Visual Representation Learning through Causal Intervention for Controllable Image Editing

Shanshan Huang[1], Haoxuan Li[2], Chunyuan Zheng[3], Lei Wang[1], Guorui Liao[1],
Zhili Gong[1], Huayi Yang[1], Li Liu[1*]
[1]School of Big Data & Software Engineering, Chongqing University
[2]Center for Data Science, Peking University [3]School of Mathematical Sciences, Peking University
{shanshanhuang, dcsliuli}@cqu.edu.cn, {hxli, cyzheng}@stu.pku.edu.cn,
{leiwangtt, guoruiliao, 202224131097, 202224131078}@stu.cqu.edu.cn

Abstract

A key challenge for controllable image editing is that visual attributes with semantic meanings are not always independent, resulting in spurious correlations in model training. However, most existing methods ignore such issues, leading to biased causal visual representation learning and unintended changes to unrelated regions or attributes in the edited images. To bridge this gap, we propose a diffusion-based causal visual representation learning framework called CIDiffuser to capture causal representations of visual attributes based on structural causal models to address the spurious correlation. Specifically, we first decompose the image representation into a high-level semantic representation for core attributes of the image and a low-level stochastic representation for other random or less structured aspects, with the former extracted by a semantic encoder and the latter derived via a stochastic encoder. We then introduce a causal effect learning module to capture the direct causal effect, that is, the difference of potential outcomes before and after intervening on the visual attributes. In addition, a diffusion-based learning strategy is designed to optimize the representation learning process. Empirical evaluations on two benchmark datasets demonstrate that our approach significantly outperforms state-of-the-art methods, enabling highly controllable image editing by modifying learned visual representations.

1. Introduction

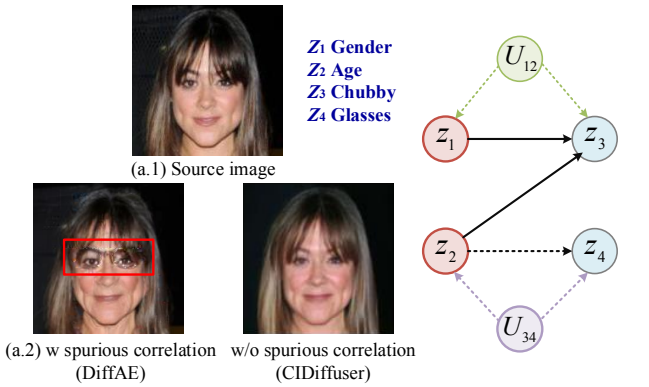
Controllable image editing aims to modify specific target attributes while keeping unrelated attributes unchanged [2]. This capability is crucial for a wide range of applications, from personalized content generation to interactive design tools [10, 12, 19, 29, 38], where fine-grained control over visual elements is necessary. Fortunately, the maturity and prevalence of deep generative models such as generative adversarial networks (GANs) [6] and variational autoencoders (VAEs) [7, 37] diffusion model [4, 8], facilitate the editing of realistic images provided in various source inputs.

Existing methods can generally be divided into two categories: condition-based image editing [10, 12, 20, 34] and representation learning-based methods [1, 5, 26, 27, 30]. The core idea of the former is to guide the editing process with the help of additional conditional information, such as text descriptions, masks, or reference images. However, acquiring such additional information is challenging and expensive. In contrast, representation learning-based methods aim to learn the latent representation of images for attribute editing, providing more precise control while reducing the reliance on external information. However, a key challenge in image editing is modification of one attribute will intentionally impact other unrelated attributes. This difficulty stems from the complex dependencies between the various attributes within an image.

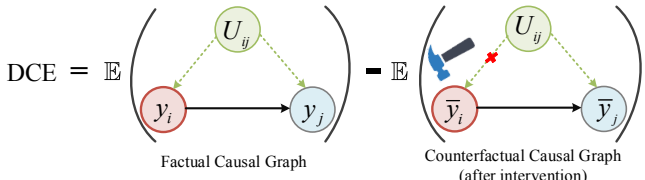
Unfortunately, many existing methods are unable to solve the above challenge. For example, in the CelebA dataset, the presence of *older faces* is usually associated with the presence of *glasses*, while *younger faces* are associated with the absence of *glasses* in the dataset, existing methods may mistakenly link *wearing glasses* with *age*. As illustrated in Fig. 1(a.2), current methods (e.g., [24]) may inadvertently add or remove *glasses* when editing *age*, as shown by the red box, even though there is no causal relationship between them.

Causal inference is an effective tool to address these issues, which is widely-used in many fields such as economy, healthcare, and e-commerce [9, 18, 35, 36]. CFI-VAE [19] introduces a causal intervention approach within the VAE framework to learn causal effects between latent rep-

*Corresponding author



(a) An example of facial image and its causal graph



(b) Calculation of the DCE

Figure 1. (a) Example of Spurious correlations in the CelebA dataset: Editing "*age*" may incorrectly influence "*glasses*". (b) Calculation of the DCE. By comparing the difference between the factual observation scenario and the counterfactual intervention scenario, the influence of confounders U_{ij} can be effectively excluded, accurately quantifying the DCE of y_i on y_j.

resentations. This model aims to improve learned representation quality by learning direct causal effects (DCE) via causal intervention. However, the representations learned by CFI-VAE lack interpretability, which means that there is no guarantee that the learned representations correspond to specific attributes in the image. Moreover, due to the inherent limitations of the VAE framework, CFI-VAE may generate incomplete latent representations because its optimization process involves a trade-off between reconstruction error and KL divergence. Such trade-off hinders the adequate capture of the causal representation, resulting in the editing process is still affected by spurious correlations.

To bridge this gap, we propose CIDiffuser, a diffusion-based framework capturing causal representations of visual attributes by adopting causal interventions on them. **First**, to fully capture image representations, we decompose the visual representation into two key components: high-level semantic representations, which handle data-generating factors and their causal relationships, and low-level stochastic representations, which capture the random or less structured aspects of the image. **Second**, inspired by previous causal effect estimation methods [31, 33, 40, 41], a structural causal model (SCM) is introduced to estimate the causal effects among these visual representations by incorporating unmeasured variables, specifically confounding biases that lead to spurious correlations, as shown in Fig. 1(a). Particularly, a causal effect learning module is introduced to remove the confounding biases by quantifying the differences measured based on DCE between the predicted outcomes before and after intervening on the attribute, as illustrated in Fig. 1(b). **Third**, we employ a diffusion model to decode both the high-level representations and low-level stochastic representations, effectively balancing the completeness of causal representations and the fidelity of the generated images. In addition, a learning strategy is devised by combining the evidence lower bound with a causal prior regularization term, a causal effect loss, and a supervised loss to optimize the representation and also to promote the alignment of learned representations with the image generating factors (i.e., visual attributes). **Finally**, extensive experiments on two public datasets show that our method outperforms existing state-of-the-art methods in capturing visual representations and controllable image editing.

2. Problem Formulation

Given a dataset $\mathcal{D}$ that consists of N images, and each image x in $\mathcal{D}$ is annotated with M labeled attributes, represented as $Y_x = \{y_{x,1}, y_{x,2}, ..., y_{x,M}\}$, such as *gender*, *age* in Fig 2. Notice that each attribute in Y_x can be either multi-values or continuous values. The goal of our model is to capture a set of causal representations $Z_x = \{z_{x,1}, z_{x,2}, ..., z_{x,M}\}$ for each label, with each causal representation in Z_x corresponding to an intervenable visual attribute in Y_x. To brief the presentation, we ignore the instance subscript. To achieve this goal, we learn a causal effect matrix $\mathbf{A}^{M\times M}$ for each image with $A_{ij} \in \mathbf{A}$ quantifying the DCE of variable z_i on z_j, i.e., the strength $z_i \rightarrow z_j$. Note that $A_{ij} = 0$ indicates the absence of a direct causal relationship from z_i to z_j and each entry A_{ij} represents the causal effect, which is not necessary to be binary. In addition, we adopt the background knowledge to pre-specify which entry is zero. Formally, we employ SCM $\mathcal{M} =< \varepsilon, z, F, P_\varepsilon >$ to encapsulate the causal relationships among the latent representations of the image dataset $\mathcal{D}$. Here $\varepsilon = \{\varepsilon_1, \varepsilon_2, ..., \varepsilon_M\}$ denotes a set of exogenous variables for the visual representations z, which is extracted from the semantic encoder $E(\cdot)$ [27] and P_ε presents its corresponding probability distribution that can be obtained via a semantic encoder. By leveraging the learned causal effect matrix $\mathbf{A}$, each causal representation of a specific visual attribute can be calculated through a corresponding SCM function $z_i = f_i(Pa(z_i), \mathbf{A}_{i,Pa(z_i)}, \varepsilon_i)$, where $Pa(z_i)$ denotes its parent set (e.g., *age, gender*), and $f_i \in F$ (i.e., a set of functions $F = \{f_1, f_2, ..., f_M\}$). Unfortunately, the estimation of $\mathbf{A}$ can be biased due to confounding bias and spurious correlations, as illustrated in Fig. 1(a).

To address this issue, we propose a causal intervention-based learning strategy to mitigate these biases. Formally,

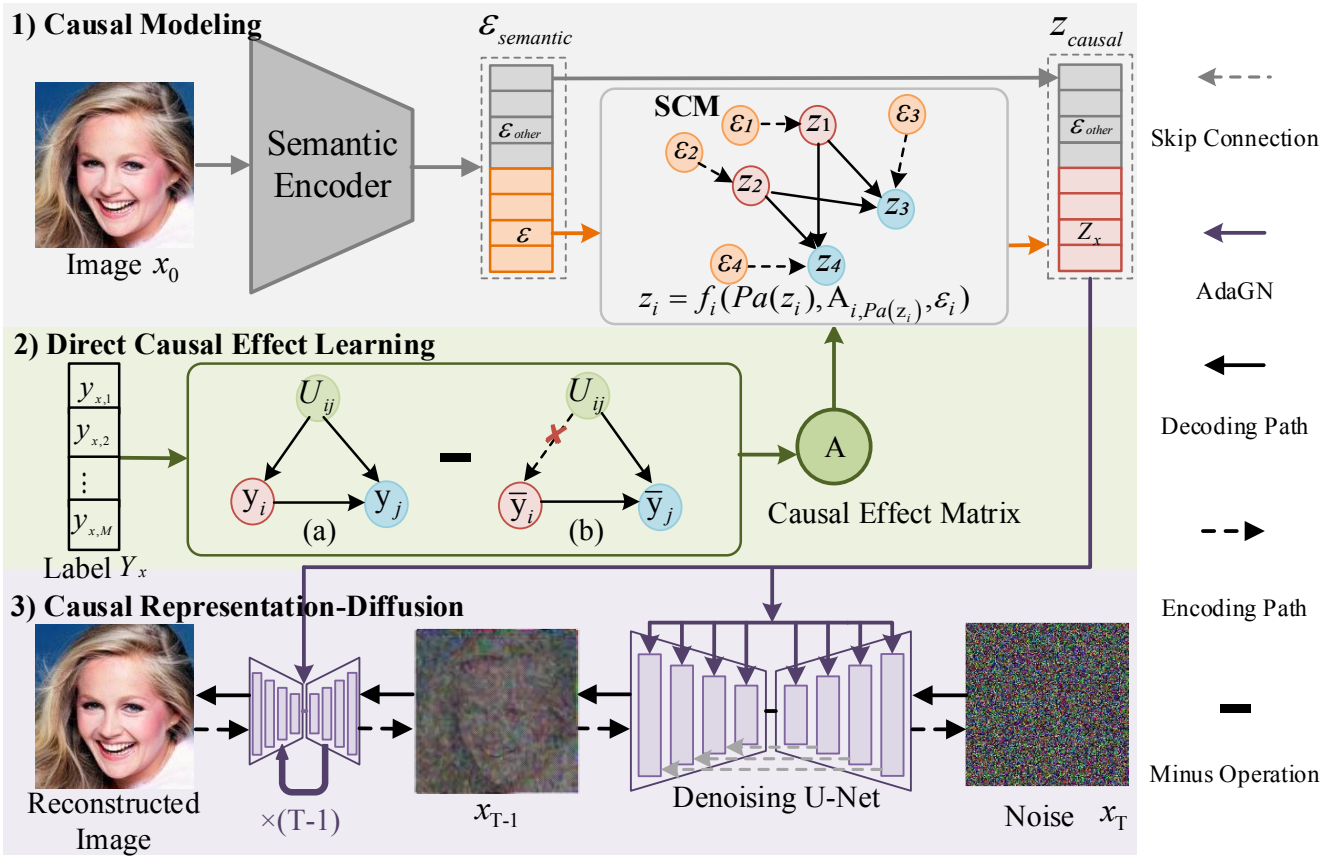


Figure 2. The framework of our CIDiffuser. The framework first encodes the input image x_0 to high-level semantic representations by a semantic encoder; Then, the direct causal effect learning module processes the semantic representations and labels to infer the causal effect matrix **A**; Next, the learned **A** is incorporated with the structural causal model (SCM) to transform these semantic representations into high-quality causal representations. These are then fused with the low-level stochastic representation derived by a stochastic encoder, and fed into the diffusion model, which undergoes a backward diffusion process to gradually guide the image transformations, ultimately reconstructing the input image.

causal intervention is defined as the *do* operation, which entails assigning a value to an attribute (e.g., *age*) directly. Then, we can calculate the DCE from attribute y_i to y_j by $DCE(y_i \rightarrow y_j) = \mathbb{E}[y_j|y_i, U_{ij}] - \mathbb{E}[y_j|do(\bar{y}_i), U_{ij}]$, as shown in fig 1(c), where $\bar{y}_i$ is the outcome value after intervention, U_{ij} denotes the confounding variables that affect both attributes, such as elderly individuals frequently wear glasses, causing both y_i (*age*) and y_j (*glasses*) to be affected by the underlying image distribution (confounding variables U_{ij}). Note that we keep the U_{ij} the same through the DCE calculation to extract the direct effect of y_i to y_j. In the facial image editing example, $y_i = 0$ means that the person in the image is *young* (0) in the dataset. $\bar{y}_i$ denotes the new value of y_i after intervention (e.g., setting $y_i = 1$ indicates the change of the *age* to *old* (1)). The difference between these two terms isolates the DCE of y_i on y_j from the effects of confounders. In this way, our method is more capable of learning high-quality causal representations for controllable image editing.

3. Methodology

Fig. 2 illustrates the whole framework of our proposed CIDiffuser. Given an image x_0 and its corresponding set of M labels $Y_x = \{y_{x,1}, y_{x,2}, ..., y_{x,M}\}$, we first adopt a semantic encoder and a stochastic encoder to extract the high-level semantic representations ϵ and low-level stochastic representations x_T, respectively. The semantic representations capture core attributes and patterns, while the stochastic representations encode random or less structured aspects of the image. Next, we define a SCM (a.k.a., causal graph) to represent the causal representation Z_{causal} as a function of learned semantic representation $\epsilon_{semantic}$. To disentangle direct causal effects from confounding biases, we introduce a direct causal effect learning module that applies causal intervention and a causal effect loss. The causal effect matrix is then integrated into the SCM to transform $\epsilon_{semantic}$ into the causal representations Z_{causal}. Finally, the causal and stochastic representations are fused and fed into the causal

decoder for image reconstruction. During the testing phase, we achieve targeted attribute editing by modifying the high-level semantic representation according to the attribute.

3.1. Semantic Encoding via Causal Modeling

To achieve a causal latent representation, we first design a semantic encoder $E(\cdot) : x_0 \rightarrow \varepsilon_{semantic}$ to map an input image x_0 to a latent representation, where $\varepsilon_{semantic} = \{\varepsilon, \varepsilon_{other}\}$ consist of two parts: the representations of the image generating factors (i.e., visual attributes of interest to user), and the representations of the other axillary factors that are necessary for image generation. We then convert the latent representation ε (also known as exogenous variables) to causal representations Z_x, which capture the causal relationships between these variables via nonlinear F in SCM. Then, we merge Z_x with ε_{other} to form the complete causal representations z_{causal}. In this formulation, each noise term $\varepsilon_i \in \varepsilon$ is the exogenous noise term for the representation z_i of the attribute in the SCM. i.e., $z_i = f((\mathbf{I} - \mathbf{A}^T)^{-1}h(\varepsilon_i))$, where $\mathbf{I}$ is the $M \times M$ identity matrix, $f(\cdot)$ and $h(\cdot)$ are transformation nonlinear functions. Note that $f(\cdot)$ is invertible, so $f^{-1}(z_i) = \mathbf{A}^T f^{-1}(z_i) + h(\varepsilon_i)$, which means that the factors z_i can be intervened in this way. Following [27], the piece-wise linear functions are used as non-linear $f(\cdot)$, which is defined as

$$f(z_i) = [w_0]_i z_i + \sum_{t=1}^{N_a} [w_t]_i (z_i - a_t)\mathbf{I}(z_i \geq a_t) + [b]_i,$$

where $w_t, t = 0, 1, ..., N_a$ and b are learnable weights and biases parameters, $a_{N_a} > ... > a_1 > a_0$ are the points of division, $\mathbf{I}(\cdot)$ denotes the indicator function.

Note that during the inference phase, controlled image editing can be achieved by intervening on specific dimensions (e.g., *age*) of the latent representations ε. Specifically, we first apply intervention operations to the latent representations ε and input the intervened representations into a nonlinear SCM to obtain causal representations z. These causal representations are then fused with stochastic representations x_t that are obtained by the diffusion process. The fused representations are then fed into a denoising diffusion implicit model (DDIM) decoder for denoising, thereby enabling controlled image editing.

3.2. Direct Causal Effect Learning Module

To address the confounding biases prevalent in the dataset, we design a direct causal effect learning (DCEL) module. This module consists of two components: a causal intervention strategy and a causal effect loss function. Specifically, we first introduce the confounding bias, and reconstruct the causal graph as illustrated in Fig. 1(a). The causal graph consists of four type nodes: confounding biases U_{ij}, causal attribute y_i, and outcome attribute y_j. Link $y_i \rightarrow y_j$ implies that there is a direct dependency between y_i and y_j. Link $y_i \leftarrow U_{ij} \rightarrow y_j$ means that confounding biases in dataset x_0 affect causal attribute y_i and outcome attribute y_j.

Therefore, we follow the paradigm of [19, 23] and exploit the causal intervention to separate the direct causal effect from the effect of confounding biases. Without loss of generality, we use a binary scenario as an example, and the proposed method can be easily extended to the continual case. Specifically, inspired by [11, 25] the causal effect can be estimated by comparing the differences of outcomes under two scenarios: the counterfactual scenario, where causal attribute y_i is intervened by set to a fixed value $\bar{y}_i$, where the influence of biases are cut off, and the factual scenario, where the causal attribute y_i remains as observed in the dataset (i.e., y_i). For those $A_{ij} \neq 0$, the casual effects from variable y_i to variable y_j, can be determined using the DCE formula: $A_{ij} = \widehat{DCE}(y_i \rightarrow y_j)$, where

$$\widehat{DCE}(y_i \rightarrow y_j) = \hat{Y}^j_{y_i}(x_0, y_k) - \hat{Y}^j_{\bar{y}_i}(x_0, y_k), k \neq i, j. \tag{1}$$

Here, $\hat{Y}^j_{y_i}(x_0, y_k)$ and $\hat{Y}^j_{\bar{y}_i}(x_0, y_k)$ represents predicted Y for the class y_j with the original image x_0 and other attribute y_k as input, and with the intervened value of y_i and $\bar{y}_i$ for class y_i, respectively. Specifically, we train a classifier C_{ij} by the causal effect loss $\mathcal{L}_d$,

$$\begin{aligned}\mathcal{L}_d = &BCE[C_{ij}((y_i, y_k), x_0), y_j] \\ &- \lambda BCE[C_{ij}((\bar{y}_i, y_k), x_0), y_j], k \neq j.\end{aligned} \tag{2}$$

Here, $\bar{y}_i$ is the outcome value after intervention, λ is a trade-off hyperparameter and the BCE is binary cross-entropy loss. Notably, for multi-value attributes, the loss function is adapted to cross-entropy loss. The first term ensures predictions are as close as possible to class y_j when the inputs of the classifier are causal attribute y_i, image x_0, and the image attributes $y_k, k \neq i, j$. In other words, the first term ensures the prediction ability for y_i to y_j. However, only minimizing the first BCE loss cannot ensure $A_{ij} \neq 0$, because the model may use only y_k and x_0 to predict y_j. Thus, we minus the second BCE loss, which changes y_i to $\bar{y}_i = 1 - y_i$. This term means that if we change the value of y_i, the prediction result will be changed, which ensures the A_{ij} is not to be zero. For the continue case, we set the $\bar{y}_i$ to zero.

When removing the negative effect of confounding bias, the learned causal effects are also influenced by class imbalance. Therefore, we improve the causal effect loss by employing influence function [3, 22] for measuring the training sample's influence on the classifier,

$$IB_{ij}(((y_i, y_k), x_0); w) = ||C_{ij}(((y_i, y_k), x_0), w) - y_j||_1 ||h||_1, \tag{3}$$

where $||\cdot||_1$ represent the 1-norm and $w = [w_1, ..., w_M]^T$ is the weight matrix of the fully connected (FC) layer,

$h = [h_1, ..., h_L]^T$ is the input of the FC layer. The meaning of this equation is the derivative of this FC layer through backward propagation. Then the causal effect loss $\mathcal{L}_d^{imb}$ can be convert to

$$\begin{aligned}\mathcal{L}_d^{imb} = \gamma_k \frac{BCE[C_{ij}((y_i, y_k), x_0), y_j]}{IB_{ij}(((y_i, y_k), x_0); w)} \\ - \lambda BCE[C_{ij}((\bar{y}_i, y_k), x_0), y_j], k \neq j.\end{aligned} \tag{4}$$

Here class-wise re-weighting term $\gamma_k = \psi n_k^{-1} / \Sigma_{k'=1}^{M} n_{k'}^{-1}$ is added to mitigate the dataset biases arising from the overall imbalanced distribution through the slow-down of the majority class loss minimization. n_k is the number of samples in the k-th class in the training dataset to measure the imbalance degree, and ψ is the hyper-parameter for an adjustment. Note that, since the influence function is derived from the loss minimization context [22], we initially use Eq. (2) for model training. After convergence, we switch to Eq. (4) to mitigate the impact of dataset biases that cause deviations in causal effect estimation.

3.3. Causal Representations-based Diffusion Model

Now, we can capture the causal representation via the learned causal effect matrix that captures the causal generation mechanism of images. Further, we utilize the causal representation-based diffusion model as stochastic encoder E_s to encode the input image x_0 into a stochastic representation x_t for generating fine image details (e.g., texture). This is achieved by $x_{t+1} = \sqrt{\bar{\alpha}_{t+1}}\hat{x}_0(x_t, t, z_{causal}) + \sqrt{1-\bar{\alpha}_{t+1}}\epsilon_\theta(x_t, t, z_{causal})$, where $\hat{x}_0(x_t, t, z_{causal}) = \sqrt{\frac{1}{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t, z_{causal}))$ is an estimate of x_0 from x_t. Here, $\bar{\alpha}_t = \Pi_{i=1}^{t}\alpha_t$ is the cumulative product of the schedule parameters α_t [24, 32], $\epsilon_\theta(x_t, t, z_{causal})$ refers to the noise prediction network, which is implemented using U-Net architecture parameterized by θ.

Next, to fuse the learned causal representations z_{causal} with stochastic representations x_t, we employ the adaptive group normalization layers (AdaGN) [4]. Specifically, the fusion is performed as follows: $\text{AdaGN}(h, t, z_{causal}) = (1 + z^s_{causal})((1 + t^s)\text{GN}(h) + t^b) + z^b_{causal}$. Here, $[z^s_{causal}, z^b_{causal}] = \text{SiLU}(\text{Linear}(z_{causal}))$, and $[t^s, t^b] = \text{SiLU}(\text{Linear}(t))$. These fused representations are then fed into a causal representations-based diffusion decoder D for image reconstruction. Following [24, 32], our decoder is a conditional DDIM that model $p_\theta(x_{t-1}|x_t, z_{causal})$ to match the inference distribution

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(\sqrt{\alpha_{t-1}}x_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \frac{x_t - \sqrt{\alpha_t}x_0}{\sqrt{1-\alpha_t}}, \sigma_t^2 I) \tag{5}$$

with the following generative process:

$$p_\theta(x_{t-1}|x_t, z_{causal}) = p(x_T)\Pi_{t=1}^{T} p_\theta(x_{t-1}|x_t, z_{causal}); \tag{6}$$

Specifically, p_θ in Eq. (6) is defined as

$$p_\theta(x_{t-1}|x_t, z_{causal}) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t, z_{causal})). \tag{7}$$

3.4. Learning Strategy

To facilitate the learning of semantic factors and ensure identifiability guarantees [13], we design a novel learning strategy for CIDiffuser based on variational inference. First, we follow [32] to develop a variational lower bound on the marginal log-likelihood of the data by applying variational inference twice,

$$\begin{aligned}&\log p_\theta(x_0) = \mathbb{E}_{q(x_1|x_0)}[\log p_\theta(x_0|x_1, z_{causal})] \\ &- \sum_{t=2}^{T} \mathbb{E}_{q(x_t|\mathbf{x}_0)}[\text{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t, z_{causal}))] \\ &- \text{KL}(q_\phi(z_{causal}|x_0)||p(z_{causal})) - \text{KL}(q(\mathbf{x}_T|x_0)||p(x_T)) \\ &:= \mathcal{L}_{diff},\end{aligned} \tag{8}$$

where $\mathcal{L}_{diff}$ represents the evidence lower bound for our CIDiffuser, $q_\phi(z_{causal}|x_0)$ is an approximate variational posterior, KL denotes the Kullback–Leibler divergence.

Then, to guarantee alignment between the underlying factors and the learned latent representations, we design a supervised loss by introducing supervision information (i.e., labels y),

$$\mathcal{L}_s = \mathbb{E}_{x_0, y}[r_s(q_\phi(z|x_0), y)], \tag{9}$$

where r_s adopts the BCE loss for binary labels and the mean squared error loss for continuous-valued labels. Additionally, the causal effect loss $\mathcal{L}_d$ is joint with $\mathcal{L}_{diff}$ and $\mathcal{L}_s$ to capture the causal representation of images, the final loss function can be expressed as

$$\mathcal{L}(E, E_s, D, S, C) = \mathcal{L}_{diff} + \kappa\mathcal{L}_s + \mu\mathcal{L}_d^{imb}, \tag{10}$$

where parameters κ, μ serve as weighting factors that are employed to balance the causal effect loss term with the supervised loss term.

4. Experiments

In this section, we conduct extensive experiments on two public datasets to evaluate the effectiveness of our method.

4.1. Experimental setting

Datasets. Evaluations are performed on two publicly available datasets: a synthetic dataset, **Pendulum** [37], and a real-world dataset, **CelebA** [21].

CelebA is a real-world dataset of facial images with 40 attributes. Following [19, 27], three datasets are constructed by selecting different subsets of image attributes, called **CelebA-smile, CelebA-age, CelebA-gender,** respectively, where **CelebA-smile** consists of six attributes: *smile*, *gender*, *narrow eye*, *mouth open*, *cheekbone*, and *chubby*, with causal relationship defined as *smile* $\rightarrow$ $\{$*narrow eye*, *mouth open*, *cheekbone*, *andchubby*$\}$, and

Table 1. Performance comparison on latent representations quality across datasets. Best results are bold, and sub-optimal results are underlined. Note that part of the results are quoted from [15].

Datasets	CelebA-smile		CelebA-age		Pendulum	
Metric	MIC	TIC	MIC	TIC	MIC	TIC
Beta-VAE	0.338	0.337	0.156	0.158	0.266	0.147
ConditionalVAE	0.788	0.661	0.898	0.787	0.938	0.805
DiffAE	0.448	0.401	0.240	0.562	0.810	0.699
Infodiff	0.413	0.384	0.594	0.556	0.742	0.719
CausalVAE	<u>0.837</u>	<u>0.716</u>	0.926	0.834	0.951	0.823
DEAR	0.526	0.530	0.347	0.343	0.329	0.306
SCM-VAE	0.751	0.689	**0.944**	<u>0.889</u>	<u>0.962</u>	0.891
CFI-VAE	0.664	0.632	0.461	0.539	0.921	**0.981**
CausalDiffAE	-	-	-	-	0.911	0.892
CIDiffuser	**0.890**	**0.869**	**0.944**	**0.903**	**0.981**	<u>0.865</u>

gender $\rightarrow$ *narrow eye*. **CelebA-age** contains six attributes: *age*, gender, *receding hairline*, *makeup*, *chubby*, and *bag under the eye*, with causal relationships given by *age* $\rightarrow$ $\{$*receding hairline*, *makeup*, *chubby*, *bag under the eye*, and *gender* $\rightarrow$ $\{$*receding hairline*, *makeup*$\}$. **CelebA-gender** focuses on five attributes: *bald*, *gender*, *mustache*, *no beard*, and *age*, with causal relationship *age* $\rightarrow$ $\{$*bald*, *mustache*, *no beard*$\}$, and *gender* $\rightarrow$ $\{$*bald*, *mustache*, *no beard*$\}$. For these three datasets, the training set consists of 162,080 samples, while the testing set comprises 40,519 samples. **Pendulum** is a synthetic dataset with 4 attributes: *pendulum angle*, *light angle*, *shadow length*, *shadow position*, where $\{$*pendulum angle*, *light angle*$\}$ $\rightarrow$ *shadow length*, $\{$*pendulum angle*, *light angle*$\}$ $\rightarrow$ *shadow position*. For this dataset, the training set includes 5,000 samples, while the testing set contains 3,000 samples. These two datasets contain unique challenges: **CelebA** exhibits significant confounding biases, like a tendency for *female* samples to be labeled as *young* (103,287 out of 156,734 samples), and *male* samples as *non-young* (30,987 out of 45,865 samples), as well as an imbalance between *chubby* and *non-chubby* samples (11,663 vs. 190,936). **Pendulum**'s confounding bias stems from the original state of the physical concept.

Evaluation Metrics. Following the previous work [16, 32, 37], We adopt two groups of metrics to evaluate the performance of CIDiffuser. On the one hand, we employ total AUROC difference (TAD), the number of attributes successfully captured (Attr), maximal information coefficient (MIC), and total information coefficient (TIC) to evaluate the disentanglement capability of learned causal representations. A higher value for these metrics indicates better disentanglement performance. On the other hand, we use Fréchet inception distance (FID), inception score (IS), and kernel inception distance (KID) to evaluate the generated images' diversity and fidelity.

4.2. Baselines

To verify the effectiveness of CIDiffuser, we compare it with the state-of-the-art methods. Generally, we adopt the baselines from two categories. Conditional VAE [28], Beta-VAE [7], DiffAE [24], PDAE [39], Infodiff [32], and DBAE [14] belong to traditional disentanglement representation learning (TRL) methods. CausalVAE [37], DEAR [27], SCM-VAE [15], CFI-VAE [19], and CausalDiffAE [17] are causal visual representation learning (CRL) methods.

4.3. Implementation Details

We implement CIDiffuser using Pytorch and the code will be released on GitHub. Following [37], the real-world datasets, CelebA are scaled to 128 $\times$ 128 resolution, while the synthetic datasets, Pendulum, are adjusted to 96 $\times$ 96 resolution. More specifically, we empirically take the ADAM as the optimizer, and the learning rate is set to $1e{-}4$. We set the batch size as 16 for model training due to computational resource constraints. Unless otherwise stated, the coefficients κ and μ are set to 0.1 and 0.5, respectively. All experiments are run on five NVIDIA GeForce RTX 4090 GPUs. The training epochs vary from dataset to dataset, with 1000 epochs for the Pendulum dataset, and 50 epochs for the CelebA-smile/age/gender datasets.

4.4. Performance Comparison

We report the overall performance comparison among the methods in Table 1 and Table 2. The visualization results are shown in Fig. 3, Fig. 4, and Fig. 5 (left). From these experimental results, we have the following observations.

First, CRL methods outperform TRL methods in disentanglement, and CausalVAE [37] achieves the most competitive performance among all the baseline methods. This is because CRL methods account for dependencies between latent factors, while TRL methods assume independence, which is often unrealistic in real-world data. *Second*, the TRL method significantly outperforms the existing CRL method in image generation quality. DiffAE [24] exhibits the best performance among all baseline methods. This is because DiffAE [24] leverages the diffusion model, which is well-suited for high-quality image generation. CRL methods, typically based on VAEs or GANs, face challenges like mode collapse or lack of generative diversity. *Third*, while DiffAE surpasses CIDiffuser in quantitative metrics, CIDiffuser generates images that better follow the image generation mechanism, as shown in Fig. 3. That is to say, interventions on specific dimensions of the latent representation (e.g., *age*, *gender*) should solely affect the outcome attributes (e.g., *beard*, *chubby*), while leaving other attributes unchanged. Conversely, modifications to the outcome attributes (e.g., *mouth open*) should not influence the attributes (e.g., *smile*, *gender*). Similar results

Table 2. Performance comparison on CelebA datasets in terms of image quality. Best in bold and sub-optimal underlined.

	TRL				CRL				Our
Methods	Beta-VAE	DiffAE	Infodiff	CausalGAN	CausalVAE	DEAR	ICM-VAE	CFI-VAE	CIDiffuser
FID ↓	169.7	**51.6**	100.7	144.1	284.3	97.2	263.3	275.7	<u>56.7</u>
IS↑	1.563	**3.669**	1.931	1.831	1.366	2.014	1.485	1.663	<u>3.053</u>
KID ↓	0.092	<u>0.041</u>	0.061	0.056	0.084	0.044	0.065	0.063	**0.027**

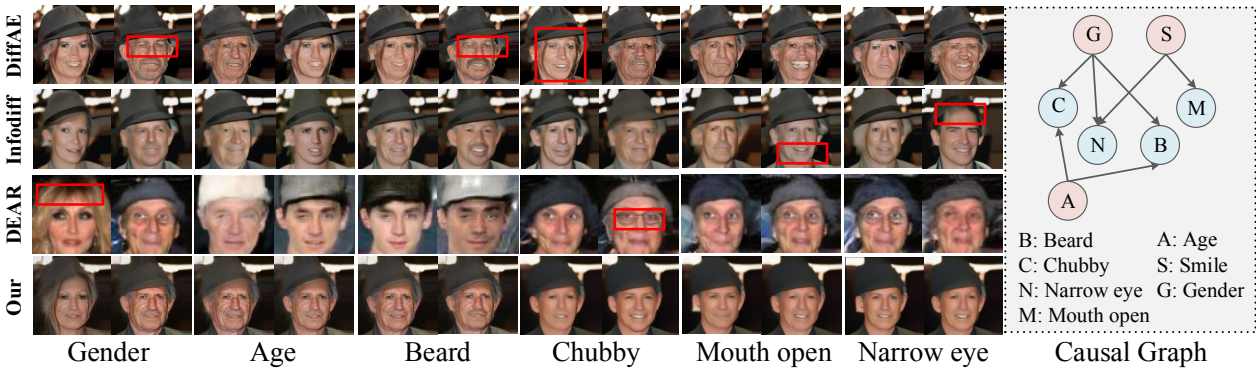


Figure 3. Comparison results on the CelebA dataset. The red boxes highlight phenomena observed in certain methods that are influenced by spurious correlations.

can be observed in the Pendulum dataset (Fig. 4). However, due to confounding biases, images generated by DiffAE [24] and Infodiff [32] often do not conform to the data generation mechanism, leading to both reverse causation issues, where the results inappropriately influence the causes, and spurious correlations, as seen in the unnecessary addition of *glasses* when editing *gender*. Overall, CIDiffuser not only excels in representation learning but also produces high-quality images, demonstrating its effectiveness.

4.5. Ablation Study

Effectiveness of Causal Modeling. To demonstrate the effectiveness of causal modeling (CM), we conducted experiments by removing it from the CIDiffuser. Table 3 demonstrates that while the CIDiffuser without the CM module performs better on the quality metrics (e.g., LQ) of latent representations, it falls short on the number of generating factors identified (i.e., Attr). This discrepancy can be attributed to the CIDiffuser model with the CM module learning a richer representation by integrating the causal structure of the data into the learning process. This integration imposes clear constraints on the latent space, enabling the model to more effectively capture causally relevant generative factors. Besides, the fidelity of images generated by CIDiffuser with the CM module is significantly higher than that of the model without the CM module. This is because the CM module ensures that the image generation process aligns with the causal mechanisms underlying physical image formation, leading to the edited image that more accurately reproduces the statistical properties and visual details of the source images.

Effectiveness of DCEL. We evaluate the performance of CIDiffuser under three different variations of the DCEL module, focusing on the presence or absence of confounding biases. Three variants are considered: one that preserves the confounding biases and class imbalance issue (i.e., without $\mathcal{L}_d$ (Eq. 2), one that suffers from the issue of class imbalance (with Eq. 2), and one that eliminates the biases (i.e., CIDiffuser with Eq. 4). Experimental results (Table 3, rows 5-7) show that while all variants capture the same number of image-generating factors, the presence of biases degrades performance in metrics like LQ, and TAD. This reveals that the confounding biases and class imbalance issue led the model to capture representations closely related to the biases rather than the essential factors relevant to the image generation process. The introduction of $\mathcal{L}_d$ (Eq. 2) eliminates the confounding biases, resulting in improvements in TAD and LQ, but remains slightly below the performance of the CIDiffuser with all bias removed. This indicates that while $\mathcal{L}_d$ mitigates the effects of false co-occurring attributes caused by confounding bias, the causal effects might still be over- or underestimated due to class imbalance issues. Further, the $\mathcal{L}_d^{imb}$ (Eq. 4) effectively refines the causal effect matrix, leading to more high-quality representations. This is evidenced by higher TAD and LQ scores, indicating successful mitigation of confounding biases and class imbalance issues. Additionally, the improvements are further supported by lower FID and KID scores, reflecting enhanced image visual quality.

Effectiveness of Supervised Loss. The impact of supervised loss on model performance is presented in Tab. 3. The results clearly indicate that the removal of supervised

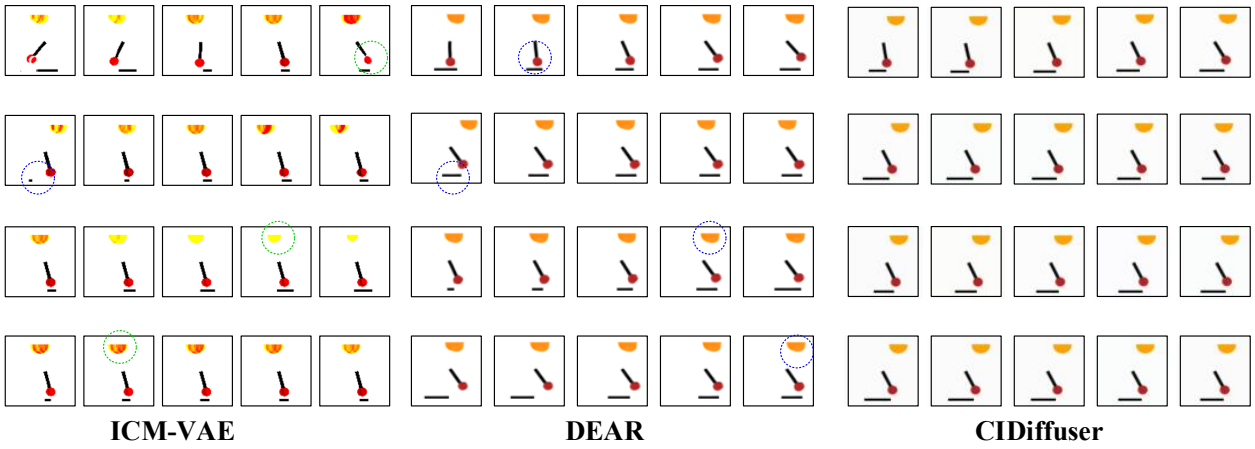


Figure 4. Performance comparison on the Pendulum dataset. Rows 1 - 4 show the results of editing the *pendulum angle*, *light position*, *shadow length*, and *shadow position*, respectively. Existing methods either compromise the quality of the generated images, as indicated by the green circles, or when editing the result attributes, it may still affect the attributes, as seen in the blue circles.

Table 3. Ablation studies on CelebA-smile. LQ is measured as AUROC for logistic regression classifiers trained on z.

Model	LQ	TAD	Attr	FID	IS	KID
w/o CM	**0.999**	0.351	4	76.7	3.036	0.047
w/o $\mathcal{L}_s$&$\mathcal{L}_d$	0.432	0.037	2	133.9	2.944	0.153
w/o $\mathcal{L}_s$	0.844	0.039	2	89.5	2.142	0.108
w/o $\mathcal{L}_d$	0.882	0.459	**6**	57.4	3.036	**0.023**
w $\mathcal{L}_d$	0.988	0.461	**6**	**56.7**	3.033	0.027
CIDiffuser	0.994	**0.481**	**6**	**56.7**	**3.053**	0.025

loss L_s leads to a significant degradation in performance, affecting both the quality of representation learning and the fidelity and diversity of generated images. This decline is primarily due to the absence of supervised loss, which hinders the alignment between the learned representations and the model's generative factors (e.g., visual attributes such as smiles, gender, and makeup). Consequently, the learned representations lack clear semantic meaning, undermining the effectiveness of disentangled representation learning. As a result, even with the incorporation of causal modeling, the generation quality remains compromised.

Impact of Causal Representations Dimension N. We investigate how varying the dimensions of causal representations affects the quality of learned representations. The experimental results, depicted in Fig. 5 (right), show a clear correlation between representation dimension and the encoder's ability to extract semantic features. As the representation dimension increases, the encoder's ability to capture the semantic features of the image is enhanced, resulting in higher fidelity during image editing. However, beyond an optimal value ($N > 64$), further increases in dimensionality degrade representation quality. This degradation occurs because higher dimensionality can introduce unwanted noise, hindering the model's ability to learn useful information. Therefore, we select 64 as the final dimension.

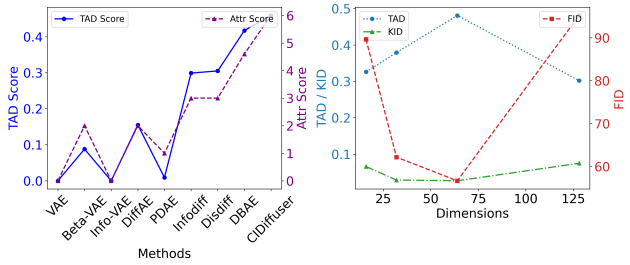


Figure 5. Performance comparison. Left: Results of representation quality of different methods on the CelebA dataset. Right: Analysis of different causal representation dimensions.

5. Conclusions

In this paper, we propose a novel causal visual representation learning framework, CIDiffuser, designed to learn high-quality causal representations for controllable image editing. Our framework incorporates a causal modeling module that ensures learned representations faithfully capture the underlying causal mechanisms of image generation. Additionally, we present a direct causal effect learning module and a novel learning strategy that effectively removes confounding biases in the dataset, promoting the capture of interpretable causal representations that are aligned with the image-generating factors (i.e., visual attributes). Extensive experiments on both synthetic and real-world datasets demonstrate the enhanced performance of our method over existing methods in representation quality and controllable image editing capabilities.

6. Acknowledgment.

This work was supported by grants from the National Key R & D Program of China (grant no. 2022YFB3303302), and the National Natural Science Foundation of China (grant nos. 62377040, 62477004, 62207007, 623B2002).

References

[1] SeungHwan An, Kyungwoo Song, and Jong-June Jeon. Causally disentangled generative variational autoencoder. In *ECAI*, 2023. 1

[2] Xi Chen, Yutong Feng, Mengting Chen, Yiyang Wang, Shilong Zhang, Yu Liu, Yujun Shen, and Hengshuang Zhao. Zero-shot image editing with reference imitation. In *NeurIPS*, 2024. 1

[3] R. Dennis Cook and Sanford Weisberg. *Residuals and Influence in Regression*. 1982. 4

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 1, 5

[5] Di Fan, Yannian Hou, and Chuanhou Gao. CF-VAE: Causal disentangled representation learning with vae and causal flows. *arXiv:2304.09010*, 2023. 1

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 2014. 1

[7] Irina Higgins, Loic Matthey, Arka Pal, Chris Burgess, Xavier Glorot, Matt Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017. 1, 6

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1

[9] Shanshan Huang, Haoxuan Li, Qingsong Li, Chunyuan Zheng, and Li Liu. Pareto invariant representation learning for multimedia recommendation. In *ACM MM*, 2023. 1

[10] Shanshan Huang, Qingsong Li, Jun Liao, Shu Wang, Li Liu, and Lian Li. Controllable image synthesis methods, applications and challenges: a comprehensive survey. *Artificial Intelligence Review*, 57(12):336, 2024. 1

[11] Shanshan Huang, Lei Wang, Jun Liao, and Li Liu. Multi-attentional causal intervention networks for medical image diagnosis. *Knowledge-Based Systems*, 299, 2024. 4

[12] Shanshan Huang, Yuanhao Wang, Zhili Gong, Jun Liao, Shu Wang, and Li Liu. Controllable image generation based on causal representation learning. *Frontiers of Information Technology & Electronic Engineering*, 25(1):135–148, 2024. 1

[13] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *AISTATS*, 2020. 5

[14] Yeongmin Kim, Kwanghyeon Lee, Minsang Park, Byeonghu Na, and Il-Chul Moon. Diffusion bridge autoencoders for unsupervised representation learning. *arXiv:2405.17111*, 2024. 6

[15] Aneesh Komanduri, Yongkai Wu, Wen Huang, Feng Chen, and Xintao Wu. Scm-VAE: Learning identifiable causal representations via structural knowledge. In *ICBD*, 2022. 6

[16] Aneesh Komanduri, Yongkai Wu, Feng Chen, and Xintao Wu. Learning causally disentangled representations via the principle of independent causal mechanisms. In *IJCAI*, 2024. 6

[17] A. Komanduri, C. Zhao, F. Chen, and X. Wu. Causal diffusion autoencoders: Toward counterfactual generation via diffusion probabilistic models. In *ECAI*, 2024. 6

[18] Meng Li and Haochen Sui. Causal recommendation via machine unlearning with a few unbiased data. In *AAAI Workshop on AICT*, 2025. 1

[19] Xiutian Li, Siqi Sun, and Rui Feng. Causal representation learning via counterfactual intervention. In *AAAI*, 2024. 1, 4, 5, 6

[20] Y. Liu, E. Sangineto, Y. Chen, L. Bao, H. Zhang, N. Sebe, B. Lepri, W. Wang, and M. De Nadai. Smoothing the disentangled latent style space for unsupervised image-to-image translation. In *CVPR*, 2021. 1

[21] Zicheng Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 5

[22] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. Influence-balanced loss for imbalanced visual classification. In *ICCV*, 2021. 4, 5

[23] Pearl and Judea. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009. 4

[24] Kittipat Preechakul, Narong Chatthee, Supasorn Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *CVPR*, 2022. 1, 5, 6, 7

[25] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *ICCV*, 2021. 4

[26] Akshay Gopinathan Reddy, L. Benin Godfrey, and Vineeth N Balasubramanian. On causally disentangled representations. *arXiv:2112.05746*, 2021. 1

[27] Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23:1–55, 2022. 1, 2, 4, 5, 6

[28] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *NeurIPS*, 2015. 6

[29] Xue Song, Jiequan Cui, Hanwang Zhang, Jingjing Chen, Richang Hong, and Yu-Gang Jiang. Doubly abductive counterfactual inference for text-based image editing. In *CVPR*, 2024. 1

[30] Rolf Suter, Dijana Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *ICML*, 2019. 1

[31] Hao Wang, Jiajun Fan, Zhichao Chen, Haoxuan Li, Weiming Liu, Tianqiao Liu, Quanyu Dai, Yichao Wang, Zhenhua Dong, and Ruiming Tang. Optimal transport for treatment effect estimation. In *NeurIPS*, 2023. 2

[32] Yingheng Wang, Yair Schiff, Aaron Gokaslan, Weishen Pan, Fei Wang, Christopher De Sa, and Volodymyr Kuleshov. Infodiffusion: Representation learning using information maximizing diffusion models. In *ICML*, 2023. 5, 6, 7

[33] Yingrong Wang, Haoxuan Li, Minqin Zhu, Anpeng Wu, Ruoxuan Xiong, Fei Wu, and Kun Kuang. Causal inference with complex treatments: A survey. *arXiv:2407.14022*, 2024. 2

[34] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *ICCV*, 2023. 1

[35] Peng Wu, Haoxuan Li, Yuhao Deng, Wenjie Hu, Quanyu Dai, Zhenhua Dong, Jie Sun, Rui Zhang, and Xiao-Hua Zhou. On the opportunity of causal learning in recommendation systems: Foundation, estimation, prediction and challenges. In *IJCAI*, 2022. Survey Track. 1

[36] Yuntian Wu, Yuntian Yang, Jiabao Sean Xiao, Chuan Zhou, Haochen Sui, and Haoxuan Li. Invariant spatiotemporal representation learning for cross-patient seizure classification. In *NeurIPS Workshop on NeuroAI*, 2024. 1

[37] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. CausalVAE: Disentangled representation learning via neural structural causal models. In *CVPR*, 2021. 1, 5, 6

[38] Min Zhang, Haoxuan Li, Fei Wu, and Kun Kuang. Metacoco: A new few-shot classification benchmark with spurious correlation. In *ICLR*, 2024. 1

[39] Zijian Zhang, Zhou Zhao, and Zhijie Lin. Unsupervised representation learning from pre-trained diffusion probabilistic models. *NeurIPS*, 2022. 6

[40] Chuan Zhou, Yaxuan Li, Chunyuan Zheng, Haiteng Zhang, Min Zhang, Haoxuan Li, and Mingming Gong. A two-stage pretraining-finetuning framework for treatment effect estimation with unmeasured confounding. In *ACM SIGKDD*, 2025. 2

[41] Minqin Zhu, Anpeng Wu, Haoxuan Li, Ruoxuan Xiong, Bo Li, Xiaoqing Yang, Xuan Qin, Peng Zhen, Jiecheng Guo, Fei Wu, et al. Contrastive balancing representation learning for heterogeneous dose-response curves estimation. In *AAAI*, 2024. 2