, Y, H)$ induced by $\mathcal{M}^*$ is shown in Figure 2. From the distribution, we can see that $Y = 1$ and $H = 1$ are

positively correlated ($P(Y=1,F=1) > P(Y=1,F=0), P(Y=0,F=1) < P(Y=0,F=0)$), and older people are more likely to have gray hair $P(H=1 \mid Y=0) > P(H=1 \mid Y=1)$.

Before the image is taken, $\mathbf{U_I}$ and $\{F, Y, H\}$ produce other unobserved generative factors $\tilde{\mathbf{U}}$, such as wrinkles, smiling, and narrow eyes at the generative level. Among them, some factors (such as wrinkles) can be produced by both $\mathbf{V}$ and $\mathbf{U_I}$, and some other factors (such as smiling) can be only produced by $\mathbf{U_I}$. Then, $f_{\mathbf{I}}$ maps all generative factors (including unobserved and observed ones) to image pixels $\mathbf{I}$ at the second stage. Looking at the image, $\{F, Y, H\}$ are deterministic and one can in principle label them through function $h$, the inverse of $f_{\mathbf{I}}^{\text{face}}$ $\{F, Y, H\}$. ■

Equipped with ASCMs, we now formalize the counterfactual image generation tasks through formal causal semantics. Suppose the true underlying ASCM is given by $\mathcal{M}^*$, which is unobserved. The goal is to query a specific type of counterfactual distribution induced by $\mathcal{M}^*$ given the input distribution $P(\mathbf{V}, \mathbf{I})$, i.e., $P^{\mathcal{M}^*}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$, where $\mathbf{X} \subseteq \mathbf{V}$. Factorizing this joint probability distribution, we can write:

$$\begin{aligned} &P^{\mathcal{M}^*}(\mathbf{I}=\mathbf{i}, \mathbf{I}_{\mathbf{x}'}=\mathbf{i}') \\ &= P^{\mathcal{M}^*}(\mathbf{I}=\mathbf{i}) P^{\mathcal{M}^*}(\mathbf{I}_{\mathbf{x}'}=\mathbf{i}' \mid \mathbf{I}=\mathbf{i}), \end{aligned} \tag{5}$$

this $\mathcal{L}_3$-quantity can be explained as follows. The initial image $\mathbf{i}$ is sampled from $P^{\mathcal{M}^*}(\mathbf{I})$ and the goal is to edit $\mathbf{i}$ to a counterfactual version $\mathbf{i}'$ with modified features $\mathbf{X} = \mathbf{x}'$, where $\mathbf{i}'$ is sampled from $P^{\mathcal{M}^*}(\mathbf{I}_{\mathbf{x}'} \mid \mathbf{I}=\mathbf{i})$ [1]. For example, the distribution $P^{\mathcal{M}^*}(\mathbf{I}, \mathbf{I}_{Y=0})$ (induced by the ASCM introduced in Example 2.2) can answer the query "generate an image describing people's face and edit the face to make the person look older".

Throughout this paper, we call this type of $\mathcal{L}_3$-distributions as *Image Counterfactual Distributions*. A particular instantiation of the image variable, such as $P(\mathbf{I}=\mathbf{i}, \mathbf{I}_{\mathbf{x}'}=\mathbf{i}')$, is called on *Image Counterfactual Query*. The explanation of image counterfactual distributions at the generative level is that given all generative factors in the initial images, what would they be had $\mathbf{X}$ taken value $\mathbf{x}'$. For instance, $P^{\mathcal{M}^*}(\mathbf{I}, \mathbf{I}_{Y=0})$ is asking what would observed factors (gender, hair color) and unobserved factors (wrinkles, smiling, narrow eyes, ...) be had the person been older.

## 3. Non-identifiability of Image counterfactual Distributions

In classic counterfactual image generation tasks, a generator $\widehat{\mathcal{M}}$ is trained to match the distribution $P(\mathbf{V}, \mathbf{I})$ (e.g., through a GAN), and then the pair of an initial image and its counterfactual can be sampled from $P^{\widehat{\mathcal{M}}}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ induced by the generator (see Figure 3a bottom). For concreteness, after sampling an initial image $\mathbf{i}$, one can get the counterfactual image $\mathbf{i}_{\mathbf{x}'}$ by manipulating the latent space or conditional signal of $\widehat{\mathcal{M}}$. However, as alluded to earlier, the Causal Hierarchy Theorem (Bareinboim et al., 2022, Thm. 1) states that counterfactual distributions cannot be computed merely from correlations. In particular, we show next the (non-)identifiability of any image counterfactual query from pure observational data:

**Corollary 3.1** (Image Causal Hierarchy Theorem)**.** *Any image counterfactual distribution is almost never uniquely computable from the observational distribution (or its samples).* ■

In other words, Corol. 3.1 states that $P^{\widehat{\mathcal{M}}}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ induced by the proxy generator may not be consistent with the true $P^{\mathcal{M}^*}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ even when the proxy generator fits the observed distributions perfectly (i.e., $P^{\widehat{\mathcal{M}}}(\mathbf{V}, \mathbf{I}) = P^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I})$). This inconsistency implies the effect of intervention $\mathbf{X} = \mathbf{x}'$ on other generative factors (features) may differ from the true model and the proxy generator.

To illustrate this issue, suppose the target query is $P^{\mathcal{M}^*}(\mathbf{I}, \mathbf{I}_{Y=0})$, where $\mathcal{M}^*$ is defined in Example 2.2. As shown in Figure 3b and the next example, even when $\widehat{\mathcal{M}}$ matches the observational distribution of the true underlying model $\mathcal{M}^*$, $\widehat{\mathcal{M}}$ can be less likely to generate counterfactual images with gray hair than $\mathcal{M}^*$ after the intervention $Y=0$.

**Example 3.2.** Consider an ASCM $\mathcal{M}'$ that is exactly the same as $\mathcal{M}^*$ introduced in Example 2.2 except that

$$f'_H = (\neg U_Y \wedge U_{H_1}) \oplus (U_Y \wedge U_{H_2}) \tag{6}$$

$\mathcal{M}'$ implies the same $P(\mathbf{V})$ as shown in Figure 2. Then, it's immediately verifiable that $P^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I}) = P^{\mathcal{M}'}(\mathbf{V}, \mathbf{I})$.

Now consider the counterfactual query $P(\mathbf{i}, \mathbf{i}_{Y=0})$, where $\mathbf{i}$ is a young male without gray hair (with generated features $F=0, Y=1, H=0$) and $\mathbf{i}'$ is an old male with gray hair (with features $F=0, Y=0, H=1$). In $\mathcal{M}^*$, $H=0$ will change to $H=1$ and $F$ will remains invariant after the intervention $do(Y=1)$ with probability 0.4, i.e.:

$$P^{\mathcal{M}^*}(F_{Y=0}=0, H_{Y=0}=1 \mid F=0, Y=1, H=0) = 0.4. \tag{7}$$

However, $H$ will never change after the same intervention in $\mathcal{M}'$ since the input of $f'_H$ does not involve $Y$, i.e.,

$$P^{\mathcal{M}'}(F_{Y=0}=0, H_{Y=0}=1 \mid F=0, Y=1, H=0) = 0. \tag{8}$$

Consequently, the true model suggests that the hair color is likely to change after making a young male look older

[1] $P^{\mathcal{M}^*}(\mathbf{I}_{\mathbf{x}'} \mid \mathbf{I}=\mathbf{i})$ serves for editing real images when the initial image $\mathbf{i}$ is a real one given by an user.

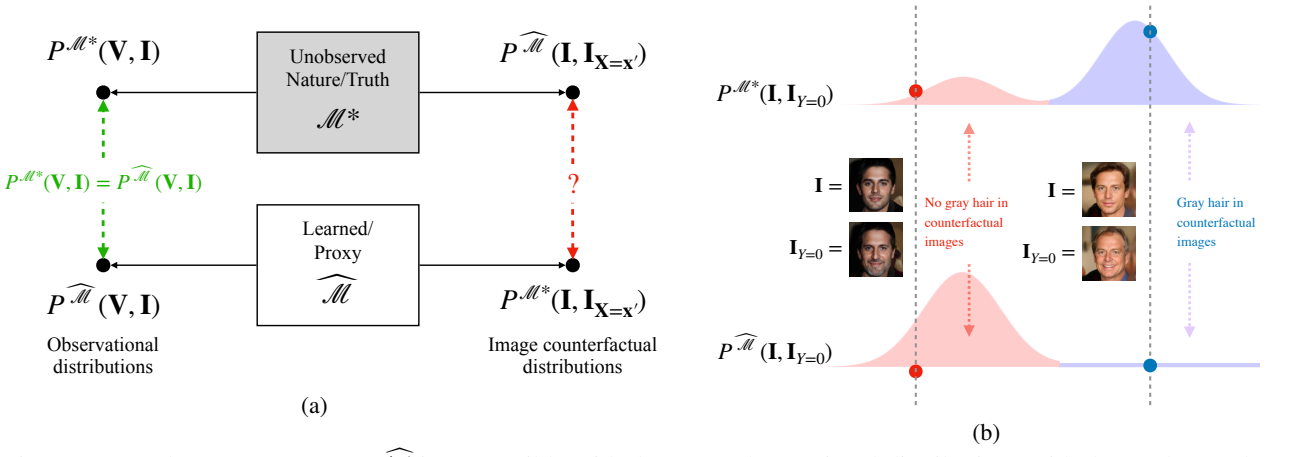



Figure 3: (a) The proxy generator $\widehat{\mathcal{M}}$ is compatible with the same observational distributions with the unobserved true model but is not guaranteed to induce the same image counterfactual distributions. (b) Two different image counterfactual distributions in Example 3.2. Each sample from a $P(\mathbf{I}, \mathbf{I}_{Y=0})$ has an initial image $\mathbf{i}$ and a counterfactual image $\mathbf{i}'_{Y=0}$. Sampling from the red part of distributions, counterfactual images do not contain gray hair. Sampling from the blue part of distributions, counterfactual images have gray hair.

with probability $0.4$ (see blue part of the distribution in the upper part of Figure 3b) while the counterfactual image generated by the proxy model would have gray hair with zero probability (blue part of upper distribution in Figure 3b). This is an instance of the aforementioned non-identifiability result (Corol. 3.1). ■

The main issue is that various generative models are capable of producing the same observational image distribution, yet they can yield qualitatively distinct counterfactual images. Broadly speaking, there is nothing in the observational distribution that indicates how an image would change under a hypothetical interventional scenario, so the counterfactual distribution remains undetermined by the observational one.

### 3.1. Identification of Image Counterfactual Distributions with Causal Diagrams

One of the realizations from the broader causal inference literature is that further assumptions are needed in order to perform counterfactual reasoning. In this section, we will leverage the causal diagram of the true underlying ASCM to discuss whether an image counterfactual distribution is uniquely computable from a combination of these assumptions and these observational distributions.

A causal diagram encodes constraints over counterfactual distributions compatible with the true and unobserved ASCM, narrowing down the hypothesis space of the proxy generator (Bareinboim et al., 2022, Sec. 1.4). It can be obtained from prior information about concepts in images. For instance, the qualitative understanding that getting older likely leads to gray hair suggests that there should be a direct edge from $Y$ to $H$ in Example 2.2. Causal diagrams

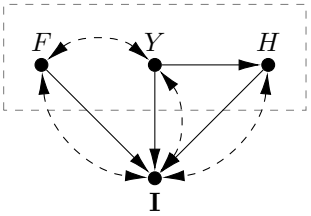



Figure 4: The causal diagram of the in $\mathcal{M}^*$ in Example 2.2

can be regarded as a causal inductive bias based on human knowledge. The complete causal diagram induced by $\mathcal{M}^*$ is shown in Figure 4; the diagram induced by $\mathcal{M}^*_0$, at the generative level, is in the dashed box. To illustrate, direct edges from $\{F, Y, H\}$ mean that these generative factors construct the image $\mathbf{I}$. The bidirected edge between $F$ and $Y$ encodes that gender and age in the dataset collected are confounded. Bidirected edges between one of the generative factors $\{F, Y, H\}$ and the images imply that some unobserved generative factors can directly affect or be confounded with $\{F, Y, H\}$.

Once qualitative knowledge about the generative process is encoded in the causal model, our new goal is to infer a target image counterfactual query $P^{\mathcal{M}^*}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ given a causal diagram $\mathcal{G}$ over $\{\mathbf{V}, \mathbf{I}\}$ and observational distributions $P(\mathbf{V}, \mathbf{L})$. We next define the notation of identifiability in the context of ASCMs.

**Definition 3.3** (Identifiability)**.** Consider the true underlying ASCM $\mathcal{M}^*$ defined over $\{\mathbf{V}, \mathbf{I}\}$ and the corresponding causal diagram $\mathcal{G}$ and observational distribution $P(\mathbf{V}, \mathbf{I})$. An image counterfactual query $P(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$ is said to be identifiable from the input $\langle P(\mathbf{V}, \mathbf{I}), \mathcal{G} \rangle$ if $P^{\mathcal{M}^{(1)}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'}) =$

$P^{\mathcal{M}^{(2)}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$ for every pair of ASCMs $\mathcal{M}^{(1)}, \mathcal{M}^{(2)} \in \Omega_{\mathbf{I}}(\mathcal{G})$ s.t. $P^{\mathcal{M}^{(1)}}(\mathbf{V}, \mathbf{I}) = P^{\mathcal{M}^{(2)}}(\mathbf{V}, \mathbf{I})$, where $\Omega^{\mathbf{I}}$ is the space of ASCMs. The distribution $P(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ is said to be identifiable if $P(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$ is identifiable for every $\mathbf{i}, \mathbf{i}' \in \mathcal{X}_{\mathbf{I}}$. ■

Compared to the previous definition of identifiability used in causal inference (e.g., (Pearl, 2009, Ch. 3)), Def. 3.3 restricts the space of SCMs to the space of ASCMs and considers image counterfactual queries. The identifiability of $P(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ is equivalent to saying that $P(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ is uniquely computable given the observational distribution and the graphical constraints encoded in $\mathcal{G}$. If satisfied, any proxy model $\widehat{\mathcal{M}}$ that is compatible with $P(\mathbf{V}, \mathbf{I})$ and $\mathcal{G}$ could be used to evaluate $P^{\mathcal{M}^*}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ when the query is identifiable. However, the following proposition implies that even with prior causal information about $\mathbf{V}$ as encoded in $\mathcal{G}$, $P(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$ is still not identifiable.

**Theorem 3.4** (ID)**.** *The image counterfactual distribution $P(\mathbf{I}, \mathbf{I}'_{\mathbf{x}'})$ is not identifiable from any combination of $\langle P(\mathbf{V}, \mathbf{I}), \mathcal{G} \rangle$.* ■

This non-identifiability challenge comes from two perspectives. First, it is unknown how $U_{\mathbf{I}}$ interacts with $\mathbf{V}$ to produce unobserved factors $\tilde{\mathbf{U}}$ while these interactions have implications for determining how the counterfactual image should look like. The next example illustrates this point.

**Example 3.5.** We split $\mathbf{U_I}$ in Example 2.2 into $\{U_S, \mathbf{U}^-_{\mathbf{I}}\}$, where $U_S$ controls the smiling generative factor and $\mathbf{U}^-_{\mathbf{I}}$ contributes to all other unobserved generative factors. Consider two ASCMs $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ with the same $\mathcal{F}\backslash\{f_{\mathbf{I}}\}$ but different $f_{\mathbf{I}}$ from $\mathcal{M}^*$ defined in Example 2.2:

$$\begin{aligned} f^{(1)}_{\mathbf{I}}(F, Y, H, U_S, \mathbf{U}^-_{\mathbf{I}}) &= f^{\text{s}}_{\mathbf{I}}(F, Y, H, U_S, \mathbf{U}^-_{\mathbf{I}}), \\ f^{(2)}_{\mathbf{I}}(F, Y, H, U_S, \mathbf{U}^-_{\mathbf{I}}) &= f^{\text{s}}_{\mathbf{I}}(F, Y, H, U_S \oplus Y, \mathbf{U}^-_{\mathbf{I}}), \end{aligned} \tag{9}$$

where $U_S$ is a fair coin and is independent with $\mathbf{U}\backslash\{U_S\}$. $\mathbf{U}^-_{\mathbf{I}}$ can be correlated with $U_Y, U_F, U_H$. $f^{\text{s}}_{\mathbf{I}}$ is the same in both $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ mapping from $\{F, Y, H, S, \mathbf{U}^-_{\mathbf{I}}\}$ to $\mathbf{I}$, where $S = U_S$ in $\mathcal{M}^{(1)}$ and $S = U_S \oplus Y$ in $\mathcal{M}^{(2)}$. $f^{\text{s}}_{\mathbf{I}}$ produces a smiling person image if and only if $S = 1$. $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ are compatible with graphical constraints encoded in the causal diagram shown in Fig. 4, and it is verifiable that $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ induce the same observational distributions.

Consider the counterfactual image query $P(\mathbf{i}, \mathbf{i}_{Y=0})$, where $\mathbf{i}$ is a non-smiling young male and $\mathbf{i}'$ is a smiling old male with gray hair. In $\mathcal{M}^{(1)}$, changing $Y$ to 0 will not affect the value of $S$ while changing $Y$ to 0 will always flip the value of $S$ in $\mathcal{M}^{(2)}$. This implies that $P^{\mathcal{M}^{(1)}}(\mathbf{i}, \mathbf{i}'_{Y=0})$ is the same as $P^{\mathcal{M}^{(2)}}(\mathbf{i}, \mathbf{i}'_{Y=0})$. ■

Second, the other perspective follows that given the observed values of a generative factor $X$ and its child $Y$, $P(y'_{x'} \mid y, x)$ is never point identifiable from the observational distribution. The next example illustrates this point.

**Example 3.6.** Consider an ASCM $\mathcal{M}^{(3)}$ that is exactly the same as $\mathcal{M}^*$ defined in Example 2.2 except for

$$f^{(3)}_H \leftarrow ((\neg Y) \wedge U_{H_1}) \vee U_{H_2} \tag{10}$$

and $P(U_{H_1} = 1) = 0.25$. Then, it is verifiable that $P^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I}) = P^{\mathcal{M}^{(3)}}(\mathbf{V}, \mathbf{I})$ and $\mathcal{M}^{(3)}$ is compatible with the graphical constraints induced by the model in Figure 4.

Consider the same counterfactual image query $P(\mathbf{i}, \mathbf{i}_{Y=0})$, and note that $H = 0$ will change to $H = 1$,

$$P^{\mathcal{M}^{(3)}}(F_{Y=0} = 0, H_{Y=0} = 1 \mid F = 0, Y = 1, H = 0) = 0.25, \tag{11}$$

after the intervention $do(Y = 0)$ with probability 0.25 in $\mathcal{M}^{(3)}$, which is different from the same quantity induced by $\mathcal{M}^*$ (Equation (7)). This implies $P^{\mathcal{M}^{(3)}}(\mathbf{i}, \mathbf{i}'_{Y=0})$ is not equal to $P^{\mathcal{M}^*}(\mathbf{i}, \mathbf{i}'_{Y=0})$. ■

## 4. Counterfactually consistent estimation of Image Counterfactual Distributions

We have seen so far that no image counterfactual distribution is identifiable given the causal diagram and the observational distribution alone. A question naturally arises considering this situation: can these non-identifiable distributions be estimated in any reasonable way? In other words, when the proxy generator ($\widehat{\mathcal{M}}$) does not induce the exact same image counterfactual distributions, what tolerance could be acceptable between $P^{\widehat{\mathcal{M}}}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ and the true $P^{\mathcal{M}^*}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$? In addition, we need an estimator to guarantee the approximation of $P^{\mathcal{M}^*}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ be within the tolerance no matter what causal relationships among generative factors are. To achieve this, we propose the following two directions to relax the exact estimation of query $P^{\mathcal{M}^*}(\mathbf{i}, \mathbf{i}_{\mathbf{x}'})$ while retaining causal principles and reasonable results.

(1) **Care set W.** As illustrated in Sec. 2, $P^{\mathcal{M}^*}(\mathbf{i}, \mathbf{i}_{\mathbf{x}'})$ takes into account how all generative factors ($\{\mathbf{V}, \tilde{\mathbf{U}}\}$) in an image would change after the intervention $do(\mathbf{X} = \mathbf{x})$ takes place. Still, in some practical situations, one may only be concerned about how some specific features behave after the intervention but not the whole image. In Example 2.2, all facial features should change causally after making the person older. To illustrate, the intervention on age should preserve the gender and smiling status, and change the hair color with probability 0.4 (Equation (7)). However, in practice, one may only care about the gender and age (i.e., $\mathbf{W} = \{F, Y\}$) after the intervention, but not whether the hair color, smiling status, and background in the image are presented the same way or not. If so, the counterfactual image can have gray hair features and smiling features with arbitrary probability. We introduce the following definition to describe the

counterfactual distributions among these selected features regarding an image counterfactual query.

**Definition 4.1** (Feature Counterfactual Query)**.** Denote $\mathbf{W}$ as a set of features one cares about and $\phi$ as a function mapping from $\mathbf{I}$ to $\mathbf{W}$ ($\mathbf{W} = \phi(\mathbf{I})$). The feature counterfactual query regarding to $P(\mathbf{i}, \mathbf{i}_{\mathbf{x}'})$ is defined as:

$$\int_{\mathbf{i}^{(1)}, \mathbf{i}^{(2)} \in \mathcal{X}_{\mathbf{I}}} \mathbf{1}\left[\phi(\mathbf{i}^{(1)}) = \mathbf{w}, \phi(\mathbf{i}^{(2)}) = \mathbf{w}'\right] dP(\mathbf{i}^{(1)}, \mathbf{i}^{(2)}_{\mathbf{x}'}) \quad (12)$$

where $\mathbf{w} = \phi(\mathbf{i})$, and $\mathbf{w}' = \phi(\mathbf{i}')$. We denote the feature counterfactual query as $\phi(P(\mathbf{i}, \mathbf{i}'_{x'}))$. ■

In other words, the feature counterfactual query is a push-forward measure from $P(\mathbf{i}, \mathbf{i}'_{x'})$ through $\phi$. The quantity in Eq. 12 integrates over all $P(\mathbf{i}^{(1)}, \mathbf{i}^{(2)}_{\mathbf{x}'})$ such that $\{\mathbf{i}^{(1)}, \mathbf{i}^{(2)}\}$ has the same cared features $\{\mathbf{w}, \mathbf{w}'\}$ with $\{\mathbf{i}, \mathbf{i}'\}$ in the target query. For concreteness, consider the counterfactual image query $P(\mathbf{i}, \mathbf{i}_{Y=0})$, where $\mathbf{i}$ is a smiling young male without gray hair and $\mathbf{i}'$ is a smiling old male with gray hair. Suppose the care set $\mathbf{W}$ contains the features gender ($F$) and age ($Y$). The feature counterfactual query $\phi(P(\mathbf{i}, \mathbf{i}'_{x'}))$ calculates the probability that the original image describes a young male and the counterfactual image describes an old male after editing. Following Equation (12), $\phi(P(\mathbf{i}, \mathbf{i}'_{x'}))$ sums over $P(\mathbf{i}^{(1)}, \mathbf{i}^{(2)}_{\mathbf{x}'})$, where $\mathbf{i}^{(1)}$ describes a young male, $\mathbf{i}^{(2)}$ describes an old male. In addition, $\mathbf{i}^{(1)}$ and $\mathbf{i}^{(2)}$ can have arbitrary hair and smiling features since those are not part of $\mathbf{W}$. Then, the feature counterfactual query induced by a proxy ASCM can be simplified using the following result.

**Lemma 4.2.** *Consider the true underlying ASCM $\mathcal{M}^*$ over $\{\mathbf{V}, \mathbf{I}\}$, and a feature set with mapping function $\phi = h^*_{\mathbf{W}}$, where $h^*_{\mathbf{W}}$ is the inverse function of $f^*_{\mathbf{I}}$ w.r.t. $\mathbf{W}$, and a a proxy ASCM $\widehat{\mathcal{M}}$ over $\{\mathbf{V}, \mathbf{I}\}$. if $P^{\widehat{\mathcal{M}}}(\mathbf{V}, \mathbf{I}) = P^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I})$, then*

$$h^*_{\mathbf{W}}(P^{\widehat{\mathcal{M}}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})) = P^{\widehat{\mathcal{M}}}(\mathbf{w}, \mathbf{w}'_{\mathbf{x}'}), \quad (13)$$

*where $\mathbf{w} = h_{\mathbf{W}}(\mathbf{i})$, and $\mathbf{w}' = h_{\mathbf{W}}(\mathbf{i}')$.* ■

This result suggests that if $\widehat{\mathcal{M}}$ agrees on the observational distribution of $\mathcal{M}^*$ and the care set $\mathbf{W}$ is a subset of observed generative factors, the feature counterfactual query is equivalent to a counterfactual query $P^{\widehat{\mathcal{M}}}(\mathbf{w}, \mathbf{w}'_{\mathbf{x}'})$ over $\mathbf{W}$ induced by $\widehat{\mathcal{M}}_0$ at the generative level. We normalize $P^{\widehat{\mathcal{M}}}(\mathbf{w}, \mathbf{w}'_{\mathbf{x}'})$ as $P^{\widehat{\mathcal{M}}}(\mathbf{w}'_{\mathbf{x}'} \mid \mathbf{w})$ following:

$$P^{\widehat{\mathcal{M}}}(\mathbf{w}'_{\mathbf{x}'} \mid \mathbf{w}) = P^{\widehat{\mathcal{M}}}(\mathbf{w}, \mathbf{w}'_{\mathbf{x}'}) / P^{\widehat{\mathcal{M}}}(\mathbf{w}) \quad (14)$$

We will focus on the *conditional feature counterfactual query* $P^{\widehat{\mathcal{M}}}(\mathbf{w}'_{\mathbf{x}'} \mid \mathbf{w})$ when the proxy model satisfies $P^{\widehat{\mathcal{M}}}(\mathbf{V}, \mathbf{I}) = P^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I})$, which implies $P^{\widehat{\mathcal{M}}}(\mathbf{w}) = P^{\mathcal{M}^*}(\mathbf{w})$. The following example illustrates this point.

**Example 4.3.** Consider the counterfactual image query $P(\mathbf{i}, \mathbf{i}_{Y=0})$, where $\mathbf{i}$ is a young male without gray hair ($F = 0, Y = 1, H = 0$) and $\mathbf{i}'$ describes an old male with gray hair ($F = 0, Y = 0, H = 1$). Suppose the care set $\mathbf{W}$ contains the feature gender ($F$) and age ($Y$) as in Example 2.2. Lem. 4.2 suggests the feature counterfactual query is

$$P^{\widehat{\mathcal{M}}}(F_{Y=0} = 0, F = 0, Y = 1) \quad (15)$$

whenever $\widehat{\mathcal{M}}$ is compatible with $\mathcal{M}^*$ w.r.t. the observational distribution. The normalized conditional feature counterfactual query is

$$P^{\widehat{\mathcal{M}}}(F_{Y=0} = 0 \mid F = 0, Y = 1), \quad (16)$$

which illustrates the probability that the gender was still male had a young male gotten older induced by the proxy model. ■

(2) **Optimal Bounds.** A complementary relaxation arises from the observation that even when a query is not point identifiable, it is still possible to compute informative bounds over the target distribution from a combination of the observational data and the causal diagram (Manski, 1990; Balke & Pearl, 1994; Zhang et al., 2022). These bounds serve as a natural measure of distance, or tolerance, between what is empirically obtainable from the data and the true, yet unobserved, counterfactual distribution. This occurs because numerous ASCMs, compatible with the observed data, can generate counterfactual distributions encompassing the bound. Any value within the optimal bound $[l, r]$ (Def. 1.3) falls within the range of some possible ground truth, contingent on the given assumptions. As assumptions are strengthened, the bounds naturally narrow. Based on the above discussion, we formally define a class of counterfactual consistent estimators of the target $P(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$.

Based on the above discussion, we formally define the Ctf-consistent estimation of an image counterfactual query.

**Definition 4.4** (Ctf-Consistent Estimator w.r.t. feature set $\mathbf{W}$)**.** Consider a feature set $\mathbf{W} \subseteq \mathbf{V}$ and its mapping function $\phi = h^*_{\mathbf{W}}$, where $h^*_{\mathbf{W}}$ is the inverse function of $f^*_{\mathbf{I}}$ regarding $\mathbf{W}$. $P^{\widehat{\mathcal{M}}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$ is said to be a *Ctf-consistent estimator* of $P^{\mathcal{M}^*}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$ w.r.t. $\mathbf{W}$ if

(1) the observational distributions induced by $\widehat{\mathcal{M}}$ and $\mathcal{M}^*$ are the same, namely, $P^{\widehat{\mathcal{M}}}(\mathbf{V}, \mathbf{I}) = P^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I})$ and

(2) the feature counterfactual query $\phi(P^{\widehat{\mathcal{M}}}(\mathbf{w}, \mathbf{w}'_{\mathbf{x}'}))$ is within the optimal bound of $P(\mathbf{w}, \mathbf{w}'_{\mathbf{x}'})$ derived by $P(\mathbf{V})$ and $\mathcal{G}$, where $\mathbf{w} = h^*_{\mathbf{W}}(\mathbf{i})$ and $\mathbf{w}' = h^*_{\mathbf{W}}(\mathbf{i}')$;

The proxy quantity $P^{\widehat{\mathcal{M}}}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ is said to be a Ctf-consistent estimator of the true $P^{\mathcal{M}^*}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ w.r.t. $\mathbf{W}$ if $P^{\widehat{\mathcal{M}}}(\mathbf{i}, \mathbf{i}_{\mathbf{x}'})$ is Ctf-consistent for every $\mathbf{i}, \mathbf{i}' \in \mathcal{X}_{\mathbf{I}}$. ■

Notice that the feature counterfactual query $\phi(P^{\widehat{\mathcal{M}}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'}))$ is equivalent to $P^{\widehat{\mathcal{M}}}(\mathbf{w}, \mathbf{w}'_{\mathbf{x}'})$ here according to Lem. 4.2. Def. 4.4 states that if (1) the observational distribution

induced by the proxy model is the same as the true model, and (2) the feature counterfactual query induced by the proxy model is within the optimal bound of $P(\mathbf{w}, \mathbf{w}'_{\mathbf{x}'})$, then the corresponding image counterfactual query can be regarded as a Ctf-consistent estimation of the true image counterfactual query. Def. 4.4 does not require that the proxy model $\widehat{\mathcal{M}}$ induces the same counterfactual image distribution $P(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ but expect $\widehat{\mathcal{M}}$ to be Ctf-consistent with $\mathcal{M}^*$ regarding the care set $\mathbf{W}$ while ignoring other observed generative factors $\mathbf{V}\backslash\mathbf{W}$ and $\tilde{\mathbf{U}}$. Specifically, the feature counterfactual distribution $P^{\widehat{\mathcal{M}}}(\mathbf{w}, \mathbf{w}'_{\mathbf{x}'})$ should be within the optimal bound but no restriction is imposed over the features for $\mathbf{V}\backslash\mathbf{W}$ and $\tilde{\mathbf{U}}$. The next example illustrates this idea.

**Example 4.5.** (Example 4.3 continued). Def. 4.4 suggests the conditional feature counterfactual query $Q = P^{\widehat{\mathcal{M}}}(F_{Y=0} = 0 \mid F = 0, Y = 1)$ induced by the proxy model $\widehat{\mathcal{M}}$ should be in the optimal bound $[r, l]$, where

$$r = l = P^{\widehat{\mathcal{M}}}(F = 0 \mid F = 0, Y = 1) = 1 \tag{17}$$

since the intervention $do(Y = 0)$ has no effect on $F$ in the causal diagram (Figure 4).This implies that the gender must remain the same after the editing. In the meantime, it does not matter whether the hair is gray ($\mathbf{V}\backslash\mathbf{W}$) or not and whether the person is smiling ($\tilde{\mathbf{U}}$) since these features are not in the care set.

Now suppose the user cares about gender, age, and hair color, namely, $\mathbf{W} = \{F, Y, H\}$ (instead of $\{F, Y\}$). Based on Def. 4.4 and Lemma 4.2, the corresponding conditional feature counterfactual query is

$$Q = P(F_{Y=0} = 0, H_{Y=0} = 1 \mid F = 0, Y = 1, H = 0), \tag{18}$$

and $Q$ illustrates the probability that the individual is still a male and has gray hair after getting older. This optimal bound analytically can be derived as (see (Pearl, 2009, Thm. 9.2.12)):

$$\begin{aligned} l &= \max\{0, 1 - \frac{P(H = 0 \mid F = 0, Y = 0)}{P(H = 0 \mid F = 0, Y = 1)}\} = 0.25 \\ r &= \min\{1, \frac{P(H = 1 \mid F = 0, Y = 0)}{P(H = 0 \mid F = 0, Y = 1)}\} = 0.5 \end{aligned} \tag{19}$$

Any $P^{\widehat{\mathcal{M}}}(\mathbf{i}, \mathbf{i}_{Y=0})$ induced by $\widehat{\mathcal{M}}$ such that $P^{\widehat{\mathcal{M}}}(\mathbf{V}, \mathbf{I}) = P^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I})$ and $Q^{\widehat{\mathcal{M}}} \in [0.25, 0.5]$ is a Ctf-consistent estimator of $P^{\mathcal{M}^*}(\mathbf{i}, \mathbf{i}_{Y=0})$. Even if $Q^{\widehat{\mathcal{M}}}$ is not equal to the true feature counterfactual query $Q^{\mathcal{M}^*} = 0.4$, the error is acceptable compared to the non-causal method currently used in practice. One may change $Y$ from 1 to 0 and keep the other features as close as possible. With such methods, the counterfactual image will never have gray hair, thus the estimation $Q = P(F_{Y=0} = 0, H_{Y=0} = 1 \mid F = 0, Y = 1, H = 0) = 0$. The causal effect of the intervention $Y = 0$ on $H$ is not reflected. ■

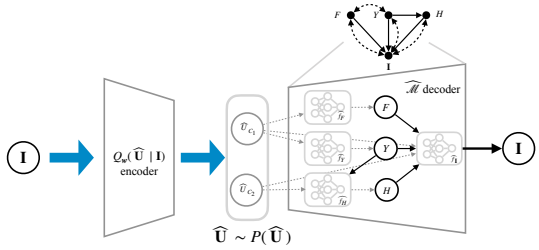



Figure 5: The ANCM network structure for Example 2.2.

From now on, our goal is to obtain a Ctf-consistent estimator of the non-identifiable target $P(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ w.r.t. the care set $\mathbf{W}$.

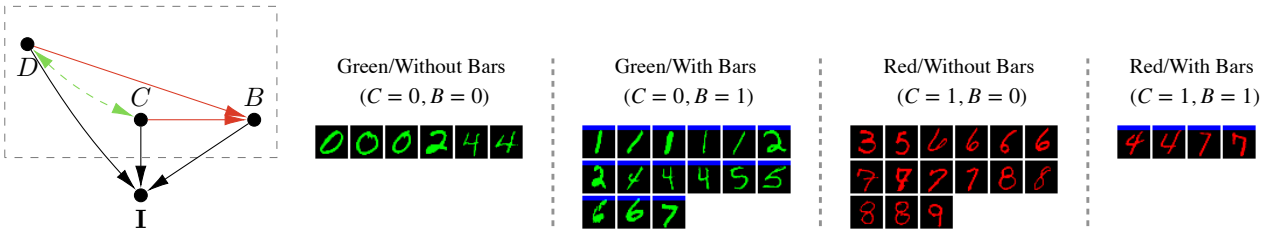
**Theorem 4.6** (Counterfactually Consistent Estimation)**.** *$P^{\widehat{\mathcal{M}}}(\mathbf{I}, \mathbf{I}'_{\mathbf{x}'})$ is a Ctf-consistent estimator with respect to $\mathbf{W} \subseteq \mathbf{V}$ of $P^{\mathcal{M}^*}(\mathbf{I}, \mathbf{I}'_{\mathbf{x}'})$ if $\widehat{\mathcal{M}} \in \Omega_{\mathbf{I}}(\mathcal{G})$ and $P^{\widehat{\mathcal{M}}}(\mathbf{V}, \mathbf{I}) = P(\mathbf{V}, \mathbf{I})$.* ■

The above result says that any proxy ASCM that is compatible with the diagram $\mathcal{G}$ and $P(\mathbf{V}, \mathbf{I})$ guarantees the estimation of the target distribution being Ctf-consistent with the true one. Specifically, in order to construct Ctf-consistent estimators, apart from fitting the generator $\widehat{\mathcal{M}}$ with the given observation $P(\mathbf{V}, \mathbf{I})$, it is sufficient to enforce the graphical constraints into $\widehat{\mathcal{M}}$.

### 4.1. Estimating and Sampling with NCMs

We learned in the previous section that one could generate Ctf-consistent samples by fitting observational distributions to an SCM $\widehat{\mathcal{M}}$ that is compatible with the given diagram (Thm. 4.6). In this section, we develop a practical method for training $\mathcal{G}$-Constrained causal deep generative models ($\mathcal{G}$-NCMs) with two primary objectives: (a) to fit the observational distribution $P(\mathbf{V}, \mathbf{I})$; (b) to sample images ($\mathbf{i}$) and their counterfactual counterparts ($\mathbf{i}'$) from them.

Towards realizing these goals, we first acknowledge that $P^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I})$ is typically not directly accessible in most settings, but rather its empirical counterpart $\widehat{P}^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I}) = \{\mathbf{v}_k, \mathbf{i}_k\}_{k=1}^n$ derived from finite datasets. Subsequently, we will train $\widehat{\mathcal{M}}$ to match this empirical distribution $\widehat{P}^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I})$. Given the substantial difference in the dimensions of variables $\mathbf{V}$ (feature labels) and $\mathbf{I}$ (images), we prefer to fit $P(\mathbf{I})$ and $P(\mathbf{V} \mid \mathbf{I})$ separately. Initially, $P(\mathbf{I})$ will be learned by minimizing the data negative log-likelihood through VAEs (Kingma & Welling, 2013). In this context, the proxy $\mathcal{G}$-NCM $\widehat{\mathcal{M}}$ serves as the decoder to approximate $P(\mathbf{I} \mid \widehat{\mathbf{U}})$ with the prior $P(\widehat{\mathbf{U}})$. Furthermore, a separate deep neural network $Q_{\boldsymbol{\omega}}(\widehat{\mathbf{U}} \mid \mathbf{I})$ is utilized to approximate the posterior $P(\widehat{\mathbf{U}} \mid \mathbf{I})$, acting as the encoder,




(a) (b)

Figure 6: The causal diagram $\mathcal{G}^{\mathbf{B}}$ and samples for "Backdoor" setting. There are more red larger digits and green smaller digits; larger digits are less likely to have a bar on top; red digits are less likely to have a bar on top.

with $\boldsymbol{\omega}$ denoting the network's parameters. The network structure corresponding to Example 2.2 is illustrated in Figure 5. The optimization objective, $L_1$, is then defined as the following evidence lower bound (ELBO) to minimize the data negative log-likelihood:

$$\begin{aligned} & L_1(\boldsymbol{\theta}, \boldsymbol{\omega}, \widehat{P}^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I})) \\ &= \mathbb{E}_{\widehat{P}^{\mathcal{M}^*}(\mathbf{V},\mathbf{I})}\left[\mathbb{E}_{Q_{\boldsymbol{\omega}}(\widehat{\mathbf{U}}|\mathbf{I})}[\log P^{\widehat{\mathcal{M}}}(\mathbf{i} \mid \widehat{\mathbf{u}})]\right] \\ &\quad - D_{KL}[q_{\boldsymbol{\omega}}(\widehat{\mathbf{U}} \mid \mathbf{I}) || P(\widehat{\mathbf{U}})], \end{aligned} \tag{20}$$

where $\boldsymbol{\theta}$ are parameters of $\widehat{\mathcal{M}}$ (see Def. 2.1) and $D_{KL}[\cdot\|\cdot]$ denotes KL divergence. To match $P(\mathbf{V} \mid \mathbf{I})$, we minimize the cross-entropy loss $L_2$ of the true labels of an image sample and its predicted labels, which can be inferred through $Q_{\boldsymbol{\omega}}(\widehat{\mathbf{U}} \mid \mathbf{I})$ and $\mathcal{M}$ like (Locatello et al., 2020b; Shen et al., 2022). Namely,

$$\begin{aligned} & L_2(\boldsymbol{\theta}, \boldsymbol{\omega}, \widehat{P}^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I})) \\ &= \mathbb{E}_{\widehat{P}^{\mathcal{M}^*}(\mathbf{V},\mathbf{I})}\left[D_{CE}(\mathbf{V}^{\widehat{\mathcal{M}}}(r(\mathbf{i})), \mathbf{v})\right] \end{aligned} \tag{21}$$

where $r(\mathbf{i})$ corresponds to the mean (vector) of $Q_{\boldsymbol{\omega}}(\widehat{\mathbf{U}} \mid \mathbf{I})$ and $D_{CE}(\cdot)$ is the cross-entropy loss. Formally, the objective for training an NCM $\widehat{\mathcal{M}}$ can be written as

$$L = L_1(\boldsymbol{\theta}, \boldsymbol{\omega}, \widehat{P}^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I})) + \lambda L_2(\boldsymbol{\theta}, \boldsymbol{\omega}, \widehat{P}^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I})) \tag{22}$$

where $\lambda$ is a parameter trying to balance the likelihood $P(\mathbf{V})$ and $P(\mathbf{I}|\mathbf{V})$. Specifically, a larger $\lambda$ prioritizes the fit of $P(\mathbf{V} \mid \mathbf{I})$ and a smaller one prioritizes the fit of $P(\mathbf{I})$ during the training stage. Alg. 1 implements more specifically the training procedure of an NCM. To illustrate, in line 1, the decoder is constructed based on the given causal diagram through Def. 1.4. In lines 2, all training parameters are initialized. And then in lines 3 to 6, the encoder and decoder are trained iteratively based on Equation (22). We refer to this approach as *ANCM*. More details about network architecture and hyperparameters used throughout this work can be found in Appendix B.

After training the ANCM, we first sample $\widehat{\mathbf{u}}$ from $P(\widehat{\mathbf{U}})$ to generate samples of the target $P(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$. The initial image sample $\widehat{\mathbf{i}}$ could be derived from $\mathbf{I}^{\widehat{\mathcal{M}}_{\mathbf{x}'}}(\widehat{\mathbf{u}})$, where $\mathbf{I}^{\widehat{\mathcal{M}}_{\mathbf{x}'}}$ is the network mapping from $\widehat{\mathbf{u}}$ to $\mathbf{i}$ in the decoder $\widehat{\mathcal{M}}$. To edit the concept $\mathbf{X} = \mathbf{x}'$, the counterfactual image sample $\widehat{\mathbf{i}}_{\mathbf{x}'}$ could be derived through $\mathbf{I}^{\widehat{\mathcal{M}}_{\mathbf{x}'}}(\widehat{\mathbf{u}})$, where $\mathbf{I}^{\widehat{\mathcal{M}}_{\mathbf{x}'}}$ is the network but evaluated through submodel $\widehat{\mathcal{M}}_{\mathbf{x}'}$ of the trained NCM.

**Algorithm 1** ANCM

**Input:** Data $\{\widehat{P}^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I}) = \{\mathbf{v}_k, \mathbf{l}_k\}_{k=1}^n\}$, causal diagram $\mathcal{G}$, temperature $\lambda$, learning rate $\eta$, training epochs $T$.

1: $\widehat{\mathcal{M}} \leftarrow \text{NCM}(\mathbf{V}, \mathcal{G})$ {from Def. 1.4}
2: Initialize parameters $\boldsymbol{\theta}$ for $\widehat{\mathcal{M}}$ and $\boldsymbol{\omega}$ for the inference network $Q_{\boldsymbol{\omega}}(\widehat{\mathbf{U}} \mid \mathbf{I})$
3: **for** $t \leftarrow 1$ to $T$ **do**
4: $\quad L \leftarrow L_1(\boldsymbol{\theta}, \boldsymbol{\omega}, \widehat{P}^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I})) + \lambda L_2(\boldsymbol{\theta}, \boldsymbol{\omega}, \widehat{P}^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I}))$
5: $\quad \boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla L$
6: $\quad \boldsymbol{\omega} \leftarrow \boldsymbol{\omega} - \eta \nabla L$
7: **end for**

## 5. Experiments

In this section, we conduct an empirical evaluation of the methods proposed in the paper, beginning with a modified Colored MNIST dataset (based on Sec. 5.1) and then moving on to CelebA-HQ dataset (Karras et al., 2018) (which describes peoples' faces) (Sec. 5.2). Further details of the model architectures are provided in Appendix B.

### 5.1. Colored MNIST with Bars

We first conduct experiments on the modified handwritten MNIST dataset (Deng, 2012), featuring colored digits and a horizontal blue bar in images. [2] The observed generative factors include $\{D, C, B\}$, where $D$ denotes the digits from 0 to 9; $C$ indicates the digit color (green for $C = 0$; red for $C = 1$); $B$ determines whether the top of the image features a blue bar ($B = 1$) or not ($B = 0$). We explore two settings, named "Backdoor" (Sec. 5.1.1) and "Frontdoor"

[2] A bar in an image refers to complete rows of blue pixels.

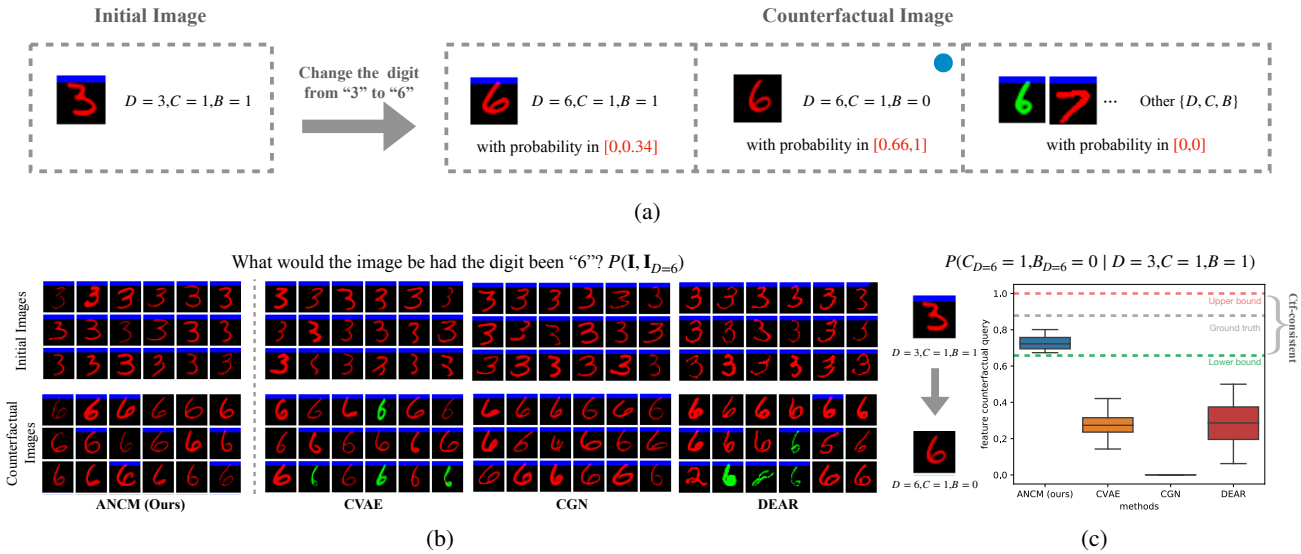



Figure 7: (a) The optimal bound of feature counterfactual queries when editing a red "3" with a bar to "6". (b) The counterfactual image generation results when editing a red "3" with a bar to "6" in the backdoor model. (c) The selected (blue circle) feature counterfactual query estimated by ANCMs and baselines.

(Sec. 5.1.2), each defined by unique causal relationships among the generative factors. This is an attractive dataset since we have full control over the generative process (SCM) and the ground truth is well-defined. For each task, we illustrate the concept of counterfactual editing and demonstrate that our method is capable of achieving great success in counterfactual editing tasks when compared with baselines.

### 5.1.1. MNIST Backdoor Model

In the Backdoor setting, the digit ($B$) and the color ($C$) are confounded with a positive correlation, but they do not directly affect each other. There are more red/larger ($\geq 5$) digits and green/smaller ($< 5$) digits in the dataset. The digit ($D$) has a negative effect on the existence of the bar ($B$). Larger digits are less likely to have a bar on the top. The color ($C$) also has a negative effect on the existence of the bar ($B$). Red digits are less likely to have a bar on top. The true and unknown ASCM $\mathcal{M}^{\mathbf{B}}$ is given by:

$$\begin{cases} D \leftarrow U_D \\ C \leftarrow \text{Bernoulli}(0.95 - 0.1U_D) \\ B \leftarrow (\mathbf{1}\left[U_D \geq 5\right] \oplus U_1) \vee (C \oplus U_2) \wedge U_3 \\ \mathbf{I} \leftarrow f_{\mathbf{I}}^{\mathbf{B}}(D, C, B, \mathbf{U_I}), \end{cases} \tag{23}$$

where the exogenous variables' distributions are:

$$\begin{aligned} U_D &\sim \text{Uniform}[0, 9] \\ U_1 &\sim \text{Bernoulli}(0.8) \\ U_2 &\sim \text{Bernoulli}(0.9) \\ U_3 &\sim \text{Bernoulli}(0.75) \end{aligned} \tag{24}$$

The mechanism $f^{\mathbf{B}}$ maps the observed generative factors $\{D, C, B\}$ and unobserved generative factors (such as the position and thickness of the digit) produced by $\mathbf{U_I}$ to the image $\mathbf{I}$. Figure 6a shows the causal diagram $\mathcal{G}^{\mathbf{B}}$ induced by $\mathcal{M}^{\mathbf{B}}$. The green edge indicates a positive effect and the red one represents a negative effect. We randomly sample 40 images from the collected samples in this setting and show them in Figure 6b.

**Task 1: Counterfactually Editing the Digits**

The first case we consider is to counterfactually edit the digit $D$. We let the cared features be the digit, color, and whether the image has a bar, namely, $\mathbf{W} = \{B, C, D\}$. This implies that we do not care about how other generative factors (i.e., position, thickness) change in the counterfactual world. For counterfactual editing, changing $D$ should not affect $C$ while it might possibly change $B$, since $D$ is confounded with $C$ but has a direct effect on $B$ (Figure 6a).

For instance, suppose we are editing a red "3" with a bar (an image with $\{D = 3, C = 1, B = 1\}$) and wonder what would happen had the digit "3" been a "6". In this case, the optimal bounds of conditional feature counterfactual distribution $P(D_{D=6}, C_{D=6}, B_{D=6} \mid D = 3, C = 1, B = 1)$ derived from the observational distribution $P(D, C, B, \mathbf{I})$ and the causal diagram $\mathcal{G}^{\mathbf{B}}$ are shown in Figure 7a. Specifically, the probability that the counterfactual image has features $\{D = 6, C = 1, B = 1\}$ is

$$P(C_{D=6} = 1, B_{D=6} = 1 \mid D = 6, C = 1, B = 1) \in [0, 0.34]. \tag{25}$$

Further, the probability that the counterfactual image has

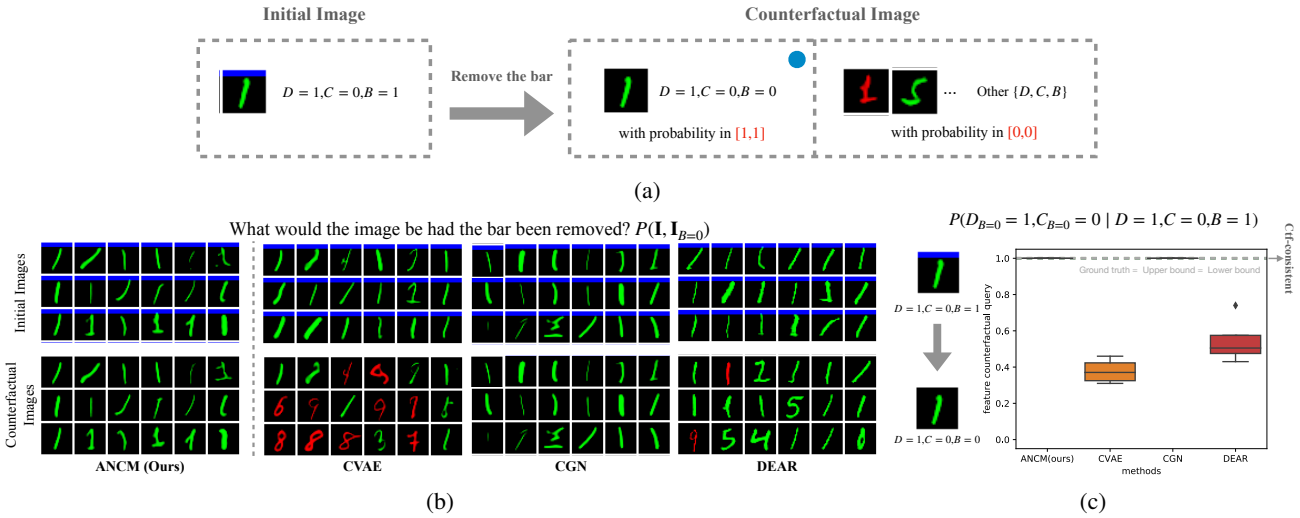



Figure 8: (a) The optimal bound of feature counterfactual queries when removing the bar for a green "1". (b) The counterfactual image generation results when removing the bar for a green "1" in the backdoor model. (c) The selected (blue circle) feature counterfactual query estimated by ANCMs and baselines.

features $\{D = 6, C = 1, B = 0\}$ is

$$P(C_{D=6} = 1, B_{D=6} = 0 \mid D = 6, C = 1, B = 1) \in [0.66, 1]. \tag{26}$$

The probability that the counterfactual image has other features (such as green "6", red "7") is zero. In other words, the counterfactual image by causal generative models can only be a red "6" with a bar ($\{D = 6, C = 1, B = 0\}$) or a red "6" without a bar ($\{D = 6, C = 1, B = 0\}$). Furthermore, the probability of the latter scenario where the bar disappears should be no less than $0.66$ due to the effect from $D$ to $B$. To achieve Ctf-consistency, we expect the generation process to follow these theoretical bounds.

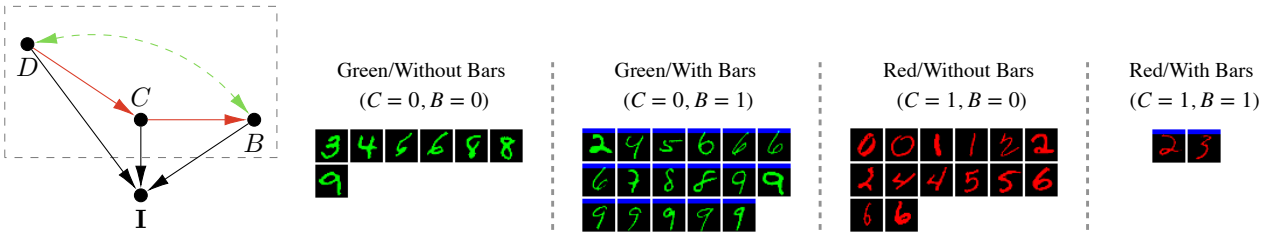
The full counterfactual image editing results are shown in Figure 7b. After changing the digit, CVAE is likely to change the color $C$ as it uses the correlation between $D$ and $C$, while $D$ and $C$ are spuriously correlated, as discussed earlier. Also, the CVAE fails to capture the causal effect from $D$ to $B$ since the bar hardly disappears after the intervention $do(D = 6)$. CGN preserves the values of both $C$ and $B$ after the intervention $D$ since it learns independent mechanisms from the generative factors to images and all generative factors are independent given the label (see more details in the Appendix B.1). The results suggest that the causal effect from $D$ to $B$ is not reflected in this estimation. DEAR follows a Markovian graph and ignores bi-direct edges in the causal diagram. Thus, DEAR fails to fit the observational distribution according to Prop. C.3 and cannot preserve the color after the intervention. ANCM preserves the original colors in counterfactual images and is likely to remove the bar, reflecting the bound value discussed above.

These results provide a qualitative suggestion that ANCM generates more realistic images that preserve the causal features found in the true generation model. Still, we would like to quantitatively understand these results, so we re-run each method 4 times and calculate the empirical probability that counterfactual images describe a red "6" without a bar after editing a red "3" with a bar to digit "6", namely $P(C_{D=6} = 1, B_{D=6} = 0 \mid D = 6, C = 1, B = 1)$. The corresponding results are shown in Figure 7c. To illustrate, the gray dashed line denotes the value given by the above ASCM $\mathcal{M}^{\mathbf{B}}$, which is unknown by any of the methods. The red line represents the upper bound of the feature counterfactual query and the green line represents the lower bound. We can see that queries generated by all baseline methods are not within the optimal bound. On the other hand, the queries generated by ANCMs are all within the optimal bound, thus they are also not far from the unknown value and are the best that can be obtained without further assumptions (over the ASCM). Both the visualization, numerical results, and theoretical results state the ANCMs are able to capture the causal effects among $\{D, C, B\}$ and produce Ctf-consistent estimators while the baselines do not.

**Task 2: Counterfactually Editing the Bars**

We now consider editing the bar $B$. We also let the care set to be $\mathbf{W} = \{B, C, D\}$. Based on the causal diagram $\mathcal{G}^{\mathbf{B}}$, we can see that changing $B$ should affect neither $D$ nor $C$. This is because $B$ is a descendant of $D$ and $C$ and not the other way around.

For concreteness, suppose we are editing an image describing a green "1" with a bar (an image with $\{D = 1, C = 0, B = 1\}$) and wonder what would happen had the bar been removed. In this case, the optimal bounds of conditional feature counterfactual distribution $P(D_{B=0}, C_{B=0} \mid D =$




(a) (b)

Figure 9: The causal diagram $\mathcal{G}^{\mathbf{F}}$ and samples for "Frontdoor" setting. Bigger digits are likely to be green; red digits are less likely to have a bar on top; there are bigger digits with bars and smaller digits without bars.

$1, C = 0, B = 1)$ are shown in Figure 8a. Specifically, the optimal bound of the probability that the counterfactual image has features $\{D = 1, C = 0, B = 0\}$ is

$$P(D_{B=0} = 1, C_{B=0} = 0 \mid D = 1, C = 0, B = 0) \in [1, 1] \tag{27}$$

Further, the probability that the counterfactual image has other features (such as red "1" and green "5") is zero. In other words, the counterfactual images must be an image describing a green "1" without a bar.

The counterfactual image editing results of ANCM and baselines are shown in Figure 8b. All methods remove bars in counterfactual images. Since $B$ is spuriously correlated with $D$ and $C$, CVAE and DEAR are also likely to change the color to red and to the larger digit when changing $B$ to zero. CGN successfully learns the independent mechanisms and changes the bars without affecting the original color and digit. We can see this branch of work (changing the intervened features but preserving others) does work for some causal relationships, but there are situations where these methods cannot work, for instance, Task 1 above (see more discussion about this in Appendix C.2). Our method ANCM also preserves these two features in counterfactual images.

These results provide a qualitative suggestion that ANCM and CGN is able to generate realistic results images that preserve the causal features found in the original model in this setting. In order to quantitatively understand these results, we re-run each method 4 times and calculate the empirical probability that counterfactual images describe a green "1" without bar after $do(B = 0)$, namely $P(D_{B=0} = 1, C_{B=0} = 0 \mid D = 1, C = 0, B = 0)$ (Figure 8c). The upper bound, lower bound, and the true value collapse to one line, which implies $P(D_{B=0} = 1, C_{B=0} = 0 \mid D = 1, C = 0, B = 0) = 1$. We can see that queries generated by CVAEs and DEAR are much smaller than the true value 1 while the queries generated by CGNs and ANCMs coincide with the true value. Both the visualization and the numerical results suggest that CGNs and ANCMs provide Ctf-consistent estimators.

#### 5.1.2. MNIST FRONTDOOR MODEL

In the Frontdoor setting, the digit ($D$) has a negative effect on the color ($C$). Larger ($\geq 5$) digits are more likely to be green. The color $C$ has a negative effect on the existence of the bar ($B$). Red digits are less likely to have a bar on top. The digit ($D$) is confounded with the existence of the bar ($B$) with a positive correlation, but do not directly affect each other. There are larger ($\geq 5$) digits with bars and smaller ($< 5$) digits without bars in the dataset. The true and unknown ASCM $\mathcal{M}^{\mathbf{F}}$ is given by:

$$\begin{cases} D \leftarrow U_D \\ C \leftarrow \text{Bernoulli}(0.05 + 0.1U_D) \\ B \leftarrow (\mathbf{1}\,[D < 5] \oplus U_2) \vee (C \oplus U_1) \wedge U_3 \\ \mathbf{I} \leftarrow f^{\mathbf{F}}_{\mathbf{I}}(D, C, B, \mathbf{U_I}), \end{cases} \tag{28}$$

where the exogenous variable distributions are:

$$\begin{aligned} U_D &\sim \text{Uniform}[0, 9] \\ U_1 &\sim \text{Bernoulli}(0.8) \\ U_2 &\sim \text{Bernoulli}(0.9) \\ U_3 &\sim \text{Bernoulli}(0.7) \end{aligned} \tag{29}$$

The mechanism $f^{\mathbf{F}}$ maps the observed generative factors $\{D, C, B\}$ and unobserved generative factors (such as the position and thickness of digits) produced by $\mathbf{U_I}$ to the image $\mathbf{I}$. Figure 9a shows the causal diagram $\mathcal{G}^{\mathbf{F}}$ induced by $\mathcal{M}^{\mathbf{F}}$. The green edge indicates a positive effect and the red one represents a negative effect. We randomly sample 40 images from the collected samples in this setting and show them in Figure 9b.

**Task 3: Counterfactually Editing the Digits**

We first consider editing the digit $D$. We let the care set be $\mathbf{W} = \{B, C, D\}$ similar to previous cases. Based on the causal diagram $\mathcal{G}^{\mathbf{F}}$, changing $D$ should directly affect $C$ and is possible to change indirectly through $B$, since $D$ is the direct parent of $C$ but not a parent of $B$ (even though it's an ancestor). For instance, suppose we are editing a green "7" with a bar (an image with $\{D = 7, C = 0, B = 1\}$) and

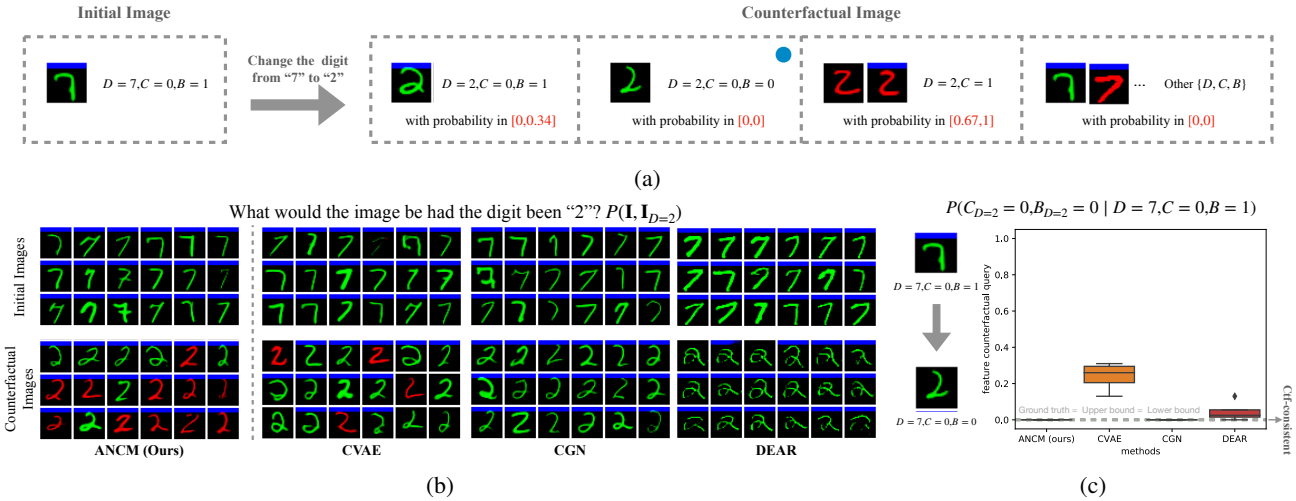



Figure 10: (a) The optimal bound of feature counterfactual queries when editing a green "7" with a bar to "2". (b) The counterfactual image generation results when editing a green "7" with a bar to "2" in the frontdoor model. (c) The selected (blue circle) feature counterfactual query estimated by CVAEs and ANCMs.

wonder what would happen had the "7" been "2". In this case, the optimal bounds of conditional feature counterfactual distribution $P(C_{D=2}, B_{D=2} \mid D = 7, C = 0, B = 1)$ derived from the observational distribution $P(D, C, B, \mathbf{I})$ and the causal diagram $\mathcal{G}^{\mathbf{F}}$ are shown in Figure 10a. Specifically, the probability that the counterfactual image has features $\{D = 2, C = 0, B = 1\}$ is

$$P(C_{D=2} = 0, B_{D=2} = 1 \mid D = 7, C = 0, B = 1) \in [0, 0.33]. \tag{30}$$

And the probability that the counterfactual image has features $\{D = 2, C = 0, B = 0\}$ is

$$P(C_{D=2} = 0, B_{D=2} = 0 \mid D = 7, C = 0, B = 1) \in [0, 0]. \tag{31}$$

Further, the probability that the counterfactual image has features $\{D = 2, C = 1\}$ is

$$P(D_{D=2} = 2, C_{D=2} = 1 \mid D = 6, C = 1, B = 1) \in [0.67, 1]. \tag{32}$$

And the probability that the counterfactual image has other features (such as green "7" and red "7") is zero. In other words, the counterfactual image by causal generative models can only be a green "2" with a bar ($\{D = 6, C = 1, B = 0\}$) or a red "2" ($\{D = 6, C = 1, B = 0\}, \{D = 6, C = 1, B = 1\}$). Since $D$ can only indirectly affect $B$ through $C$, changing the digit will not influence the bar if the color remains the same. Thus, we expect that the counterfactual image generated by estimation methods would not be a green "2" without a bar.

The counterfactual image editing results of ANCM and baselines are shown in Figure 10b. ANCM and CVAE methods generate red digits with bars and red digits without bars, which implies they capture the effect from $D$ to $C$ and $B$. However, CVAE generates counterfactual images describing green "2" without bars since $D$ and $B$ are correlated. ANCM does not change the existence of the bar when $C$ remains the same. This implies that ANCM captures the indirect effect from $D$ to $B$ as discussed above. CGN fails to capture the effect from $D$ to $C$ and $B$ and simply keep $C$ and $B$ the same. DEAR also fails to capture the correct causal relationships since the graph encoded in the network is incorrect and the data distribution cannot be fit.

To quantitatively understand these results, we re-run each method 4 times and calculate the empirical probability that counterfactual images describing a green "2" without bars after editing a green "7" with a bar to digit "2", namely $P(C_{D=2} = 0, B_{D=2} = 0 \mid D = 7, C = 0, B = 1)$. The results are shown in Figure 10c. The upper bound, lower bound, and the feature counterfactual query generated by the true model collapse to one line, which implies $P(C_{D=2} = 0, B_{D=2} = 0 \mid D = 7, C = 0, B = 1) = 0$. We can see that the query estimated by CVAE and DEAR are greater than 0. The query estimated by CGN is 0 but this is because CGNs fail to capture any (indirect or direct) effect from the digits to bars. In contrast, the queries generated by ANCMs are exactly 0. Both the visualization and the numerical results state that ANCMs capture the full causal invariances among $\{D, C, B\}$ and produce Ctf-consistent estimators while the other methods do not.

**Task 4: Counterfactually Editing the Colors**

We then consider editing the color $\mathbf{C}$ of digits. Similarly, we let $\mathbf{W} = \{B, C, D\}$. Based on the causal diagram $\mathcal{G}^{\mathbf{F}}$, we can see that changing $C$ should not affect $D$ and is possible to change $B$ since $C$ is confounded with $D$

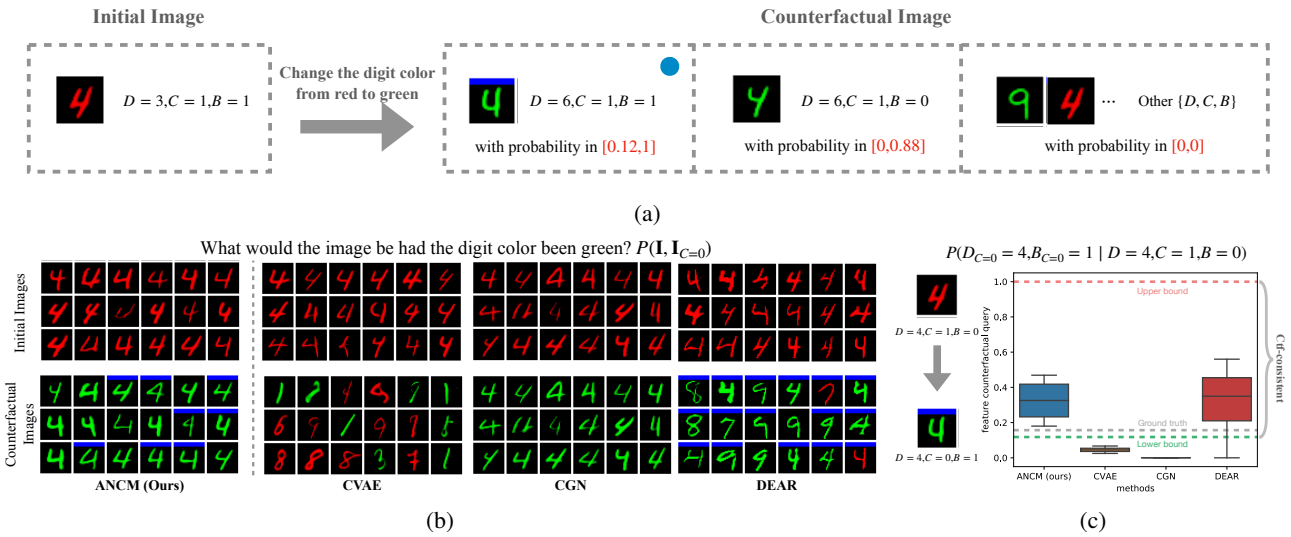



Figure 11: (a) The optimal bound of feature counterfactual queries when editing a red "4" without a bar to green. (b) The counterfactual image generation results when editing a red "4" with a bar to green in the frontdoor model. (c) The selected (blue circle) feature counterfactual query estimated by ANCMs and baselines.

and has a direct effect on $B$ in Figure 9a. For example, suppose we are editing a red "4" without a bar (an image with $\{D = 4, C = 1, B = 0\}$) and wonder what would happen had the color been green. In this case, the optimal bounds of conditional feature counterfactual distribution $P(D_{C=0}, B_{C=0} \mid D = 4, C = 1, B = 0)$ derived from the observational distribution $P(D, C, B, \mathbf{I})$ and the causal diagram $\mathcal{G}^{\mathbf{F}}$ are shown in Figure 11a. Specifically, the probability that the counterfactual image has features $\{D = 4, C = 0, B = 1\}$ is

$$P(D_{C=0} = 4, B_{C=0} = 1 \mid D = 4, C = 1, B = 0) \in [0.12, 1]. \tag{33}$$

Futher the probability that the counterfactual image has features $\{D = 4, C = 1, B = 0\}$ is

$$P(D_{C=0} = 4, B_{C=0} = 0 \mid D = 4, C = 1, B = 0) \in [0, 0.88]. \tag{34}$$

And the probability that the counterfactual image has other features (such as green "9" and red "4") is zero. In other words, the counterfactual image by causal generative models can only be a green "4" and the bar will appear with a probability at least 0.12 after changing the color to green.

The counterfactual image editing results of ANCM and the baselines are shown in Figure 11b. CVAE and DEAR are also likely to change the digit since $D$ is spuriously correlated with each other and DEAR fails to fit the given distribution. CGN fails to capture the direct causal effect from the color to the bar and preserves the digits and the bars as the same. ANCM preserves the original digit in counterfactual images after editing and the change of $B$ shows up.

These results provide a qualitative suggestion that ANCM generates more realistic results images that preserve the causal features found in the original model. To quantitatively understand these results, we re-run each method 4 times and calculate the empirical probability that counterfactual images describe a green "4" with a bar after editing a red "4" without a bar to green, namely $P(D_{C=0} = 4, B_{C=0} = 1 \mid D = 4, C = 1, B = 0)$. The results are shown in Sec. 5.1.2. We can see that ANCMs present Ctf-consistent estimators while queries generated by CVAEs and CGN are nearly 0, which is not within the bound. DEARs roughly present Ctf-consistent estimators for $P(D_{C=0} = 4, B_{C=0} = 1 \mid D = 4, C = 1, B = 0)$, but the former visualization results demonstrate DEAR cannot preserve the digit, which means it cannot provide Ctf-consistent estimators for the whole distribution. Both the visualization and the numerical results state that ANCMs capture the causal invariances among $\{D, C, B\}$ while the baseline methods do not.

### 5.2. Celeba-HQ

In CelebA-HQ experiment, we consider two causal diagrams as shown in Fig. 12. In the first experiment, we consider generative factors $Smile$ $(S)$ and $Open\ Mouth$ $(O)$, and in the second experiment, we consider $Female$ $(F)$, $Young$ $(Y)$ and $Grayhair$ $(H)$. The first target counterfactual queries are "What would the image be had the person opened the mouth?", and the second is "What would the image be had the person been older?". The feature sets are $\mathbf{W} = \{S, O\}$ and $\mathbf{W} = \{F, Y, H\}$ in these two settings, respectively. We also compare ANCM (ours) against the CVAE and DEAR baselines. CGN is not compared here since the variables of CGN are restricted to $Shape, Texture$ and $Background$. Meanwhile, Diffuse-

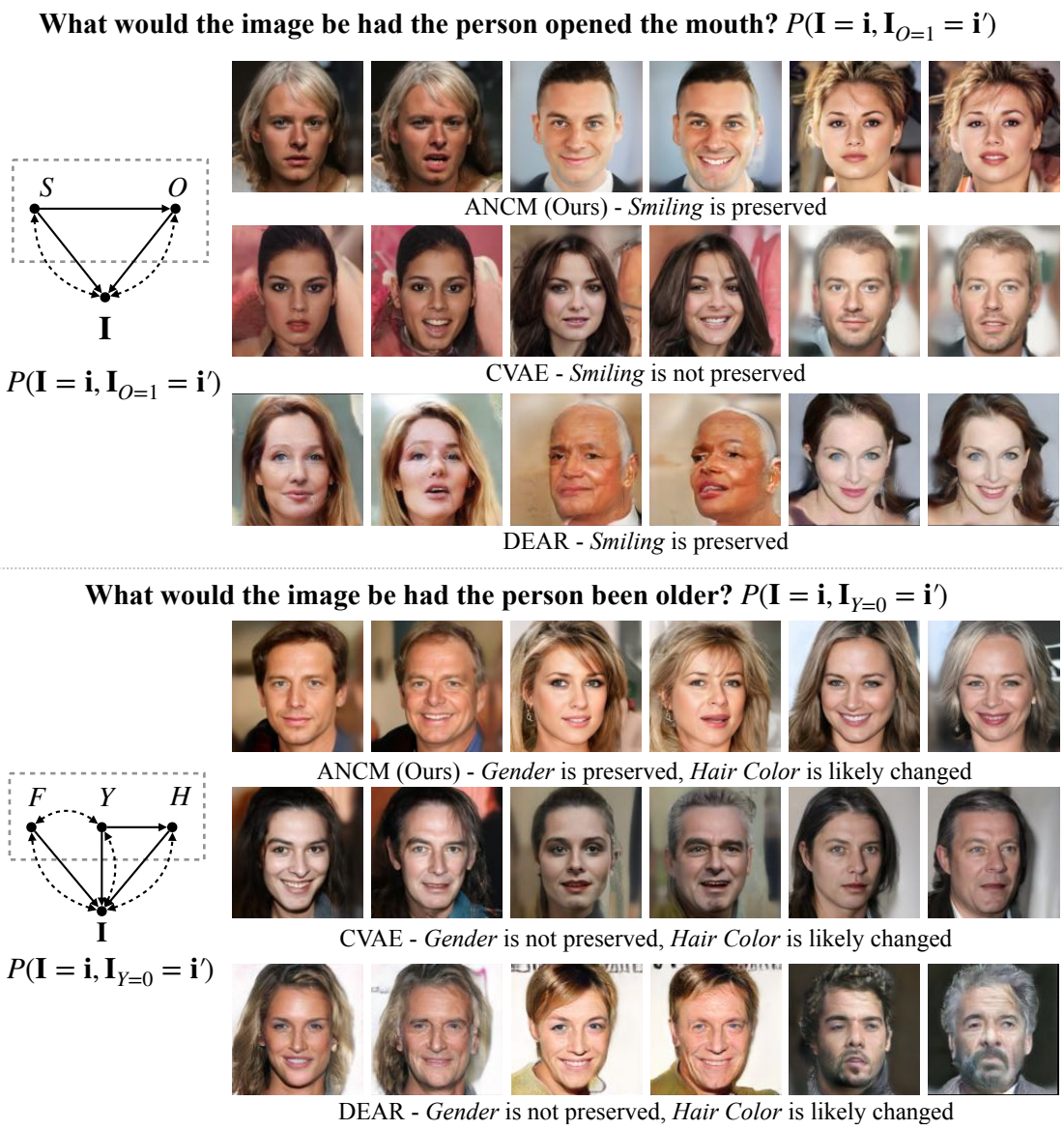



Figure 12: Editing results of the CelebaHQ Experiment.

VAE (Pandey et al., 2022) is leveraged for ANCM and CVAE here to refine samples to high quality since VAEs often produce blurry images that lack high-frequency information (Dosovitskiy & Brox, 2016).

The empirical results are shown in Fig. 12. In the first setting, the feature set $\mathbf{W} = \{S, O\}$ implies the counterfactual query is $P(S, O, S_{O=1})$, namely, "Would the person smile (or not) had the person opened the mouth?". The constraints induced by the ground truth model imply that changing the mouth should not affect smiling since $O$ is the direct child of $S$ and not the other way around. As shown Figure 12, the smiling features are preserved after the editing by ANCM and DEAR. However, CVAE only captures the correlation between these factors, thus the non-smiling person changes to smiling after editing of mouth. On the other hand, ANCM produces higher-quality images compared to DEAR.

The second causal diagram indicates the correlations between gender and age in Example 1.1. The dataset has more face images of young females and old males. More specifically, 71% of the young people are female and 66% of the old people are male. The features set $\mathbf{W} = \{F, Y, H\}$ implies the counterfactual distribution is "What would the gender and the hair color of the person be had the person been older?". The causal constraints suggest that the gender of the person should be preserved and the likelihood of gray hair should increase. ANCM matches these causal relationships while baselines may change the original gender as shown in Figure 12, which is of course undesirable.

## 6. Conclusions

We study the problem of counterfactual image generation and editing through formal causal language. Our goal in this paper is to provide guarantees that when a particular feature is edited within an image, the resulting changes in other generative factors are faithfully reflected in the counterfactual images produced. We formally showed that image counterfactual distributions are not identifiable from a combination of observational data and prior causal knowledge about the generating model represented as a causal diagram. In such non-identifiable cases, we propose new estimators (Ctf-consistent) that come accompanied with guarantees that the generated counterfactual images remain causally consistent with the true image counterfactual distribution for any causal relationship between generative factors. We developed an efficient algorithm to train neural causal models and sample counterfactual images. Finally, we demonstrate our methods are able to generate high-quality counterfactual images for synthetic images.

Building on the machinery developed in this paper, and the understanding gained from it about image counterfactual generation, we identify some future challenges to extend these results in a broader range of practical settings. First, the set of features of interests, referred to as the "care set" in the paper ($\mathbf{W}$), guarantees the preservation of invariances and causal relations across counterfactual conditions. Although we consider settings where the set $\mathbf{W}$ is labeled, the challenge of handling unlabeled data in many practical situations remains significant. Second, another important area for future research is enhancing the efficiency and scalability of the inferences made in this paper.

**Impact statement.** This paper presents work whose goal is to advance the field of machine learning. There are many societal implications of our work and we hope to be beneficial, as elaborated next. Reflecting on the broader literature, we propose the first method capable of providing formal guarantees over counterfactual image generation and editing. The main advancement of our work lies in its emphasis on preserving the causal relationships among features, enabling sound, robust, and more realistic counterfactual generation. This approach differs significantly from the existing literature, which primarily focuses on reflecting the intervened features in the image. The critical distinction centers on what happens with the other features that were not intervened upon, and determining which features are shared or not between factual and counterfactual worlds. Although almost never formally articulated, there are two prevalent approaches to this problem in the prior literature. Some works remain silent regarding the counterfactual status of the non-intervened features. This means that the neural network might leverage the correlation between features found in the factual world, leading to the various spurious results discussed earlier. For instance, instructing a generative AI to change a specific feature of an individual might result in a completely different person with other features, such as a different gender or race, despite they are not being causally related. This occurs because the neural model tends to leverage the correlation between factors found in the observational data, which is oblivious to their causal relationship. Other works attempt to ensure that the non-intervened features are preserved across factual and counterfactual worlds. However, this approach is also inadequate in settings where some of the features exert causal influence on others, and the generative AI should accordingly ascertain these relations. For instance, making a person older should logically lead to changes in hair color (or its amount) in both factual and counterfactual images.

After all, we believe the results stemming from this work have broad implications for the development of the next generation of generative AI. First, we note that the training datasets used for large generative models are almost never balanced (see, for example, (Buolamwini & Gebru, 2018)), which implies spurious correlations across features and the generated images. In practice, this often leads to more frequent, unexpected inaccuracies and biases in these models (e.g., refer to (Plecko & Bareinboim, 2022). ) Understanding and accounting for the causal relationships among generative factors is fundamental for the accuracy and fairness of these models. Second, the lack of proper treatment of the causal invariances required for sound counterfactual reasoning translates into the impossibility of providing any sort of guarantees over what these models generate as output and their plausibility, a certainly undesirable state of affairs.

## Acknolwedgements

This research was supported in part by the NSF, ONR, AFOSR, DARPA, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation. We thank Kevin Xia for the feedback provided in the early versions of this manuscript.

## References

Augustin, M., Boreiko, V., Croce, F., and Hein, M. Diffusion visual counterfactual explanations. *Advances in Neural Information Processing Systems*, 35:364–377, 2022.

Avin, C., Shpitser, I., and Pearl, J. Identifiability of Path-Specific Effects. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, pp. 357–363. Morgan-Kaufmann Publishers, 2005.

Avrahami, O., Lischinski, D., and Fried, O. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18208–18218, 2022.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Balke, A. and Pearl, J. Counterfactual Probabilities: Computational Methods, Bounds, and Applications. In de Mantaras, R. L. and D.˜Poole (eds.), *Uncertainty in Artificial Intelligence 10*, pp. 46–54. Morgan Kaufmann, San Mateo, CA, 1994.

Bareinboim, E., Forney, A., and Pearl, J. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, pp. 1342–1350, 2015.

Bareinboim, E., Correa, J. D., Ibeling, D., and Icard, T. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 507–556. Association for Computing Machinery, New York, NY, USA, 2022.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Brehmer, J., De Haan, P., Lippe, P., and Cohen, T. S. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35:38319–38331, 2022.

Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Chai, L., Wulff, J., and Isola, P. Using latent space regression to analyze and leverage compositionality in {gan}s. In *International Conference on Learning Representations*, 2021.

Chen, T. Q., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. In *NeurIPS*, 2018.

Child, R. Very deep {vae}s generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*, 2021.

Correa, J., Lee, S., and Bareinboim, E. Nested counterfactual identification from arbitrary surrogate experiments. *Advances in Neural Information Processing Systems*, 34: 6856–6867, 2021a.

Correa, J., Lee, S., and Bareinboim, E. Nested counterfactual identification from arbitrary surrogate experiments. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 6856–6867. Curran Associates, Inc., 2021b.

Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castricato, L., and Raff, E. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, pp. 88–105. Springer, 2022.

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

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., and Das, P. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Advances in neural information processing systems*, 31, 2018.

Donahue, J., Krähenbühl, P., and Darrell, T. Adversarial feature learning. In *International Conference on Learning Representations*, 2017.

Dosovitskiy, A. and Brox, T. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, 29, 2016.

Dumoulin, V., Belghazi, I., Poole, B., Lamb, A., Arjovsky, M., Mastropietro, O., and Courville, A. Adversarially learned inference. In *International Conference on Learning Representations*, 2017.

Falcon, W. and Cho, K. A framework for contrastive self-supervised learning and designing a new approach. *arXiv preprint arXiv:2009.00104*, 2020.

Gal, R., Patashnik, O., Maron, H., Bermano, A. H., Chechik, G., and Cohen-Or, D. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022.

Goetschalckx, L., Andonian, A., Oliva, A., and Isola, P. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the ieee/cvf international conference on computer vision*, pp. 5744–5753, 2019.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Goyal, Y., Feder, A., Shalit, U., and Kim, B. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*, 2019a.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., and Lee, S. Counterfactual visual explanations. In *International Conference on Machine Learning*, pp. 2376–2384. PMLR, 2019b.

Härkönen, E., Hertzmann, A., Lehtinen, J., and Paris, S. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33: 9841–9850, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Heckman, J. J. Randomization and Social Policy Evaluation. In Manski, C. and Garfinkle, I. (eds.), *Evaluations: Welfare and Training Programs*, pp. 201–230. Harvard University Press, Cambridge, MA, 1992.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., and Salimans, T. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022.

Hyvärinen, A. and Pajunen, P. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

Jaber, A., Zhang, J., and Bareinboim, E. Causal identification under Markov equivalence. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, pp. 978–987. AUAI Press, Aug 2018.

Jaber, A., Zhang, J., and Bareinboim, E. Identification of conditional causal effects under Markov equivalence. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 11512–11520. Curran Associates, Inc., 2019.

Jaber, A., Kocaoglu, M., Shanmugam, K., and Bareinboim, E. Causal discovery from soft interventions with unknown targets: Characterization and learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9551–9561. Curran Associates, Inc., 2020.

Jaber, A., Ribeiro, A., Zhang, J., and Bareinboim, E. Causal identification under markov equivalence: Calculus, algorithm, and completeness. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 3679–3690. Curran Associates, Inc., 2022.

Jahanian*, A., Chai*, L., and Isola, P. On the "steerability" of generative adversarial networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HylsTT4FvB.

Joshi, S., Koyejo, O., Vijitbenjaronk, W., Kim, B., and Ghosh, J. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217. PMLR, 2020.

Khorram, S. and Fuxin, L. Cycle-consistent counterfactuals by latent transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10203–10212, 2022.

Kim, H. and Mnih, A. Disentangling by factorising. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2649–2658. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kim18b.html.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kocaoglu, M., Shanmugam, K., and Bareinboim, E. Experimental design for learning causal graphs with latent

variables. In *Advances in Neural Information Processing Systems 30*, pp. 7018–7028. Curran Associates, Inc., 2017a.

Kocaoglu, M., Snyder, C., Dimakis, A. G., and Vishwanath, S. Causalgan: Learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023*, 2017b.

Kocaoglu, M., Snyder, C., Dimakis, A. G., and Vishwanath, S. CausalGAN: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations*, 2018.

Kocaoglu, M., Jaber, A., Shanmugam, K., and Bareinboim, E. Characterization and learning of causal graphs with latent variables from soft interventions. In *Advances in Neural Information Processing Systems 32*, pp. 14346–14356, Vancouver, Canada, 2019. Curran Associates, Inc.

Kwon, G. and Ye, J. C. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18062–18071, 2022.

Lachapelle, S., Rodriguez, P., Sharma, Y., Everett, K. E., Le Priol, R., Lacoste, A., and Lacoste-Julien, S. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *First Conference on Causal Learning and Reasoning*, 2021.

Lee, S., Correa, J., and Bareinboim, E. Generalized transportability: Synthesis of experiments from heterogeneous domains. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, 2020. AAAI Press.

Li, A., Jaber, A., and Bareinboim, E. Causal discovery from observational and interventional data across multiple environments. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Locatello, F., Bauer, S., Lucic, M., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. *ArXiv*, abs/1811.12359, 2019a.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019b.

Locatello, F., Poole, B., Raetsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. Weakly-supervised disentanglement without compromises. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6348–6359. PMLR, 13–18 Jul 2020a. URL https://proceedings.mlr.press/v119/locatello20a.html.

Locatello, F., Tschannen, M., Bauer, S., Rätsch, G., Schölkopf, B., and Bachem, O. Disentangling factors of variations using few labels. In *International Conference on Learning Representations*, 2020b. URL https://openreview.net/forum?id=SygagpEKwB.

Manski, C. F. Nonparametric bounds on treatment effects. *American Economic Review, Papers and Proceedings*, 80: 319–323, 1990.

Mao, C., Xia, K., Wang, J., Wang, H., Yang, J., Bareinboim, E., and Vondrick, C. Causal transportability for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7521–7531, 2022.

Mooij, J. M., Magliacane, S., and Claassen, T. Joint causal inference from multiple contexts. *The Journal of Machine Learning Research*, 21(1):3919–4026, 2020.

Moraffah, R., Karami, M., Guo, R., Raglin, A., and Liu, H. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(1):18–33, 2020.

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

Pandey, K., Mukherjee, A., Rai, P., and Kumar, A. DiffuseVAE: Efficient, controllable and high-fidelity generation from low-dimensional latents. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., and Lischinski, D. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2085–2094, 2021.

Pawlowski, N., Coelho de Castro, D., and Glocker, B. Deep structural causal models for tractable counterfactual inference. *Advances in Neural Information Processing Systems*, 33:857–869, 2020.

Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2000.

Pearl, J. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pp. 411–420, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558608001.

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

Pearl, J. and Mackenzie, D. *The Book of Why*. Basic Books, New York, 2018.

Peters, J., Janzing, D., and Schlkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

Plecko, D. and Bareinboim, E. Causal fairness analysis. Technical Report R-90, Causal Artificial Intelligence Lab, Columbia University, Jul 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rodríguez, P., Caccia, M., Lacoste, A., Zamparo, L., Laradji, I., Charlin, L., and Vazquez, D. Beyond trivial counterfactual explanations with diverse valuable explanations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1056–1065, 2021.

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., and Norouzi, M. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.

Samangouei, P., Saeedi, A., Nakagawa, L., and Silberman, N. Explaingan: Model explanation via decision boundary crossing transformations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 666–681, 2018.

Sanchez, P. and Tsaftaris, S. A. Diffusion causal models for counterfactual estimation. In Schölkopf, B., Uhler, C., and Zhang, K. (eds.), *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 647–668. PMLR, 11–13 Apr 2022.

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

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.

Shen, X., Liu, F., Dong, H., Lian, Q., Chen, Z., and Zhang, T. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23:1–55, 2022.

Shen, Y., Gu, J., Tang, X., and Zhou, B. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9243–9252, 2020.

Shpitser, I. and Pearl, J. Effects of Treatment on the Treated: Identification and Generalization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, Montreal, Quebec, 2009. AUAI Press.

Shpitser, I. and Sherman, E. Identification of Personalized Effects Associated With Causal Pathways. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, pp. 530–539, 2018.

Sohn, K., Lee, H., and Yan, X. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Spirtes, P., Glymour, C. N., and Scheines, R. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.

Squires, C., Wang, Y., and Uhler, C. Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1039–1048. PMLR, 2020.

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

Van Looveren, A. and Klaise, J. Interpretable counterfactual explanations guided by prototypes. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II 21*, pp. 650–665. Springer, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Verma, S., Boonsanong, V., Hoang, M., Hines, K. E., Dickerson, J. P., and Shah, C. Counterfactual explanations and algorithmic recourses for machine learning: a review. *arXiv preprint arXiv:2010.10596*, 2020.

Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

von Kügelgen, J., Besserve, M., Wendong, L., Gresele, L., Kekić, A., Bareinboim, E., Blei, D. M., and Schölkopf, B. Nonparametric identifiability of causal representations from unknown interventions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Wang, P. and Vasconcelos, N. Scout: Self-aware discriminant counterfactual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8981–8990, 2020.

Xia, K., Lee, K.-Z., Bengio, Y., and Bareinboim, E. The causal-neural connection: Expressiveness, learnability, and inference. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 10823–10836. Curran Associates, Inc., 2021. URL https://causalai.net/r80.pdf.

Xia, K., Pan, Y., and Bareinboim, E. Neural causal models for counterfactual identification and estimation. In *International Conference on Learning Representations*, 2022.

Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9593–9602, 2021.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. Self-attention generative adversarial networks. In *International conference on machine learning*, pp. 7354–7363. PMLR, 2019.

Zhang, J. and Bareinboim, E. Fairness in decision-making–the causal explanation formula. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pp. 2037–2045, 2018.

Zhang, J., Jin, T., and Bareinboim, E. Partial counterfactual identification from observational and experimental data. In *Proceedings of the 39th International Conference on Machine Learning (ICML-22)*, 2022.

Zhang, J., Squires, C., Greenewald, K., Srivastava, A., Shanmugam, K., and Uhler, C. Identifiability guarantees for causal disentanglement from soft interventions. *arXiv preprint arXiv:2307.06250*, 2023.

# A. Proofs

In this section, we provide proof of the statements in the main body of the paper.

## A.1. Proofs of Corollary 3.1

First, we bring forth the formal definition of layer 1, 2, and 3 valuations, which shows how the SCM valuates observational, interventional distributions, and counterfactual distributions respectively.

**Definition A.1** (Layer 1, 2, 3 Valuation)**.** An SCM $\mathcal{M}$ induces layer $\mathcal{L}_3(\mathcal{M})$, a set of distributions over $\mathbf{V}$, each with the form $P(\mathbf{Y}_*) = P(\mathbf{Y}_{1[\mathbf{x}_1]}, \mathbf{Y}_{2[\mathbf{x}_2],\dots})$ such that

$$P^{\mathcal{M}}(\mathbf{y}_{1[\mathbf{x}_1]}, \mathbf{y}_{2[\mathbf{x}_2]}, \dots) = \int_{\mathcal{X}_{\mathbf{U}}} \mathbb{1}\left[\mathbf{Y}_{1[\mathbf{x}_1]}(\mathbf{u}) = \mathbf{y}_1, \mathbf{Y}_{2[\mathbf{x}_2]}(\mathbf{u}) = \mathbf{y}_2, \dots\right] dP(\mathbf{u}), \tag{35}$$

where $\mathbf{Y}_{i[\mathbf{x}_i]}(\mathbf{u})$ is evaluated under $\mathcal{F}_{\mathbf{x}_i} := \{f_{V_j} : V_j \in \mathbf{V} \setminus \mathbf{X}_i\} \cup \{f_X \leftarrow x : X \in \mathbf{X}_i\}$. The specific set of distributions $P(\mathbf{Y}_{\mathbf{x}})$, where there is only one event, is defined as layer $L_2(\mathcal{M})$. The specific distribution $P(\mathbf{V})$, where $\mathbf{X}$ is empty, is defined as layer $L_1(\mathcal{M})$. ■

Then, we provide the formal Causal Hierarchy Theorem (CHT) here, which states that the layers of the hierarchy remain distinct for almost any SCM.

*Fact* 1 (Causal Hierarachy Theorem (CHT) (Bareinboim et al., 2022, Thm. 1)). Let $\Omega^*$ be the set of all SCMs. We say that Layer $j$ of the causal hierarchy for SCMs collapses to Layer $i$ ($i < j$) relative to $\mathcal{M}^* \in \Omega^*$ if $L_i(\mathcal{M}^*) = L_i(\mathcal{M})$ implies that $L_j(\mathcal{M}^*) = L_j(\mathcal{M})$ for all $\mathcal{M} \in \Omega^*$. then, with respect to the Lebesgue measure over (a suitable encoding of $L_3$-equivalence classes of) SCMs, the subset in which Layer $j$ of NCMs collapses to Layer $i$ is measure zero.

The CHT says that causal questions in the higher layers cannot be answered with knowledge and data restricted to lower layers. More specifically, we can almost surely find a $\mathcal{M}$ for a fixed $\mathcal{M}^*$ such that $L_1(\mathcal{M}^*) = L_1(\mathcal{M})$ ($\mathcal{M}^*$ and $\mathcal{M}$ agree with observational distributions), but $L_2(\mathcal{M}^*) \neq L_2(\mathcal{M}), L_3(\mathcal{M}^*) \neq L_3(\mathcal{M})$ (as illustrated in Figure 13), which implies that quantities in layer 2 and 3 are not uniquely computed.

In this paper, we focus on image counterfactual distributions induced by a special class of SCMs, augmented SCMs (ASCMs) over generative factors $\mathbf{V}$ and $\mathbf{I}$. To prove Corollary 3.1, we aim to prove that it is almost surely that we can find an ASCM $\mathcal{M}' \in \Omega^{\mathbf{I}}$ for the true ASCM $\mathcal{M}^* \in \Omega^{\mathbf{I}}$ such that $L_1(\mathcal{M}^*) = L_1(\mathcal{M}')$, but $L_3(\mathcal{M}^*) \neq L_3(\mathcal{M}')$. Notice that we cannot trivially apply CHT to prove this statement since $\mathcal{M}^*$ and $\mathcal{M}'$ must come from the space of ASCMs $\Omega^{\mathbf{I}}$.

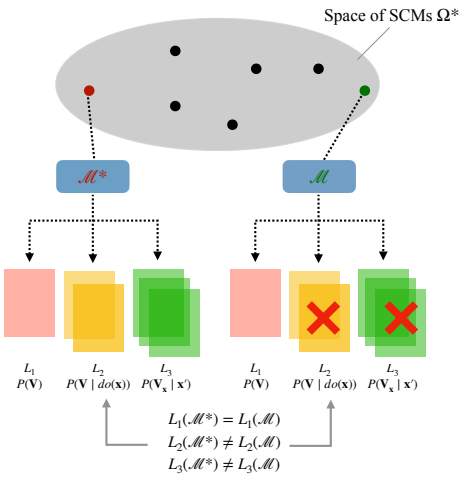



Figure 13: Causal Hierarchy Theorem (CHT).

We provide the following lemma to connect image counterfactual queries induced by an ASCM $\mathcal{M}$ to counterfactual quantities induced by $\mathcal{M}_0$, which is a standard SCM. Let $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(1)}$ be two ASCMs such that $P^{\mathcal{M}^{(1)}}(\mathbf{V}, \mathbf{I}) = P^{\mathcal{M}^{(2)}}(\mathbf{V}, \mathbf{I})$. Then $P^{\mathcal{M}^{(1)}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'}) = P^{\mathcal{M}^{(2)}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$ only if $P^{\mathcal{M}_0^{(1)}}(\mathbf{v}, \mathbf{v}'_{\mathbf{x}'}) = P^{\mathcal{M}_0^{(2)}}(\mathbf{v}, \mathbf{v}'_{\mathbf{x}'})$. ■

*Proof.* Denote $h_{\mathbf{V}}^{(1)}$ and $h_{\mathbf{V}}^{(2)}$ as the function mapping from $\mathbf{I}$ to $\mathbf{V}$ of $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ respectively. We can derive $h_{\mathbf{V}}^{(1)} = h_{\mathbf{V}}^{(2)}$ from Equation (3) since $P^{\mathcal{M}^{(1)}}(\mathbf{V} \mid \mathbf{I}) = P^{\mathcal{M}^{(2)}}(\mathbf{V} \mid \mathbf{I})$. Take the mapping function $\phi$ as $h_{\mathbf{V}}^{(1)} = h_{\mathbf{V}}^{(2)} = h_{\mathbf{V}}$, we have

$$P^{\mathcal{M}^{(1)}}(\mathbf{v}^*, \mathbf{v}'^*_{\mathbf{x}'}) = h_{\mathbf{V}}(P^{\mathcal{M}^{(1)}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})) \tag{36}$$

$$P^{\mathcal{M}^{(2)}}(\mathbf{v}^*, \mathbf{v}'^*_{\mathbf{x}'}) = h_{\mathbf{V}}(P^{\mathcal{M}^{(2)}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})) \tag{37}$$

Since $P(\mathbf{v}^*, \mathbf{v}'^*_{\mathbf{x}'})$ is the output of a function with input $P(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$, $P^{\mathcal{M}^{(1)}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'}) = P^{\mathcal{M}^{(2)}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$ only if $P^{\mathcal{M}_0^{(1)}}(\mathbf{v}, \mathbf{v}'_{\mathbf{x}'}) = P^{\mathcal{M}_0^{(2)}}(\mathbf{v}, \mathbf{v}'_{\mathbf{x}'})$. □

With Fact. 1 and Lem. A.1, we prove Corollary 3.1.

**Corollary A.2** (Image Causal Hierarchy Theorem)**.** *Any image counterfactual distribution is almost never uniquely computable from the observational distribution (or its samples).* ■

*Proof.* Let $\mathcal{M}^*$ be the true underlying ASCM consisting of the generative SCM $\mathcal{M}_0^*$ and $f_{\mathbf{I}}^*$ and $P(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ be arbitrary target image counterfactual distribution. According to CHT, there always exists another generative SCM $\mathcal{M}'_0$ such that $P^{\mathcal{M}'_0}(\mathbf{V}) = P^{\mathcal{M}_0^*}(\mathbf{V})$ but $P^{\mathcal{M}'_0}(\mathbf{v}, \mathbf{v}'_{\mathbf{x}'}) \neq P^{\mathcal{M}_0^*}(\mathbf{v}, \mathbf{v}'_{\mathbf{x}'})$ for some $\mathbf{v}, \mathbf{v}'$. We construct an $\mathcal{M}'$ over $\mathcal{M}_0^{\prime *}$ as following: (1) choose $U_{\mathbf{I}}$ to satisfy $P^{\mathcal{M}'}(\mathbf{V}, \mathbf{U}_{\mathbf{I}}) = P^{\mathcal{M}^*}(\mathbf{V}, \mathbf{U}_{\mathbf{I}})$; (2) let $f'_{\mathbf{I}} = f^*_{\mathbf{I}}$. Then $P^{\mathcal{M}'}(\mathbf{V}, \mathbf{I}) = P^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I})$. According to Lem. A.1, $P^{\mathcal{M}'_0}(\mathbf{v}, \mathbf{v}'_{\mathbf{x}'}) \neq P^{\mathcal{M}_0^*}(\mathbf{v}, \mathbf{v}'_{\mathbf{x}'})$ implies that $P^{\mathcal{M}_0}P(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'}) \neq P^{\mathcal{M}'}P(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$ for any $\mathbf{i}$ and $\mathbf{i}'$ such that $\mathbf{v} = h(\mathbf{i})$ and $\mathbf{v}' = h(\mathbf{i}')$. Then image counterfactual distribution induced by $\mathcal{M}^*$ and $\mathcal{M}'$ are not the same. In other words, we could always construct $\mathcal{M}'$ in the above way for $\mathcal{M}^*$ such that $P^{\mathcal{M}'}(\mathbf{V}, \mathbf{I}) = P^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I})$ but $P^{\mathcal{M}'}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'}) \neq P^{\mathcal{M}^*}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$, which implies that $P(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ is not uniquely computable from $P(\mathbf{V}, \mathbf{I})$. □

### A.2. Proofs of Theorem 3.4

In Sec. 2, we explains that the mechanism $f_{\mathbf{I}}$ plays two roles in the generation process of image variable $\mathbf{I}$: (1) constructing unobserved generative factors $\tilde{\mathbf{U}}$ from $\mathbf{U}_{\mathbf{I}}$ and $\mathbf{V}$; (2) mixing all generative factors $\mathbf{V}$ and $\tilde{\mathbf{U}}$ to images pixels. We first split $f_{\mathbf{I}}$ into factors constructing function $\tau$ and mixing function $f$, and write down this generation process explicitly:

$$\begin{aligned} \tilde{\mathbf{U}} &\leftarrow \tau(\mathbf{V}, \mathbf{U}_{\mathbf{I}}), \\ \mathbf{I} &\leftarrow \tilde{f}(\mathbf{V}, \tilde{\mathbf{U}}). \end{aligned} \tag{38}$$

We assume there exists an unobserved generative factor $\tilde{U} \in \tilde{\mathbf{U}}$ such that $\tilde{U}$ is invertible from the image $\mathbf{I}$, i.e. there exists function $h_{\tilde{U}}(\mathbf{I}) = \tilde{U}$. In other words, a change of $\tilde{U}$ will certainly lead to a change in images. It is reasonable to assume that there exists at least such an unobserved generative factor contributing to the image, regardless of the observed generative factors $\mathbf{V}$. Otherwise, the image may become excessively deterministic from labeled variable $*V$.

With the intuition in Example 3.5, we show the non-identifiability (Theorem 3.4) stands from the mixing of unobserved generative factors.

**Theorem 3.4** (ID)**.** *The image counterfactual distribution $P(\mathbf{I}, \mathbf{I}'_{\mathbf{x}'})$ is not identifiable from any combination of $\langle P(\mathbf{V}, \mathbf{I}), \mathcal{G} \rangle$.* ■

*Proof.* We prove this theorem by constructing two ASCMs $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ that induce the given $P(\mathbf{V}, \mathbf{I})$ but differ from the image counterfactual query $P(\mathbf{I}, \mathbf{I}'_{\mathbf{x}'})$. The true $h_{\mathbf{X}}$ is known since $P(\mathbf{V} \mid \mathbf{I})$ is given.

Consider the image counterfactual query $P(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$ such that $h_{\mathbf{X}}(\mathbf{i}) = \mathbf{x} \neq h_{\mathbf{X}}(\mathbf{i}') = \mathbf{x}'$. This query implies that the original feature $\mathbf{X} = \mathbf{x}$ in the original image $\mathbf{i}$ is changed to $\mathbf{X} = \mathbf{x}'$ in the counterfactual image $\mathbf{i}'$. More specifically, denote a changing variable as $X^{\Delta} \in \{X \in \mathbf{X} \mid x \neq x'\}$.

Consider the unobserved generative factor $\tilde{\mathbf{U}}^+$ that is invertible from $\mathbf{I}$ and denote $\tilde{\mathbf{U}}^- = \tilde{\mathbf{U}} \backslash \tilde{U}^+$. W.L.O.G, we assume $\mathbf{V}$ and $\tilde{U}$ are binary variables.

Suppose $P^{\mathcal{M}^{(1)}}(\tilde{U}^+ = 0 \mid \mathbf{v}) = a$ and $P^{\mathcal{M}^{(1)}}(\tilde{U}^+ = 0 \mid \mathbf{v}') = b$ where $\mathbf{v} = h_{\mathbf{V}}^{(1)}(\mathbf{i})$ and $\mathbf{v}' = h_{\mathbf{V}}^{(1)}(\mathbf{i}')$. Notice that

$a$ and $b$ can be arbitrary probability values since $P(\mathbf{V}, \mathbf{I})$ is arbitrary. W.L.O.G, we assume $a \leq b$. According to Equation (38), $f_{\mathbf{I}}^{(1)}$ in $\mathcal{M}^{(1)}$ can be split as follows:

$$\begin{aligned} \tilde{U}^{+} &\leftarrow \tau^{(1)}(\mathbf{V}, \mathbf{U}_{\mathbf{I}}), \\ \tilde{\mathbf{U}}^{-} &\leftarrow \tau^{-}(\mathbf{V}, \mathbf{U}_{\mathbf{I}}), \\ \mathbf{I} &\leftarrow \tilde{f}(\mathbf{V}, \tilde{U}^{+}, \tilde{\mathbf{U}}^{-}). \end{aligned} \tag{39}$$

Let $U_{\mathbf{I}}^{+} \in \mathbf{U}_{\mathbf{I}}$ be an unobserved parent of $\tilde{\mathbf{U}}^{+}$ and $\mathcal{X}_{\tilde{\mathbf{U}}^{+}} = \{0, 1, 2, 3, 4\}$. Becasue $\tilde{U}^{+}$ is a binary variable, $\tau^{(1)}$ with input $\mathbf{V} = \mathbf{v}$ and $\mathbf{V} = \mathbf{v}'$ can be always rewritten as:

$$\begin{aligned} \tau^{(1)}(\mathbf{v}, \mathbf{U}_{\mathbf{I}}) &= \begin{cases} 0 & U_{\mathbf{I}}^{+} = 0, 1; \\ 1 & U_{\mathbf{I}}^{+} = 2, 3, 4; \end{cases} \\ \tau^{(1)}(\mathbf{v}', \mathbf{U}_{\mathbf{I}}) &= \begin{cases} 0 & U_{\mathbf{I}}^{+} = 0, 1, 2; \\ 1 & U_{\mathbf{I}}^{+} = 3, 4; \end{cases} \end{aligned} \tag{40}$$

where $P(U_{\mathbf{I}}^{+} = 0) = c, P(U_{\mathbf{I}}^{+} = 1) = a - c, P(U_{\mathbf{I}}^{+} = 2) = b - a, P(U_{\mathbf{I}}^{+} = 3) = 1 - b - c, P(U_{\mathbf{I}}^{+} = 4) = c$, where $0 < c < \min(1 - b, a)$ and $U_{\mathbf{I}}^{+}$ is independent with $\mathbf{U}_{\mathbf{I}} \backslash U_{\mathbf{I}}^{+}$. We can verify $P^{\mathcal{M}^{(1)}}(\tilde{U}^{+} = 0 \mid \mathbf{v}) = a$ and $P^{\mathcal{M}^{(1)}}(\tilde{U}^{+} = 0 \mid \mathbf{v}') = b$.

Then we illustrate how to construct $\mathcal{M}^{(2)}$ based on $\mathcal{M}^{(1)}$. First, let $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ equips the same generative SCM, namely $\mathcal{M}_0^{(1)} = \mathcal{M}_0^{(2)}$. Second, we construct $f_{\mathbf{I}}^{(2)}$ as follows:

$$\begin{aligned} \tilde{U} &\leftarrow \tau^{(2)}(\mathbf{V}, \mathbf{U}_{\mathbf{I}}) \\ \tilde{\mathbf{U}}^{-} &\leftarrow \tau^{-}(\mathbf{V}, \mathbf{U}_{\mathbf{I}}), \\ \mathbf{I} &\leftarrow \tilde{f}(\mathbf{V}, \tilde{U}^{+}, \tilde{\mathbf{U}}^{-}), \end{aligned} \tag{41}$$

where $\tau^{(1)}(\mathbf{V}, \mathbf{U}_{\mathbf{I}}) = \tau^{(2)}(\mathbf{V}, \mathbf{U}_{\mathbf{I}})$ when $\mathbf{V} \neq \mathbf{v}'$ and $\mathbf{V} \neq \mathbf{v}$, and

$$\begin{aligned} \tau^{(2)}(\mathbf{v}, \mathbf{U}_{\mathbf{I}}) &= \begin{cases} 0 & U_{\mathbf{I}}^{+} = 0, 1; \\ 1 & U_{\mathbf{I}}^{+} = 2, 3, 4; \end{cases} \\ \tau^{(2)}(\mathbf{v}', \mathbf{U}_{\mathbf{I}}) &= \begin{cases} 0 & U_{\mathbf{I}}^{+} = 1, 2, 3; \\ 1 & U_{\mathbf{I}}^{+} = 0, 4; \end{cases} \end{aligned} \tag{42}$$

In other words, $f_{\mathbf{I}}^{(1)}$ and $f_{\mathbf{I}}^{(2)}$ is only different from the constructing function of $\tilde{U}^{+}$ when $\mathbf{V} = \mathbf{v}'$, namely $\tau^{(1)}(\mathbf{v}', \cdot) \neq \tau^{(2)}(\mathbf{v}', \cdot)$. Third, We construct $P^{(2)}(\mathbf{U})$ to satisfy $P^{\mathcal{M}^{(2)}}(\mathbf{U}) = P^{\mathcal{M}^{(1)}}(\mathbf{U})$.

It is verifiable that

$$\begin{aligned} &P^{\mathcal{M}^{(2)}}(\tilde{U}^{+} = 0 \mid \mathbf{v}) = a, \\ &P^{\mathcal{M}^{(2)}}(\tilde{U}^{+} = 0 \mid \mathbf{v}') = b, \\ &P^{\mathcal{M}^{(2)}}(\mathbf{V}) = P^{\mathcal{M}^{(1)}}(\mathbf{V}) \end{aligned} \tag{43}$$

since $P(U_{\mathbf{I}}^{+} = 0) = P(U_{\mathbf{I}}^{+} = 3)$ and $\mathcal{M}_0^{(1)} = \mathcal{M}_0^{(2)}$. Thus, $P^{\mathcal{M}^{(1)}}(\mathbf{V}, \tilde{U}^{+}, \tilde{\mathbf{U}}^{-}) = P^{\mathcal{M}^{(2)}}(\mathbf{V}, \tilde{U}^{+}, \tilde{\mathbf{U}}^{-})$ which implies $P^{\mathcal{M}^{(1)}}(\mathbf{V}, \mathbf{I}) = P^{\mathcal{M}^{(2)}}(\mathbf{V}, \mathbf{I})$. Also, $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ induces the same causal diagram since $\mathcal{M}_0^{(1)} = \mathcal{M}_0^{(2)}$ and $P^{\mathcal{M}^{(1)}}(\mathbf{U}) = P^{\mathcal{M}^{2)}}(\mathbf{U})$.

On the other hand, let $A_{x^{\Delta}, \tilde{u}^{+}} = \{\mathbf{i} \mid \tilde{f}(\mathbf{V}, \tilde{U}^{+}, \tilde{\mathbf{U}}^{-}), x^{\Delta} \in \mathbf{v}, \tilde{U}^{+} = \tilde{u}^{+}\}$. To illustrate, $A_{x^{\Delta}, \tilde{u}}$ is the range of the function $\tilde{f}(x^{\Delta}, \tilde{u}^{+}, \cdot)$. Notice that $A_{x^{\Delta},0}$, $A_{x^{\Delta},1}$, $A_{x'^{\Delta},0}$, $A_{x'^{\Delta},1}$ are disjoint with each other since $\tilde{U}$ is invertible from $\mathbf{I}$.

Now we argue that $P^{\mathcal{M}^{(1)}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'}) \neq P^{\mathcal{M}^{(2)}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$. Take $\mathbf{i} \in A_{x^{\Delta},0}$ and $\mathbf{i}' \in A_{x'^{\Delta},1}$. Based on Def. A.1, $P(\mathbf{u})$ contributes to $P(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$ only if $\mathbf{1}[\mathbf{I}(\mathbf{u}) = \mathbf{i}] \wedge \mathbf{1}[\mathbf{I}_{\mathbf{x}'}(\mathbf{u}) = \mathbf{i}']$. According to the construction of $\mathcal{M}^{(2)}$, $\mathbf{I}^{(1)}(\mathbf{u}) = \mathbf{I}^{(2)}(\mathbf{u}) = \mathbf{i}$ when $\mathbf{V} = \mathbf{v}$ and $\tilde{U}^{+} = 0$ or $\tilde{U}^{+} = 1$. Given $\mathbf{u}$ such that $\mathbf{I}^{(1)}(\mathbf{u}) = \mathbf{I}^{(2)}(\mathbf{u}) = \mathbf{i}$, we claim that $\mathbf{I}_{\mathbf{x}'}^{(1)}(\mathbf{u}) < \mathbf{I}_{\mathbf{x}'}^{(2)}(\mathbf{u})$. The reason is that $\tilde{U}_{\mathbf{I}}^{+} = 0$ or $\tilde{U}_{\mathbf{I}}^{+} = 1$ makes that $\tau^{1}(\mathbf{v}, \mathbf{u}_{\mathbf{I}}) = 0$ from Equation (40), thus $\mathbf{I}_{\mathbf{x}'}(\mathbf{u}) \notin A_{x'^{\Delta},1}$. Then $P^{\mathcal{M}^{(1)}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'}) = 0$. However, $\tilde{U}_{\mathbf{I}}^{+} = 0$ leads $\tau^{2}(\mathbf{v}, \mathbf{u}_{\mathbf{I}}) = 1$ from Equation (42). Then $0 = P^{\mathcal{M}^{(1)}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'}) < P^{\mathcal{M}^{(2)}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$. □

The above proof constructs two ASCMs that are compatible with arbitrary $P(\mathbf{V}, \mathbf{I})$ and $\mathcal{G}$, but induce different image counterfactual distributions. In the construction, the non-identifiability only comes from the $f_{\mathbf{I}}$ since the ASCMs have the same generative ASCMs over $\mathbf{V}$.

### A.3. Proofs of Theorem 4.6

We first prove Lem. 4.2.

**Lemma A.3.** *Consider the true underlying ASCM $\mathcal{M}^{*}$ over $\{\mathbf{V}, \mathbf{I}\}$, and a feature set with mapping function $\phi = h_{\mathbf{W}}^{*}$, where $h_{\mathbf{W}}^{*}$ is the inverse function of $f_{\mathbf{I}}^{*}$ w.r.t. $\mathbf{W}$, and a a proxy ASCM $\widehat{\mathcal{M}}$ over $\{\mathbf{V}, \mathbf{I}\}$. if $P^{\widehat{\mathcal{M}}}(\mathbf{V}, \mathbf{I}) = P^{\mathcal{M}^{*}}(\mathbf{V}, \mathbf{I})$, then*

$$h_{\mathbf{W}}^{*}(P^{\widehat{\mathcal{M}}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})) = P^{\widehat{\mathcal{M}}}(\mathbf{w}, \mathbf{w}'_{\mathbf{x}'}), \tag{13}$$

*where $\mathbf{w} = h_{\mathbf{W}}(\mathbf{i})$, and $\mathbf{w}' = h_{\mathbf{W}}(\mathbf{i}')$.* ■

*Proof.* $P^{\widehat{\mathcal{M}}}(\mathbf{V}, \mathbf{I}) = P^{\mathcal{M}^{*}}(\mathbf{V}, \mathbf{I})$ implies that for any $\mathbf{i} \in \mathcal{X}_{\mathbf{I}}, \mathbf{w} \in \mathcal{X}_{\mathbf{W}}$, $P^{\widehat{\mathcal{M}}}(\mathbf{w} \mid \mathbf{i}) = P^{\mathcal{M}^{*}}(\mathbf{w} \mid \mathbf{i})$. Based on

Equation (3), $h^{*}_{\mathbf{W}}(P^{\widehat{\mathcal{M}}}(\mathbf{i}, \mathbf{i}'_{x'})) = h^{*}_{\mathbf{W}}$. Then according to Def. 4.1,

$$h^{*}_{\mathbf{W}}(P^{\widehat{\mathcal{M}}}(\mathbf{i}, \mathbf{i}'_{x'}))$$
$$= \int_{\mathbf{i}^{(1)}, \mathbf{i}^{(2)} \in \mathcal{X}_{\mathbf{I}}} \mathbf{1}\left[h^{*}_{\mathbf{W}}(\mathbf{i}^{(1)}) = \mathbf{w}, h^{*}_{\mathbf{W}}(\mathbf{i}^{(2)}) = \mathbf{w}'\right] dP(\mathbf{i}^{(1)}, \mathbf{i}^{(2)}_{\mathbf{x}'}) \tag{44}$$
$$= \int_{\mathbf{u} \in \mathcal{X}_{\mathbf{U}}} \mathbf{1}\left[h^{*}_{\mathbf{W}}(\mathbf{I}(\mathbf{u})) = \mathbf{w}, h^{*}_{\mathbf{W}}(\mathbf{I}_{\mathbf{x}'}(\mathbf{u})) = \mathbf{w}'\right] dP(\mathbf{u}) \tag{45}$$

Def. A.1

$$= \int_{\mathbf{u} \in \mathcal{X}_{\mathbf{U}}} \mathbf{1}\left[\mathbf{W}(\mathbf{u}) = \mathbf{w}, \mathbf{W}_{\mathbf{x}'}(\mathbf{u})) = \mathbf{w}'\right] dP(\mathbf{u}) \tag{46}$$
$$h^{*}_{\mathbf{W}}(\mathbf{I}) = h^{\widehat{\mathcal{M}}}_{\mathbf{W}}$$
$$= P^{\widehat{\mathcal{M}}}(\mathbf{w}, \mathbf{w}_{\mathbf{x}}) \tag{47}$$

Def. A.1

□

Then we prove Thm. 4.6.

**Theorem 4.6** (Counterfactually Consistent Estimation)**.** *$P^{\widehat{\mathcal{M}}}(\mathbf{I}, \mathbf{I}'_{\mathbf{x}'})$ is a Ctf-consistent estimator with respect to $\mathbf{W} \subseteq \mathbf{V}$ of $P^{\mathcal{M}^*}(\mathbf{I}, \mathbf{I}'_{\mathbf{x}'})$ if $\widehat{\mathcal{M}} \in \Omega_{\mathbf{I}}(\mathcal{G})$ and $P^{\widehat{\mathcal{M}}}(\mathbf{V}, \mathbf{I}) = P(\mathbf{V}, \mathbf{I})$.* ■

*Proof.* According to Def. 4.4 and Lem. 4.2, $P^{\widehat{\mathcal{M}}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$ is a causal Ctf-consistent estimation if $P^{\widehat{\mathcal{M}}}(\mathbf{w}, \mathbf{w}'_{\mathbf{x}'})$ is in the optimal bound $[l, r]$ of $P(\mathbf{w}, \mathbf{w}'_{\mathbf{x}'})$ derived from $\mathcal{G}$ and $P(\mathbf{V}, \mathbf{I})$. We recall Def. 1.3 and write down the min/max of the optimization problem for the optimal bound in this setting.

$$\max_{\mathcal{M} \in \mathbb{M}(\mathcal{G})} / \min_{\mathcal{M} \in \mathbb{M}(\mathcal{G})} \quad \phi(P^{\widehat{\mathcal{M}}}(\mathbf{w}, \mathbf{w}'_{\mathbf{x}'})) \tag{48}$$
$$s.t. \quad P^{\mathcal{M}}(\mathbf{V}) = P(\mathbf{V}). \tag{49}$$

It is straightforward that $P^{\widehat{\mathcal{M}}}(\mathbf{w}, \mathbf{w}'_{\mathbf{x}'})$ must be in $[l, r]$ since $\widehat{\mathcal{M}}$ is in the feasible set: $\widehat{\mathcal{M}} \in \mathbb{M}(\mathcal{G}), P^{\widehat{\mathcal{M}}}(\mathbf{V}) = P(\mathbf{V})$ □

## B. Experimental Details

This section provides details about our experiments and models. Our models are primarily written in PyTorch (Paszke et al., 2017), and trained with PyTorch Lightning (Falcon & Cho, 2020).

### B.1. Colored MNIST and Bar

We first provide more details of the architectures of the proposed ANCM and the other three baselines: CVAE, CGN, and DEAR.

**ANCM (ours)**. As illustrated in Alg. 1 and Figure 5, VAE-ANCM use the VAE architecture to maximize $P(\mathbf{I})$ and $P(\mathbf{V} \mid \mathbf{I})$ separately, where the decoder is a $\mathcal{G}$-constrained NCM $\widehat{\mathcal{M}}$ (Def. 1.4). Each function $\widehat{f}_V$ $\widehat{\mathcal{F}}$ in $\widehat{\mathcal{M}}$ is a feed-forward neural network with 2 hidden layers of width 64 with layer normalization applied (Ba et al., 2016). Each exogenous variable $\widehat{U} \in \widehat{\mathbf{U}} \backslash \widehat{\mathbf{U}_{\mathbf{I}}}$ is a 4-dimensional standard normal distribution. The dimension of $\widehat{\mathbf{U}_{\mathbf{I}}}$ is set as 512. The architectures of encoder $Q_{\mathbf{w}}(\tilde{\mathbf{U}} \mid \mathbf{I})$ and $\widehat{f}_{\mathbf{I}}$ designed based on ResNet (He et al., 2016) and are shown in Tab. 1.

**CVAE (Sohn et al., 2015)**. With a latent vector $\mathbf{Z}$ and data samples $\mathbf{i}$, the CVAE is trained to maximize the evidence lower bound (Kingma & Welling, 2013) log-likelihood of the conditional image distributions, namely, $P(\mathbf{I} \mid \mathbf{X})$. Specifically, the optimization objective is as follows:

$$\begin{aligned} ELBO(\boldsymbol{\theta}, \boldsymbol{\omega}) =& \mathbb{E}_{Q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{I},\mathbf{X})}[\log P_{\boldsymbol{\theta}}(\mathbf{I} \mid \mathbf{Z}, \mathbf{X})] \\ & - D_{KL}[Q_{\boldsymbol{\omega}}(\mathbf{Z} \mid \mathbf{I}, \mathbf{X}) \| P(\mathbf{Z})] \end{aligned} \tag{50}$$

where $\mathbf{X}$ is the intervened set ($D$ in this setting), $p(\mathbf{Z})$ is the guassian prior distribution of the latent vector $\mathbf{Z}$, the encoder $Q_{\omega}(\mathbf{Z} \mid \mathbf{I}, \mathbf{X})$ is modeled as a neural network mapping from $\{\mathbf{I}, \mathbf{X}\}$ to $\mathbf{Z}$ parametrized by $\boldsymbol{\omega}$, and the decoder $p_{\boldsymbol{\theta}}(\mathbf{I} \mid \mathbf{Z}, \mathbf{X})$ is modeled as a neural network mapping from $\{\mathbf{Z}, \mathbf{X}\}$ to $\mathbf{I}$. Trained with the re-parameterization trick (Kingma & Welling, 2013), the decoder is capable of generating samples from $P(\mathbf{I} \mid \mathbf{X})$ ideally. We choose the dimension of $\mathbf{Z}$ as 512, and the decoder and encoder are chosen the same as ANCM.

**CGN (Sauer & Geiger, 2021)**. CGN proposes to encode an SCM over variables $Shape, Texture, Background$, and $Label$ into the proxy generative model. Given the label of the image, $Shape, Texture, Background$ are independent. Formally, the mechanism of this SCM is designed as follows:

$$\begin{cases} Label \leftarrow f_l(U_l) \\ Shape \leftarrow \widehat{f}_s(Label, U_d) \\ Texture \leftarrow \widehat{f}_t(Label, U_s) \\ Background \leftarrow \widehat{f}_b(Label, U_b) \\ \mathbf{I} \leftarrow \widehat{f}_{\mathbf{I}}(Shape, Texture, Background), \end{cases} \tag{51}$$

where mechanism $f_s, f_t, f_b$ is designed to learn the conditional distribution $P(V \mid Label)$ with prior knowledge, where $V \in \{Shape, Texture, Background\}$. The composition mechanism $\widehat{f}_{\mathbf{I}}$ is not learned but defined analytically. After fitting the given observational distribution $P(Label, \mathbf{I})$, the intervention can be performed by changing the $Label$. In Colored MNIST and Bar experiments, the digit is regarded as $Shape$; the color is regarded as $background$ and the color is regarded as $Background$. We use the same VAE structure as ANCM to learn mechanism

| Encoder | Decoder $\widehat{f}_{\mathbf{I}}$ |
|---|---|
| $3 \times 3$ 32 conv, $28^2 \uparrow 32^2$, $3ch \to 32ch$ | Concat $\{\mathbf{v}, \mathbf{l}, \mathbf{u}_{\mathbf{I}}\}$, 1*1*512 fully-connected |
| 7 ResBlock, $32^2 \downarrow 16^2$, $32ch \to 64ch$ | 1 ResBlock, $1^2 \uparrow 4^2$, $512ch \to 256ch$ |
| 7 ResBlock, $16^2 \downarrow 8^2$, $64ch \to 128ch$ | 2 ResBlock, $4^2 \downarrow 8^2$, $256ch \to 128ch$ |
| 3 ResBlock, $8^2 \downarrow 4^2$, $128ch \to 256ch$ | 3 ResBlock, $8^2 \downarrow 16^2$, $128ch \to 64ch$ |
| 3 ResBlock, $4^2 \downarrow 1^2$, $256ch \to 512ch$ | 7 ResBlock, $16^2 \downarrow 32^2$, $64ch \to 32ch$ |
| 2 ResBlock, 1*1*512 fully-connected | 8 ResBlock, $3 \times 3$ 3 conv. $32^2 \uparrow 28^2$, $64ch \uparrow 3ch$ |

Table 1: Architectures of the networks for MNIST and Bar experiments. $3 \times 3$ represents a $3 \times 3$ convolutional kernel is adopted. $a^2 \uparrow b^2$ denotes $a \times a$ resolution is upsampled to $b \times b$. $b^2 \downarrow a^2$ denotes $b \times b$ resolution is downsampled to $a \times a$. $i\ ch \to j\ ch$ denotes the channel changes.

$f_s, f_t, f_b$ and the composition mechanism is designed as:

$$\mathbf{I} = \text{Concat}(Shape \odot Texture, Background) \quad (52)$$

where $\odot$ represents an elementwise multiplication operation. Theoretically, CGN learns the independent mechanism from $Shape, Texture, Background$ to the image. After performing interventions on one variable, others should be preserved in the image. This work can represent a branch of works that tries to change some specific features but remains others. However, they do not work on general causal relationships among generative factors (see more discussion in Appendix C.2).

**DEAR (Shen et al., 2022).** DEAR is designed to learn causal representations under the supervision of annotations and Markovian causal diagrams. It encodes the given graph to the latent space by a mask adjacency matrix. DEAR also fits the $P(\mathbf{I})$ and $P(\mathbf{V} \mid \mathbf{I})$ separately similar to ANCM. $P(\mathbf{I})$ is fitted with BiGAN that is capable of learning representation and generating data simultaneously (Donahue et al., 2017; Dumoulin et al., 2017). And $P(\mathbf{V} \mid \mathbf{I})$ is fitted through a regularizer that predicts annotations from the representations. The intervention can be performed by ancestral sampling in the latent space. In this work, we do not aim for learning the representations but only use their way to encode graphs into generative networks for comparison. When implementing DEAR, we simply ignore the bidirected edges in the given graph and encode the graph with directed edges. Theoretically, since DEAR relies on the Markovian assumption, it cannot fit the given observational distribution perfectly. This work can represent a branch of works with the Markovian assumption (see Appendix C.4). In theory, these methods that rely on the Markovian assumption can fail to provide Ctf-consistent estimator. See the following example:

**Example B.1.** Suppose one tries to use a Markovian model $\widehat{M}$ to fit observational distributions shown in Fig. 2 and provide counterfactual image samples for $\mathcal{M}^*$ in Example 2.2. The relationships between $F$ and $Y$ in $\widehat{\mathcal{M}}$ should be (1) $F$ and $Y$ are independent; (2) $F \leftarrow Y$; (3) $F \to Y$;

First, case (1) fails to fit the given observational distribution since they are strongly correlated from the data. Notice that the optimal bound of $P(f_{y'} \mid f, y)$ is $[1, 1]$, which implies $F$ is guaranteed not to change after the intervention on $Y$. $\widehat{\mathcal{M}}$ with case (2) fails to provide such estimation. To illustrate, the optimal bound of $P(y_{f'} \mid f, y)$ is $[1, 1]$, which implies $Y$ is guaranteed not to change after the intervention on $F$. However, $Y$ has a direct effect on $F$ when $\widehat{\mathcal{M}}$ has case (2) property, which means $Y$ is likely to change after the intervention. Case (3) fails to provide such estimations for the same reason. $\widehat{\mathcal{M}}$ cannot provide in-bound estimation of $P(y_{f'} \mid f, y)$. ■

We choose DEAR since it leverages self-attention (Vaswani et al., 2017) and SAGAN (Zhang et al., 2019) architectures for the discriminator and generator and then can generate high-quality images, which is important for the CelebA-HQ experiments. We choose the same network architecture and hyperparameters in the original paper for pendulum experiments.

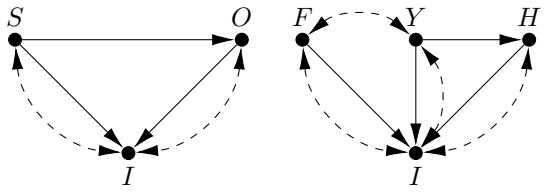
Encoders and decoders for ANCM, CVE, and CGN are trained with a learning rate of $10^{-4}$, and they are optimized with Adam optimizer (Kingma & Ba, 2014). All training processes are performed with a batch size of 100. We choose the temperature $\lambda$ as 100 initially and gradually decrease it during training.

### B.2. CelebA-HQ

The CelebA-HQ experiment is conducted on the CelebA-HQ (Karras et al., 2018) dataset describing human faces. The causal diagram of the ground truth is not given, however, causal diagrams shown in Fig 14a and 14b are used as inductive bias based on prior knowledge about human faces. To illustrate, in *Smiling setting* (Fig 14a), Smiling ($S$) has a positive effect on Open_Mouth ($O$). But $S$ and $O$ can be confounded with $\mathbf{I}$ by unknown generative factors $\mathbf{U}_{\mathbf{I}}$. In *Age setting* (Fig 14b), Young $Y$ are confounded by Female $F$ like Example 1.1 and $Y$ has a negative effect on gray hair color. Similarly, these three generative factors can be confounded with the image variable.

| Encoder | Decoder $\widehat{f_{\mathbf{I}}}$ |
|---|---|
| $3 \times 3$ 64 conv, $128^2 \uparrow 128^2$, $3ch \to 64ch$ | Concat $\{\mathbf{v}, \mathbf{l}, \mathbf{u_I}\}$, fully 1*1*1024 fully-connected |
| 1 ResBlock, $128^2 \downarrow 64^2$, $64ch \to 64ch$ | 1 ResBlock, $1^2 \uparrow 4^2$, $1024ch \to 512ch$ |
| 3 ResBlock, $64^2 \downarrow 32^2$, $64ch \to 128ch$ | 2 ResBlock, $4^2 \downarrow 8^2$, $512ch \to 256ch$ |
| 3 ResBlock, $32^2 \downarrow 16^2$, $128ch \to 128ch$ | 2 ResBlock, $8^2 \downarrow 16^2$, $256ch \to 128ch$ |
| 7 ResBlock, $16^2 \downarrow 8^2$, $128ch \to 256ch$ | 7 ResBlock, $16^2 \downarrow 32^2$, $128ch \to 128ch$ |
| 3 ResBlock, $8^2 \downarrow 4^2$, $256ch \to 512ch$ | 2 ResBlock, $32^2 \downarrow 64^2$, $128ch \to 64ch$ |
| 3 ResBlock, $4^2 \downarrow 1^2$, $512ch \to 1024ch$ | 2 ResBlock, $64^2 \downarrow 128^2$, $64ch \to 64ch$ |
| 2 ResBlock, 1*1*1024 fully-connected | 1 ResBlock, $3 \times 3$ 3 conv. $128^2 \uparrow 128^2$, $64ch \to 3ch$ |

Table 2: Architectures of the networks for CelebA-HQ experiments. $3 \times 3$ represents a $3 \times 3$ convolutional kernel is adopted. $a^2 \uparrow b^2$ denotes $a \times a$ resolution is upsampled to $b \times b$. $b^2 \downarrow a^2$ denotes $b \times b$ resolution is downsampled to $a \times a$. $i\ ch \to j\ ch$ denotes the channel changes.




(a) (b)

Figure 14: Causal diagrams used in CelebA-HQ experiments.

In the Smiling setting, we are given the observed distribution $P(S, O, \mathbf{I})$ (image with labels of $S$ and $O$). We are asking the counterfactual image distribution $P(\mathbf{I}, \mathbf{I}_{O=1})$ (what would the image be had the person opened the Mouth?). We set the care set as $\{S, O\}$. In other words, the counterfactual question we ask is "Given an image, would this person smile had the person opened the mouth?". Based on the graphical constraints induced by Fig 14a and the observed distributions, the optimal bound of $P(S = s, O = o, S_{O=o} = s)$ is $[P(S = s, O = o), P(S = s, O = o)]$; the $P(S = s, O = o, S_{O=o} = s')$ are $[0, 0]$, where $o \neq o'$ and $s \neq s'$. In other words, the smiling feature should be preserved no matter how $O$ changes. In the Age setting, we are given the observed distribution $P(F, Y, H, \mathbf{I})$ (image with labels of $F$, $Y$ and $H$) and we are asking the counterfactual image distribution $P(\mathbf{I}, \mathbf{I}_{Y=0})$. We set the care set as $\{F, Y, H\}$. In other words, the counterfactual question we ask is "Given an image, would this person have gray hair and change gender had the person become old?". Based on the graphical constraints induced by Fig 14b and the observed distributions, changing $Y$ should not change $F$. And $P(F = f, Y = 1, H = h, F_{Y=0} = f, H_{Y=0} = h')$ is bounded by $[r, l]$, where $r = \max(0, P(F = f, Y = 1, H = h) - P(F = f, Y = 1)P(H = h \mid Y = 0))$ and $l = \min(P(F = f, Y = 1, H = h), P(F = f, Y = 1)P(H = h' \mid Y = 0))$. Thus $P(F_{Y=0} = 0, H_{Y=0} = 1 \mid F = 0, Y = 1, H = 0)$ is bounded by $[0.179, 0.1808]$ from the observational data. To illustrate, given a young male image, the probability he has gray hair is at least 0.179 but at most 0.181.

Similarly to the last section, our causal method ANCM is compared with CVAE and DEAR in this section. We do not implement CGN since the generative variables for CGN are limited in $Shape, Texture, Background$, In this experiment, the composition function from face generative factors to images cannot be designed by hand.

We leverage DiffuseVAE (Sanchez & Tsaftaris, 2022), a two-stage method incorporating VAE and the Diffusion Probabilistic Models (DDPM) techniques, to get high-quality images. At the first stage, CVAEs are trained with association information and have the ability to edit target generative factors. ANCMs are trained with causal information (causal diagram) and have the ability to intervene on generative factors causally. However, both of them produce blurred images, $\hat{\mathbf{i}}$, due to the naive VAE structure. In the second stage, these blurred image samples are refined by DDPM to high quality $\mathbf{i}$. More details about the training procedure and architectures are provided in the next section. Figure 15 illustrates the internal two-stage image samples for DiffuseVAE in the Smiling setting. In the VAE stages, ANCMs and CVAE generate the initial images. After performing intervention $O = 1$ on each image by corresponding models, the counterfactual images are produced. Notice that in these VAE stages, image samples exhibit blurred properties. At the DDPM stage, both initial images and counterfactual images are refined to the final results.

The experimental results are shown in Fig. 12. Other additional results are provided in Fig. 17 (Smiling setting) and Fig. 18 (Young setting). In the Smiling setting, all methods open the person's mouth. ANCM preserves $S$ after changing $O$. DEAR also successfully changes $O$ without changing $S$ since the causal diagram among $S$ and $O$ is Markovian. However, CVAE changes the smiling features at the same time. Other features, such as background are not guaranteed to be the same since they might be confounded

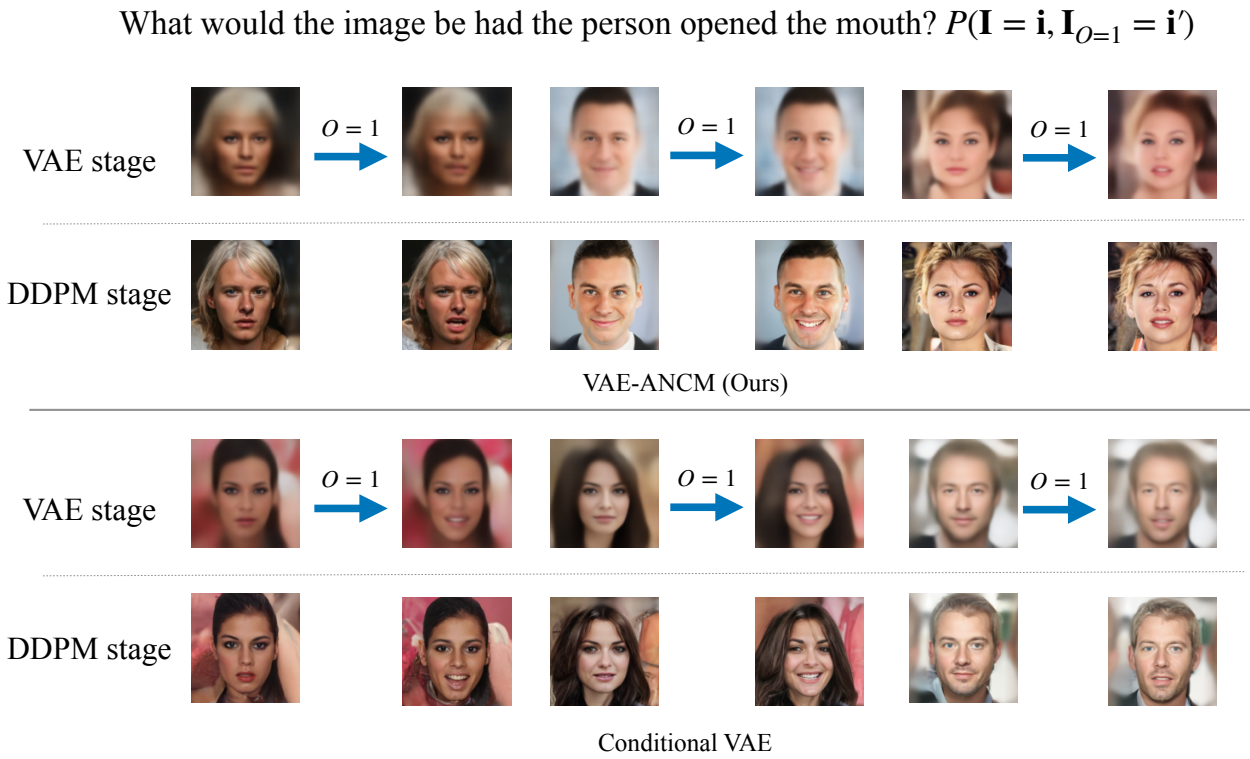



Figure 15: Internal image samples of the CelebaHQ Experiment in the Smiling setting.

with $S$ and $O$.

In the Age setting, all methods make the person's appearance older. ANCMs preserve $F$ after changing $Y$. However, the other two baselines change the gender features at the same time. DEAR does not work in this setting since the causal diagram is non-Markovian. Gray hair is possible to appear after changing the Age for all three methods.

### B.2.1. Models and Hyperparameters

As we discussed above, the training process has two stages: (1) the VAE stage; (2) the DDPM stage. The training procedure at the VAE stage is the same as the VAE training introduced in the Colored MNIST and Bar experiments. We choose the different architecture for encoders and $\widehat{f}_{\mathbf{I}}$ as shown in Tab. 2. We set the dimension of the latent vector $\mathbf{z}$ as 1024 for CVAEs. For ANCMs, exogenous variables of $\mathbf{I}$ are $1024/\max(c, 1)$-dimensional standard normal distributions, where $c$ is the number of bidirected arrows to $\mathbf{I}$. Other exogenous variables are 10-dimensional standard normal distributions. For DEAR, the network architecture and hyperparameters are chosen as the same in (Shen et al., 2022).

In the second stage, the DDPM procedure is adopted to refine the decoder output at the VAE stage. DDPMs (Sohn et al., 2015; Ho et al., 2020) are deep generative models that consist of a forward process and reverse process with $T$ time-steps. The forward process is modeled as a Markov chain that gradually perturbs $\mathbf{i}^{t-1}$ (the image at step $t-1$) with gaussian noise to $\mathbf{i}^t$ (the image at step $t$), where $\mathbf{i}^0$ (image at step 0) is the original image sample. Denoting the image variable at time step $t$ as $\mathbf{I}^t$ ($\mathbf{I}^0 = \mathbf{I}$), the forward model can be expressed as

$$q(\mathbf{I}^{1:T} \mid \mathbf{I}_0) = \prod_{t=1}^{T} q(\mathbf{I}^t \mid \mathbf{I}^{t-1}) \tag{53}$$

The reverse process is also modeled as a Markov chain that gradually denoise $\mathbf{i}^{t-1}$ to $\mathbf{i}_t$ at each time step and finally recovers the original $\mathbf{i}$. Formally,

$$p_\varphi(\mathbf{I}^{0:T}) = p(\mathbf{I}^T) \prod_{t=1}^{T} p_\varphi(\mathbf{I}^{t-1} \mid \mathbf{I}^t) \tag{54}$$

where $p_\varphi$ implies the density is modeled by a neural network parameterized by $\varphi$ and $p(\mathbf{I}^T)$ is often chosen to be an isotropic Gaussian distribution. To maximize the log-likelihood of $P(\mathbf{I})$, the following objective is minimized suggested by (Ho et al., 2020):

$$\begin{aligned} &\mathbb{E}_q[D_{KL}(q(\mathbf{I}^T \mid \mathbf{I}^0) \| p(\mathbf{I}^T)) + \\ &\sum_{t>1} D_{KL}(q(\mathbf{I}^{T-1} \mid \mathbf{I}^T, \mathbf{I}^0) \| p_\varphi(\mathbf{I}^T)) \\ &- \log p_\varphi(\mathbf{I}^0 \mid \mathbf{I}^1)] \end{aligned} \tag{55}$$

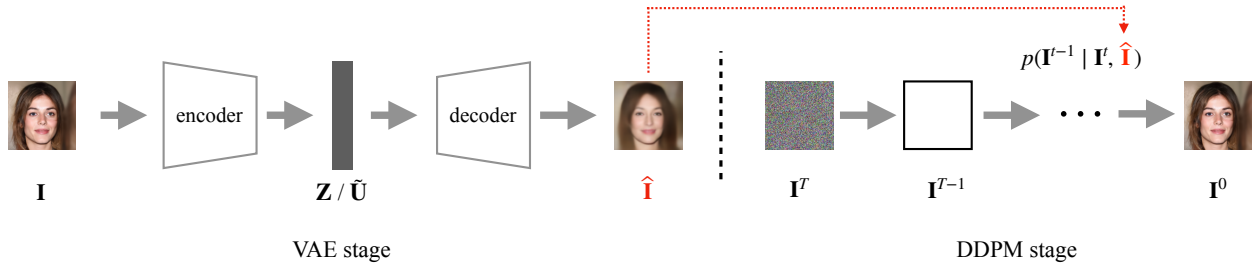



Figure 16: Two stages of DiffuseVAE.

When the image $\mathbf{I}$ has an associated conditioning signal $\mathbf{C}$, for example, a low-resolution image (Ho et al., 2022; Saharia et al., 2022) or a classification label, one can maximize the log-likelihood of the conditional distribution $P(\mathbf{I} \mid \mathbf{C})$ with

$$\begin{aligned} &\mathbb{E}_q[D_{KL}(q(\mathbf{I}^T \mid \mathbf{I}^0) \| p(\mathbf{I}^T)) \\ &+ \sum_{t>1} D_{KL}(q(\mathbf{I}^{T-1} \mid \mathbf{I}^T, \mathbf{I}^0, \mathbf{C}) \| p_\omega(\mathbf{I}^T)) \\ &- \log p_\omega(\mathbf{I}^0 \mid \mathbf{I}^1, \mathbf{C})] \end{aligned} \tag{56}$$

In our experiments, we leverage the reconstructions of the VAEs as the conditional signal at the second stage as suggested by DiffuseVAE. For concreteness, we freeze the encoder and decoder of the VAE trained at the first stage and train $p_\omega$ individually with objective Eq. 56, where $\mathbf{C}$ is the reconstructions of VAEs. The whole training procedure is illustrated in Fig. 16.

All encoders and decoders are trained with a learning rate of $10^{-4}$ and are optimized with Adam optimizer (Kingma & Ba, 2014). All training processes are performed with a batch size of 64. We choose the temperature $\lambda$ as 100 and gradually decrease it. At the DDPM stage, we choose the same hyperparameters and architecture as (Pandey et al., 2022).

## C. Connection to related works

This paper systematically establishes the counterfactual image editing tasks within the framework of causal language. In this section, we explicitly discuss some subtleties of extant research about image editing and counterfactual data generation. By integrating the causal framework proposed in this paper, we (1) interpret the empirical findings and current evaluation metrics already present in the literature; (2) show other methods may not be as general as the method discussed in this paper due to their different assumptions and problem settings. Before that, we first present the following propositions for discussion.

**Proposition C.1.** *Let the care set $\mathbf{W} = \mathbf{X}$. Any proxy model $\widehat{\mathcal{M}}$ provides Ctf-consistent estimator regarding $\mathbf{W}$ such that $P^{\widehat{\mathcal{M}}}(\mathbf{V}, \mathbf{I}) = P^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I})$* ■

*Proof.* For any ASCM, the normalized feature counterfactual distribution $P(\mathbf{W}'_{\mathbf{x}'} = \mathbf{w}' \mid \mathbf{W} = \mathbf{w}) = P(\mathbf{X}'_{\mathbf{x}'} = \mathbf{w}' \mid \mathbf{X} = \mathbf{w}) = 1$ if the intervened value is $\mathbf{x}'$ consistent with $\mathbf{w}'$. Otherwise, this quantity is 0. This implies any ASCM provides Ctf-consistent estimators when $\mathbf{W} = \mathbf{X}$ from Def. 4.4. □

**Proposition C.2.** *There exist settings in which the lower bound of normalized feature counterfactual query $P^{\mathcal{M}^*}(\mathbf{w}'_{\mathbf{x}'} \mid \mathbf{w})$ is strictly bigger than 0 and smaller than 1 and $v \neq v'$, where $V \in \mathbf{W} \backslash \mathbf{X}$.* ■

*Proof.* See Example. 4.3 and 4.5. To illustrate, the optimal bound of the feature counterfactual query $Q$ (Eq. 18) is 0.25 and the value of $H$ differs in observation $\mathbf{w}$ and counterfactual $\mathbf{w}'$. □

**Proposition C.3.** *Consider the true underlying ASCM $\mathcal{M}^*$ with a semi-Markovian graph $\mathcal{G}$. There exist settings where any model $\widehat{\mathcal{M}}$ in which the Markovianity assumption (also known as causal sufficiency) is enforced cannot satisfy (1) $P^{\widehat{\mathcal{M}}}(\mathbf{V}, \mathbf{I}) = P^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I})$ and (2) any feature counterfactual query is in the optimal bounds.* ■

*Proof.* See the counterexample provided in Example B.1. □

### C.1. Counterfactual Visual Explanation

Counterfactual visual explanation aims for the following question: "Given an image $\mathbf{i}$ for which a vision system predicts class $C = c$, how $\mathbf{i}$ could change such that the system would output a different specified class $C = c'$?". Earlier methods construct adversarial sample $\mathbf{i}'$ as the counterfactual image of $\mathbf{i}$. Namely, the goal is to query a permutation $\delta$ such that $\mathbf{i} + \delta$ is labeled as $C = c'$ by the given predictor. (Wang & Vasconcelos, 2020) search the region contributes most to the prediction in pixel levels. Other approaches map the given image to the feature space and find minimal and sufficient input features to flip the prediction (Dhurandhar

et al., 2018; Goyal et al., 2019b; Van Looveren & Klaise, 2021). However, the optimization process of search adversarial samples can push $\mathbf{i}'$ far away from the data manifold of images. To enhance the realisticness of counterfactual images, VAEs (Goyal et al., 2019a; Joshi et al., 2019; Rodríguez et al., 2021), GANs (Samangouei et al., 2018; Khorram & Fuxin, 2022) and Diffusions (Augustin et al., 2022) are incorporated to push counterfactual images close to the training distributions. In the course of these studies, additional properties (as shown below) have been suggested to assess the generated counterfactual image, aside from realism (Verma et al., 2020; Moraffah et al., 2020). It is noteworthy that these properties do not necessarily have to be satisfied simultaneously.

**Validity.** The counterfactual image should be labeled as the target $C = c'$ by the given predictor. Suppose one wants to edit an image $\mathbf{i}$ describing a young person to an older look in Example 1.1. Validity states that the pixels of counterfactual image $\mathbf{i}'$ should truly constitute old features.

**Sparsity.** The permutation $\delta$ should affect a minimal number of features. Following the previous example, sparsity encourages that counterfactual image $\mathbf{i}'$ preserves features as much as possible.

**Proximity.** The counterfactual image $\mathbf{i}'$ should stay as close to the initial pixel-level image. Proximity also motivates the counterfactual image not to change too much. Nevertheless, compared to sparsity, proximity focuses more on the change of each pixel. For example, the 180-degree rotation of an initial image satisfies sparsity (preserves almost all features) but does not satisfy proximity since each individual pixel varies.

**Diversity.** The counterfactual image $\mathbf{i}'$ should be as diverse as possible. To illustrate, $\mathbf{i}'$ should not be unique but are motivated to be different from each other. Continue the previous example. The diversity encourages that the old version images of the original young one scan have different old extent and various other features.

Our causal framework and estimation method provide explanations for these metrics. First, The invertibility of $f_{\mathbf{I}}$ in ASCMs supports the validity theoretically. To illustrate, the generative factor $Old = 1$ causes image pixels containing old features through $f_{\mathbf{I}}$, and images are predicted correctly as old class through the invertible function $h$. Second, the sparsity and proximity are challenged. The achievement of sparsity and proximity implies that only the intervened generative factors $\mathbf{X}$ change while other generative factors keep the same. This implies that $\mathbf{X}$ does not have a causal effect on other generative factors, which only happens in the causal diagram that all other generative factors are non-descendants of $\mathbf{X}$. Formally, based on Prop. C.2, sparsity and proximity should not always be satisfied. Third, the non-identifiability results (Thm. 3.4) support the diversity. Thm. 3.4 states that ASCMs that are compatible with the same observational distribution and the same causal diagram can induce different image counterfactual distributions. Diversity can be obtained by sampling different counterfactual images from $P(\mathbf{I}_{\mathbf{x}'} \mid \mathbf{I} = \mathbf{i})$.

In summary, these designed metrics are aligned with our causal framework of counterfactual image editing. However, we should notice that sparsity and proximity are limited since they ignore the causal effect from the intervened concept to other concepts in images and only work to some specific causal diagram. We propose Ctf-consistent estimators (Def. 4.4) for general causal relationships. For any causal relationships among generative factors, Def. 4.4 is guaranteed to offer in-bound estimation regarding the cared feature set $\mathbf{W}$.

### C.2. Manipulate latent spaces

Generative networks, such as GANs (Goodfellow et al., 2020; Karras et al., 2018; Brock et al., 2019; Karras et al., 2019), VAEs (Kingma & Welling, 2013; Child, 2021; Vahdat & Kautz, 2020), and Diffusions (Ho et al., 2020; Song et al., 2021) fit the distribution $P(\mathbf{I})$ to learn a non-linear mapping from latent variables (generative factors) $\mathbf{Z}$ to the real image variable $\mathbf{I}$. Feeding sampled $\mathbf{z}$ from the latent space, GANs, VAEs, and Diffusions are capable of producing photo-realistic image $\mathbf{i}$. By manipulating the latent space values $\mathbf{z}$ to $\mathbf{z}'$, features in the original sample $\mathbf{i}$ controlled by $\mathbf{z}$ are modified by $\mathbf{z}'$ and counterfactual image $\mathbf{i}'$ are produced. One key point in this procedure is to find a proper way to guide the manipulation of latent vector $\mathbf{z}$.

Shen et al. (2020) construct a hyperplane between class $V_1 = v_1$ and class $V_1 = v_1'$ (for example, old and young) and move $\mathbf{z}$ along with the vertical direction to the hyperplane. This could be regarded as a process of fitting $P(\mathbf{V} \mid \mathbf{I})$. The work only cares about the intervened generative factor $V_1$, namely the cared feature set $\mathbf{W}$ only contains a single variable in our framework. Also, Shen et al. (2020) proposes conditional manipulation when users care about multiple concepts in images. Conditional manipulation changes only one target and other concepts are preserved by enforcing the same distance to hyperplanes constructed by other concepts.

Similarly, several other works perform latent vector manipulation with such linear attribute vector editing directions (Goetschalckx et al., 2019; Jahanian* et al., 2020; Karras et al., 2019). Chai et al. (2021) train a regression model mapping from images to corresponding latent vectors and utilize it to compose the initial image with desired features. Khorram & Fuxin (2022) proposes a cycle-consistent procedure to learn the transformation from the initial latent vector $\mathbf{z}$ to the counterfactual latent vector $\mathbf{z}'$. Recently, text

information has been leveraged into image editing as well. The image description in written text is beneficial to the encoding process and guiding the manipulation in latent space (Radford et al., 2021; Avrahami et al., 2022; Crowson et al., 2022; Gal et al., 2022; Kwon & Ye, 2022; Patashnik et al., 2021) and the natural editing instruction text can be directly used to prompt the transition from the original images to counterfactual images (Brooks et al., 2023).

From Prop. C.1, all these similar methods provide counterfactually consistent estimators when the case set is equal to the intervened set. However, these works ignore the effect of the intervened features on other features. Even some paper discusses the situation where other features should be preserved (such as conditional manipulation) while this only applies to restrict causal relationships among generative factors (Prop. C.2). This point can be found in both Example 4 and colored MNIST and bars experiments (Sec. 5.1). Example 4 shows that the probability that the hair color changed to gray had the person gotten older should be at least 0.25. However, the approaches mentioned above cannot guarantee this counterfactual quantity and even may never change the gray hair. On the other hand, experiments in the backdoor setting show that the probability that the bar disappeared had the digit changed should be at least 0.66. ANCMs are counterfactually consistent with the true model while these approaches (such as baseline CGN) will totally ignore this causal effect.

### C.3. Causal Representation Learning

Another branch of work does not focus on how to guide $\mathbf{z}$ in the latent space of existing models but aims to derive an explainable latent space while training. To illustrate, these works expect that each dimension of $\mathbf{z}$ represents a specific generative factor. Then manipulation can be achieved simply by modifying these corresponding dimensions of the intervened generative factors. These explainable latent vectors are called *disentangled representations* for generative factors. Even though there is a lack of a formal definition of disentanglement, the key intuition is that a disentangled representation should separate the distinct informative factors of variations in the data (Bengio et al., 2013). Early approaches largely enforce statistical independence among latent variables $\mathbf{Z}$ (Higgins et al., 2017; Kim & Mnih, 2018; Chen et al., 2018). However, (Locatello et al., 2019a) argues that unsupervised disentanglement learning is impossible without additional inductive bias on the model or data. This impossible result is covered in Thm. 3.4 when the causal diagram shows independence among all generative factors. Thus, recent works shift to weakly supervised settings to overcome the non-identifiability (Locatello et al., 2020a;b; Lachapelle et al., 2021).

Despite the successful disentanglement learning in literature, they assume all generative factors are independent of each other, which only works for a special causal diagram. More recently, some methods propose to encode the structural causal model into the latent space and aim for causal disentangled representation learning (Yang et al., 2021; Shen et al., 2022; Brehmer et al., 2022; Zhang et al., 2023; von Kügelgen et al., 2023). These works aim to learn disentangled representations of latent generative factors and even the causal diagram from interventional data. After learning the causal representations, these work demonstrate their ability to edit images causally edit images to some extent by manipulating the latent space.

We should point out that the main contributions (learning the causal representations) of these papers are not the same (even orthogonal to) this work. The goal of this work is not to learn the disentangled latent representations behind the images but to argue how we counterfactually generate images even when the label of generative factors and the causal diagram are directly given to us. In other words, even though the causal representations are successfully learned by these works, there is still work that needs to be done for editing images. The non-identifiability result (Thm. 3.4) implies samples obtained by manipulating the representations are hardly from the true image counterfactual distribution. There is no explanation or metrics to evaluate how these samples are causally consistent with the true model. Thm. 4.6 provides such explanations, which says such methods provide in-bound estimations when the causal constraints are correctly encoded. In addition, we should also notice the SCMs that these works encoded into the neural networks are not as general as our approaches (this will be illustrated in the next section).

### C.4. Estimation Counterfactual distributions by Encoding SCMs into Neural Networks

A growing literature uses modern neural methods to estimate counterfactual queries for high-dimensional data (Kocaoglu et al., 2018; Pawlowski et al., 2020; Sanchez & Tsaftaris, 2022; Sauer & Geiger, 2021; Yang et al., 2021; Shen et al., 2022; Brehmer et al., 2022; Zhang et al., 2023; Von Kügelgen et al., 2021). For instance, CausalGAN (Kocaoglu et al., 2017b) is the first work to encode the constraints induced by the causal diagram into generators of GANs. Implemented through GANs, flow-based and Diffusions architectures, (Pawlowski et al., 2020) and (Sanchez & Tsaftaris, 2022) evaluates counterfactual quantities through a three-step procedure from Pearl (2000, Thm. 7.1.7): abduction, action and prediction. (Sauer & Geiger, 2021) generate counterfactual images over a proxy SCM that only contains three generative factors *shape*, *texture*, and *background*. Several perspectives limit these works in general settings.

**Identifiability discussion.** The identifiability of the query

should be discussed before estimation, otherwise, it is unclear how the error between the estimation and the ground truth will be. Our work is the first one to talk about the identifiability of counterfactual queries when the mechanism $f_{\mathbf{I}}$ is invertible (Thm. 3.4). Even Nasr-Esfahany & Kiciman (2023) talks bout the identifiability of bijective SCMs, however, the invertibility is different from ASCMs and bijective SCMs. To illustrate, bijective SCMs assume the unobserved parents $\mathbf{U}_{V_j}$ are determined by both $V_j$ and parents of $V_j$ for any $V_j$. On the other hand, while ASCMs only the $f_{\mathbf{I}}$ is invertible and generative factors $\mathbf{V}$ are fully determined by the image variable $\mathbf{I}$.

**Non-parametric SCMs with semi-Markovian diagrams.** Restrict assumptions are made for SCMs in existing works. Plenty of works assume their SCMs are Markovian, which implies the absence of unobserved confounding among generative factors. While this assumption may hold in specific settings, the same is certainly strong and does not hold in many others (Kocaoglu et al., 2018; Pawlowski et al., 2020; Sanchez & Tsaftaris, 2022; Sauer & Geiger, 2021; Shen et al., 2022). Prop. C.3 states that Markovian proxy models can fail to counterfactually estimate image counterfactual distributions in general. Also, the experiments show these Markovian proxy models (such as DEAR) fail to provide counterfactually consistent edits. Other works focus on parametric SCMs over generative factors, such as linear mechanisms and additive noise, while we study a more general class of non-parametric models (Yang et al., 2021; Shen et al., 2022).

**High-quality image samples.** Naive VAEs and GANs are adopted as the network structure for many existing neural causal estimation methods (Kocaoglu et al., 2018; Pawlowski et al., 2020; Sauer & Geiger, 2021; Yang et al., 2021; Brehmer et al., 2022), which is not capable of generating high-quality image samples. (Sauer & Geiger, 2021) leverages pre-trained BigGAN (Brock et al., 2019) to generate high-quality images that can be decomposed to shape, texture, and background. Shen et al. (2022) overcome the blur issue of VAE generation by self-attention techniques. Sanchez & Tsaftaris (2022) and our work achieve SOTA using diffusion models.

## D. Frequently Asked Questions

Q1. Is it reasonable to expect that the causal diagram is available? Why don't you learn it from the data?

**Answer**. First, the assumption of the causal diagram is made out of necessity. The Image CHT result (Thm. 3.1 in the main body) formally states that image counterfactual distributions are almost never recoverable from the observational data alone. Causal assumptions should be made to make progress. The causal diagram is a well-known flexible data structure that is used throughout the literature to encode a qualitative description of the generating model, which is often much easier to obtain than the actual mechanisms of the generating SCM (Pearl, 2000; Spirtes et al., 2000; Peters et al., 2017). Even though non-identifiability is still present given the causal diagrams (Thm. 3.4), the underlying causal assumptions are beneficial to offer the range of some possible ground truth (in the form of optimal bounds). As assumptions are strengthened, the bounds naturally narrow. The goal of this paper is not to decide which set of assumptions is the best but rather to provide the toolkit for AI engines to perform the inferences once the assumptions have already been made as well as understanding the trade-off between assumptions and the guarantees provided by the method.

Second, the true underlying causal diagrams cannot be learned only from the observational distribution in general. More specifically, there almost surely exist situations that $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ induce the same observational distribution but are compatible with different causal diagrams (see (Bareinboim et al., 2022, Sec. 1.3) for details). With higher layer distributions (such as distributions from $L_2$), it is possible to recover an equivalence class of diagrams (Kocaoglu et al., 2017a; 2019; Jaber et al., 2020; Li et al., 2023; von Kügelgen et al., 2023). In practice, the interventional distribution is not available in many image editing tasks. As a comparison, the causal diagram itself can be more easily obtained from human knowledge in common image editing tasks. For example, getting old will lead to gray hair appearing in Example 1.1. Our approach leverages such human inductive bias towards generative factors to obtain counterfactually consistent estimators.

Q2. How could the graphical assumption be relaxed? What if the causal diagram from human knowledge is incomplete or even wrong?

**Answer**. Relaxing the graphical assumption is out of the scope of this work since the goal of this paper is first to establish a causal framework of counterfactual image editing and perform counterfactually consistent estimation under general well-understood assumptions. Still, building on the current understanding, we believe that finding a solution to the problem when only partial information about the graph is available is an important direction for future work. One such possible approach is to utilize equivalence classes of causal diagrams, learnable from data, to perform inferences (Jaber et al., 2018; 2019; 2022; Mooij et al., 2020; Squires et al., 2020).

On the other hand, interestingly, even if the assumed

causal diagram $\bar{\mathcal{G}}$ is not aligned with the ground truth $\mathcal{G}^*$, the generative model performs some sort of "hallucination" through the given input diagrams. More specifically, as long as the proxy model $\widehat{\mathcal{M}}$ is able to fit the observational distribution with constraints induced by $\bar{\mathcal{G}}$, $\widehat{\mathcal{M}}$ offers counterfactually consistent estimator for the imaging world. For example, suppose the age will directly change human genders in a hallucinating world. Encoding edge $Gender \leftarrow Age$ into $\widehat{\mathcal{M}}$ will achieve this hallucination.

Q3. Is the non-identifiability results proved in this paper a known result in causal representation learning?

**Answer**. From a theoretical perspective, we formalize the counterfactual image editing problem with a causal framework. We list our contributions as follows and will discuss them one by one.

First, we prove the non-identifiability of image counterfactual distributions from only the observational distribution (Thm. 3.1). In causal representation learning literature, there is a known result saying that disentangled causal representations are not identifiable given only the i.i.d samples from the observational distribution (Locatello et al., 2019b; Lachapelle et al., 2021; Hyvärinen & Pajunen, 1999; Khemakhem et al., 2020). However, this result is not the same as the first contribution (a), i.e., Thm. 3.1. Specifically, the goal of our work is not to infer the latent causal representations but to query a $\mathcal{L}_3$-layer distributions $P(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$. To illustrate, the supervision information of generative factors (labels of $\mathbf{V}$) is directly given to us. Even if causal representations can be successfully learned for generative factors $\mathbf{V}$, the impossibility of querying the image counterfactual distributions still holds. See Appendix C.3 for a more detailed comparison with the causal representation learning literature.

Second, we show the image counterfactual distributions are not identifiable given both observational distributions (Thm. 3.4). Inferring counterfactual queries from lower-layer distributions and causal assumptions is a fundamental question in causal inference. In the more traditional literature of causal inference, there are different symbolic methods for solving these problems in various settings and under different assumptions (Heckman, 1992; Pearl, 2001; Avin et al., 2005; Shpitser & Pearl, 2009; Shpitser & Sherman, 2018; Zhang & Bareinboim, 2018; Correa et al., 2021b). Recently, NCMs (Xia et al., 2022) also demonstrate its ability for identification through solving an optimization procedure. This work (Thm. 3.4) is the first one to show the non-identifiability of image counterfactual queries given observational distribution and the causal diagram. We refer more discussion about this identifiability result to Appemdix .C.4.

Third, we propose counterfactual (Ctf) consistent estimators (Def. 4.4) for such non-identifiable situations and give a sufficient condition to obtain this estimator (Thm. 4.6). This is one of the most important parts of this work. Even in non-identifiable settings, Ctf-consistent estimators offer a way to causally edit images in a reasonable way. As illustrated in Sec. 4, Ctf-consistent estimators ensure the feature counterfactual queries are in the optimal bound, which implies the error from the estimation to the ground truth is controlled, and this is the best we can do given the observational distribution and the causal diagram. On the other hand, a Ctf-consistent estimator could also be a metric to evaluate if other methods perform causal editing regarding the cared feature set. And this metric interprets the former metric designed for counterfactual visual explanation, as elaborated in Appendix .C.1.

From an application perspective, we design an algorithm ANCM to provide Ctf-consistent estimators for the target image counterfactual distributions and get samples from it. ANCM encodes the causal constraints induced by the given causal diagram into neural networks. Even some existing works also aim to encode SCMs into the deep generative networks, however, they are restricted in several perspectives (see Appendix C.4). Experiments in Sec. 5 and Appensix B also demonstrate our method provides causally consistent editings while existing methods do not in general settings.

Q4. Is the editing goal of this paper to change the intervened features in the image and keep other features the same?

**Answer**. This is a good question and the answer is not necessarily. The situation is a bit more nuanced than that. The goal of this work is to provide causally consistent editing results with the underlying ground truth. Many existing works try to edit certain features and prevent this edit from affecting other features (Shen et al., 2020; Goetschalckx et al., 2019; Jahanian* et al., 2020; Karras et al., 2019; Chai et al., 2021; Khorram & Fuxin, 2022). However, this situation only works for very specific causal relationships among generative factors and there are settings where this type of editing does not provide counterfactually consistent editing (see Prop. C.2). For example, editing $Age$ but not changing $Gender$ is counterfactually consistent with the true model in Example 2.2. However, if one always keeps the gray hair in the counterfactual image the same as the original image, this editing is not counterfactually consistent as illustrated in Example 4.5. This point is shown in a more practical way through the

colored MNIST and bars experiment (Sec. 5.1). After changing the digit, the causal effect from Digit to Bar should be reflected. One cannot always keep the bar the same as the initial image. Further discussion is provided in Appensix. C.2.

In contrast, this paper proposes Ctf-consistent estimators (Def. 4.4) that work for general causal relationships among generative factors.

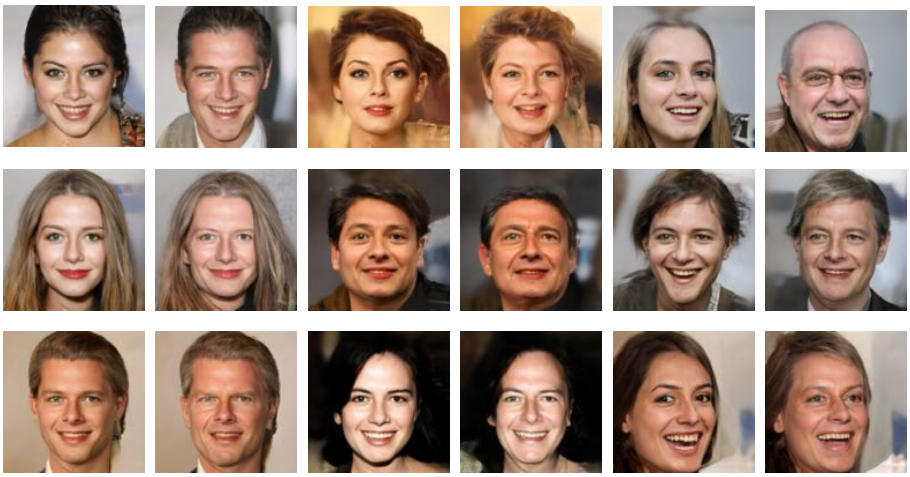
Q5. Why do you assume non-Markovianity? Is this more general than solutions that assume Markovianity?

**Answer**. For clarification, the Markovianity assumption is saying that exogenous variables in SCMs $U_j$ are independent of each other and each exogenous variable affects at most one endogenous variable. In graphical terms, there are no bi-directed edges between variables in the causal diagram.

We do not assume "non-Markovianity", but rather we do not make any assumption on Markovianity at all. Typically, Markovianity is the assumption and not the other way around. Markovianity requires that there exists no unobserved confounding, but works that do not assume Markovianity do not enforce such a requirement and are therefore more general because they work in both Markovian and non-Markovian cases. Markovianity is often assumed to simplify the problem setting, but it is unrealistic in some settings to expect that data on all potential confounding variables are collected in a study. Generative models that are enforced with Markovianity structure to proximate the true ASCM that induces non-Markovianity graphs are not guaranteed to provide Ctf-consistent estimators (see Prop. C.3).

What would the image be had the person been older? $P(\mathbf{I} = \mathbf{i}, \mathbf{I}_{Y=0} = \mathbf{i}')$

ANCM (Ours)

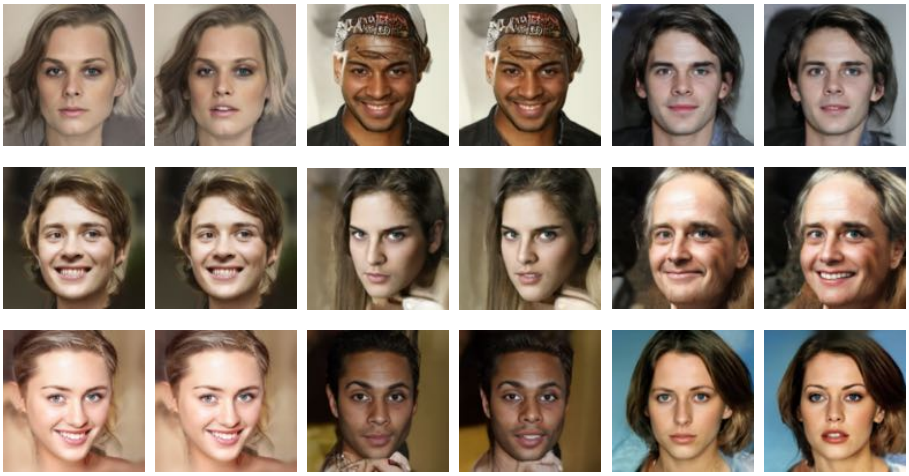


CVAE

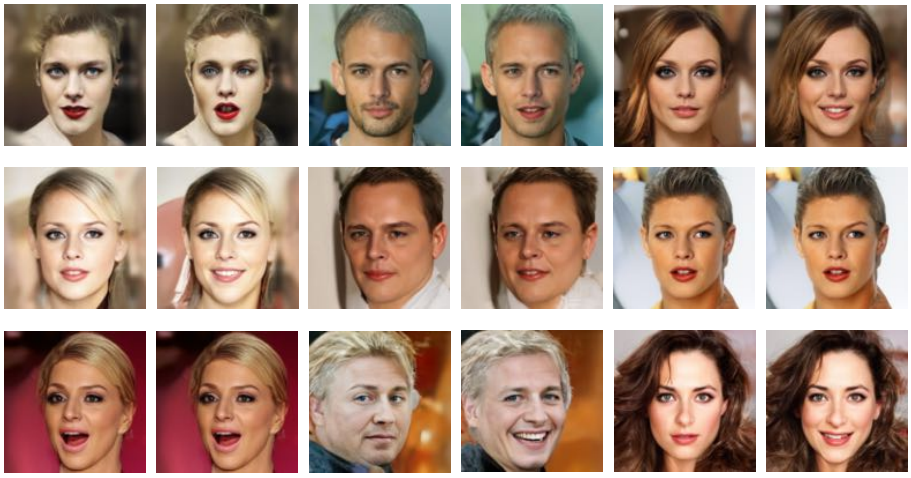
DEAR

Figure 17: Additional results of the CelebaHQ Experiment in the Smiling setting.

What would the image be had the person opened the mouth? $P(\mathbf{I} = \mathbf{i}, \mathbf{I}_{O=1} = \mathbf{i}')$



ANCM (Ours)



CVAE

DEAR

Figure 18: Additional results of the CelebaHQ Experiment in the Age setting.