CAD-VAE: Leveraging Correlation-Aware Latents for Comprehensive Fair Disentanglement

Chenrui Ma[1], **Xi Xiao**[2], **Tianyang Wang**[2], **Xiao Wang**[3], **Yanning Shen**[1*]

[1]University of California, Irvine, Irvine, CA 92697, USA
[2]University of Alabama at Birmingham, Birmingham, AL 35294, USA
[3]Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA

Abstract

While deep generative models have significantly advanced representation learning, they may inherit or amplify biases and fairness issues by encoding sensitive attributes alongside predictive features. Enforcing strict independence in disentanglement is often unrealistic when target and sensitive factors are naturally correlated. To address this challenge, we propose **CAD-VAE** (**C**orrelation-**A**ware **D**isentangled **VAE**), which introduces a correlated latent code to capture the information shared between the target and sensitive attributes. Given this correlated latent, our method effectively separates overlapping factors without extra domain knowledge by directly minimizing the conditional mutual information between target and sensitive codes. A relevance-driven optimization strategy refines the correlated code by efficiently capturing essential correlated features and eliminating redundancy. Extensive experiments on benchmark datasets demonstrate that CAD-VAE produces fairer representations, realistic counterfactuals, and improved fairness-aware image editing.

Code — https://github.com/merry7cherry/CAD-VAE
Extended version — https://arxiv.org/abs/2503.07938

Introduction

Deep generative models have achieved remarkable success in capturing complex data distributions for applications ranging from image synthesis (Bai et al. 2024; Ma et al. 2025c,b) to video generation (Montanaro et al. 2024). In particular, variational autoencoders (VAEs)(Higgins et al. 2017; Kingma and Welling 2022; Mathieu et al. 2019; Ma et al. 2025a) have provided a principled approach to representation learning, where data are encoded into compact latent variables that effectively capture meaningful factors of variation. However, while these latent representations have enabled impressive performance in numerous tasks, concerns about fairness have emerged, as models can inadvertently learn and amplify biases present in training data(Jang and Wang 2024).

Such fairness issues arise when the target label and sensitive label become entangled due to societal or dataset biases (Lahoti et al. 2020; Liu et al. 2021). To address these problems, existing methods commonly fall into two categories. Invariant learning techniques aim to remove sensitive attributes from the learned representation, often via adversarial training or additional regularization (Lahoti et al. 2020; Roy and Boddeti 2019). By contrast, disentanglement approaches encourage the model to partition its latent space into separate codes for target and sensitive information, seeking statistical independence among them (Creager et al. 2019; Liu, Sun, and Zhao 2023). Although these solutions have made progress, they typically assume minimal correlation between target and sensitive factors or enforce strict separation via mutual information penalties (Chen et al. 2018). However, multiple works (Jang and Wang 2024; Park et al. 2020) have demonstrated that achieving fully fair disentanglement is fundamentally impossible under realistic conditions. First, many datasets contain unwanted correlations between the target label and sensitive attributes due to societal bias, making it infeasible to preserve all predictive cues while completely discarding sensitive information (Dressel and Farid 2018). Second, certain features inherently influence both target and sensitive attributes, so perfectly partitioning features into disjoint latent spaces is unachievable without compromising prediction accuracy (Kohavi 1996). In these circumstances, any attempt at full disentanglement faces an inevitable trade-off between fairness and utility.

A natural way to handle this correlation is to explicitly model how target and sensitive attributes overlap. For instance, some methods rely on causal graphs to separate task-relevant features and capture their relationships with sensitive variables (Kim et al. 2021; Sánchez-Martin, Rateike, and Valera 2022; Zhu et al. 2023; Hwa et al. 2024). However, constructing such graphs requires extensive domain knowledge, which is often challenging to acquire in real-world scenarios.

Motivated by the limitations of existing methods, an correlated latent code is introduced to capture the shared information between target and sensitive attributes. Our approach advances existing methods (Jang and Wang 2024; Park et al. 2020) by a directly minimizing conditional mutual information mechanism to achieve disentanglement and an explicit relevance learning strategy to learn the correlated latent code efficiently and properly, as summarized follow:

1. We propose a novel correlation-aware representation learning framework that directly minimizes the conditional mutual information between target and sensitive property,

*Corresponding author: yannings@uci.edu.

conditioned on the correlated latent code, effectively addressing the conflict between predictive objectives and disentanglement.

2. We introduce an explicit relevance-driven optimization strategy that precisely regulates the correlated latent code, ensuring it captures only the essential shared information without extra domain knowledge.
3. We validate our approach through comprehensive experiments on multiple benchmark datasets, demonstrating its superiority in achieving correlation-aware disentanglement, enhancing fair prediction performance, and improving both counterfactual generation and fairness-aware image editing, as well as its broad applicability in the context of Vision-Language Models (VLM).

Related Work

Fair Disentanglement Learning

Fair disentanglement methods aim to separate representations into target-related and sensitive-related latent codes rather than directly removing sensitive information (Liu, Sun, and Zhao 2023; Madras et al. 2018; Xu et al. 2018; Wang et al. 2024). Early works such as β-VAE (Higgins et al. 2017) and FactorVAE (Kim and Mnih 2018) introduced mechanisms for semantic decomposition of latent factors, with FactorVAE promoting independence across dimensions by reducing total correlation. Building on this foundation, FairFactorVAE (Liu, Sun, and Zhao 2023) further restricts sensitive leakage within the disentanglement process.

Subsequent studies refine these ideas by emphasizing flexible or guided decomposition. FFVAE (Creager et al. 2019) adapts latent structures to better isolate sensitive attributes, while GVAE (Ding et al. 2020) employs adversarial constraints to suppress unwanted information. Other strategies incorporate structural priors such as orthogonality in ODVAE (Sarhan et al. 2020) or distance-covariance minimization in FairDisCo (Liu et al. 2022). These approaches collectively illustrate progress in disentanglement, yet also reveal the challenge of fully separating target and sensitive information when these factors are inherently correlated.

Correlation-Aware Learning

Despite advancements in disentanglement, perfect independence between latent codes is difficult to achieve due to natural correlations between sensitive and target attributes (Mehrabi et al. 2021; Jang and Wang 2024). For example, facial attributes in CelebA (Liu et al. 2015), such as "mustache," correlate with both gender and attractiveness, complicating clean separation. Correlation-aware learning frameworks seek to address this by leveraging causal graphs to categorize latent variables according to their relationships with sensitive attributes (Kim et al. 2021; Sánchez-Martin, Rateike, and Valera 2022; Zhu et al. 2023; Hwa et al. 2024). However, causal-graph construction requires strong domain knowledge, and inaccurate assumptions may hinder independence (Jang and Wang 2024). FADES (Jang and Wang 2024) mitigates this dependency by grouping samples across attributes to approximate conditional mutual information and capture shared sensitive-relevant structure. While effective in some contexts, this indirect method may still allow leakage and offers limited control over relevance allocation. These limitations motivate approaches that directly optimize conditional independence with explicit guidance for balancing sensitive-relevant information.

Counterfactual Fairness

Counterfactual fairness (CF) evaluates whether predictions remain stable when sensitive attributes are hypothetically altered (Kusner et al. 2017). Causal inference is widely used to generate counterfactual instances (Zhou et al. 2024; Jung et al. 2025; Zhu et al. 2023; Chiappa 2019; Wu, Zhang, and Wu 2019), enabling comparisons between factual and hypothetical outcomes. Graph-based CF models (Kim et al. 2021; Li et al. 2025) rely on predefined causal structures to produce realistic counterfactuals, but these structures demand precise domain knowledge; inaccurate models may lead to implausible counterfactuals, such as depicting a female subject with a mustache. Although CF offers strong theoretical grounding, its reliance on domain expertise limits practical deployment. To overcome this challenge, our approach introduces a correlated latent code with an explicit relevance-learning mechanism, allowing the model to autonomously learn attribute relationships and enhance counterfactual fairness without external causal assumptions.

Preliminary

Conditional Independence and Mutual Information

Proposition 1 (Conditional Independence)**.** *Let A, B, and C be random variables. We say that A is **conditionally independent** of B given C, denoted $A \perp B \mid C$, if and only if their conditional joint probability distribution factorizes as follows:*

$$p(A, B \mid C) = p(A \mid C)p(B \mid C). \tag{1}$$

Directly measuring the degree of conditional independence by computing the divergence between the two sides of Eq. (1) is often intractable in practice, especially in the context of deep learning models where the underlying distributions are complex and high-dimensional. Instead, a common and more tractable approach is to use an information-theoretic surrogate measure.

Definition 1 (Conditional Mutual Information)**.** *The **Conditional Mutual Information** (CMI) between two random variables A and B given a third random variable C measures the expected amount of information that A and B share, conditioned on C. It is defined as the expected Kullback-Leibler (KL) divergence between the conditional joint distribution and the product of the conditional marginal distributions:*

$$I(A; B \mid C) = \mathbb{E}_{p(C)}\Big[D_{\mathrm{KL}}\Big(p(A, B \mid C = c) \Big\| p(A \mid C = c)p(B \mid C = c)\Big)\Big]. \tag{2}$$

This can be expressed over the entire distribution of C as:

$$I(A; B \mid C) = \int D_{\mathrm{KL}}\Big(p(A, B \mid C) \Big\| p(A \mid C)p(B \mid C)\Big)\, dp(C). \tag{3}$$

CMI provides a principled way to measure conditional dependence due to its fundamental properties.

Lemma 1 (Properties of CMI). *Conditional mutual information is non-negative, i.e.,* $I(A;B \mid C) \geq 0$. *Furthermore,* $I(A;B \mid C) = 0$ *if and only if* A *and* B *are conditionally independent given* C ($A \perp B \mid C$). *See Appendix 1 for a detailed proof.*

Lemma 2 (Symmetry of CMI). *Conditional mutual information is symmetric in its primary arguments:*

$$I(A;B \mid C) = I(B;A \mid C). \tag{4}$$

Variational Autoencoder

A Variational Autoencoder (VAE) (Kingma and Welling 2022) is a generative model that learns a latent representation of data. It uses an **encoder** network, $q_\phi(z \mid x)$, to map an input sample x to a latent distribution, and a **decoder** network, $p_\theta(x \mid z)$, to reconstruct the input from a latent sample z.

The model is trained by minimizing the negative Evidence Lower Bound (ELBO):

$$\mathcal{L}_{\text{VAE}}(\theta,\phi) = \underbrace{\mathbb{E}_{q_\phi(z|x)}\big[-\log p_\theta(x \mid z)\big]}_{\text{Reconstruction Loss}} + \underbrace{\text{KL}\Big(q_\phi(z \mid x) \,\Big\|\, p_\theta(z)\Big)}_{\text{KL Divergence}}, \tag{5}$$

where the first term measures reconstruction accuracy and the second term is a regularizer that pushes the learned latent distribution $q_\phi(z \mid x)$ towards a prior $p_\theta(z)$, which is typically a standard Gaussian $\mathcal{N}(0, I)$.

Total Correlation Loss

To enforce statistical independence among latent variables, **FactorVAE** (Kim and Mnih 2018) introduces a penalty on the **Total Correlation (TC)**. TC is the Kullback-Leibler (KL) divergence between the aggregate posterior, $q(z)$, and the product of its marginals, $\prod_j q(z_j)$:

$$L_{TC} = \text{KL}\left(q(z) \middle\| \prod_j q(z_j)\right) \tag{6}$$

As this term is intractable to compute directly, it is approximated using a discriminator, D, which is trained to distinguish between samples from $q(z)$ and samples from the product of marginals (approximated by permuting dimensions across a batch). The encoder, in turn, is trained to minimize the following adversarial loss, thereby fooling the discriminator and reducing the TC:

$$L_{TC} \approx \mathbb{E}_{q(z)}\left[\log \frac{D(z)}{1 - D(z)}\right] \tag{7}$$

Method

We first present the problem definition, model components, and architecture, which serve as crucial foundations for the subsequent sections.

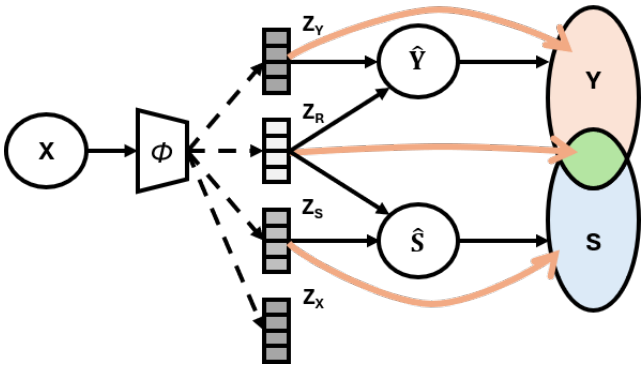


Figure 1: **Illustration of the data flow.** The orange lines connect the information in the observed space and their corresponding latent codes.

CAD-VAE

Let $\mathcal{D} = \{(x_i, y_i, s_i)\}_{i=1}^N$ denote a dataset consisting of triplets, where x_i denotes an input sample (e.g., an image), y_i is the label of x_i corresponding to target property Y, and s_i is the label corresponding to the sensitive property S. The value range of Y and S is $\mathcal{Y}$ and $\mathcal{S}$, respectively, i.e., $y \in \mathcal{Y}$ and $s \in \mathcal{S}$.

As discussed in Introduction, correlated information between the target attribute Y and the sensitive attribute S is pervasive in disentanglement learning. To address this, we introduce an additional latent code z_R to explicitly model this correlated information.

The goal is to learn a latent representation that factorizes the information relevant to Y, the information relevant to S, the shared information between Y and S, and the background or irrelevant factors. As defined below:

- z_X: captures task-irrelevant information.
- z_Y: encodes the information strongly correlated with Y.
- z_S: encodes the information strongly correlated with S.
- z_R: represents the *shared* information between Y and S.

Hence, for a single observation, the corresponding latent variable set is $z := (z_X, z_Y, z_S, z_R)$.

The latent code z_R isolates the overlapping information between Y and S, which allows the primary latent codes z_Y and z_S to remain free of unwanted correlations while preserving the model's predictive power (Creager et al. 2019; Kim et al. 2021; Jang and Wang 2024). From a causal perspective as illustrated in Figure 1, if Y and S are conditionally independent given z_R, then z_R acts as their common cause, thereby promoting the independence of z_Y and z_S.

To learn such latent code, we employ the Variational Autoencoder (VAE) framework as our backbone. The model is trained to minimize the negative ELBO, $\mathcal{L}_{\text{VAE}}$, as defined in Eq. (5).

In addition, we introduce four classifiers to enforce different constraints, including:

- Enforcing z_Y and z_S to capture sufficient information ensures that attributes Y and S can be recovered correspondingly in alignment with z_R.

- Eliminating information leakage (see in subsequent section)
- Encouraging z_R encapsulate only the correlated information $Y \cap S$ (see in subsequent section)

Here, we first present the training method and loss function of each classifier.

- $f_y(z_Y, z_R)$ is a classifier that predicts $\hat{y}$ from (z_Y, z_R);
- $f_s(z_S, z_R)$ is a classifier that predicts $\hat{s}$ from (z_S, z_R);
- $f_{y\_op}(z_S)$ is an *opponent* classifier that attempts to predict $\hat{y}$ from z_S;
- $f_{s\_op}(z_Y)$ is an *opponent* classifier that attempts to predict $\hat{s}$ from z_Y.

Let ω_y, ω_s, ω_{y_op}, and ω_{s_op} denote the parameters of these four classifiers, respectively.

We define

$$\min_{\phi,\omega_y}\Big[\mathcal{L}_y(\omega_y,\phi)\Big]=\mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[-\log f_y\big(\hat{y}\mid z_Y,z_R\big)\Big], \quad (8)$$

where z_Y and z_R are sampled from the encoder $q_\phi(z \mid x)$: $(z_Y, z_R) \sim q_\phi(z \mid x)$. The parameters ω_y and the encoder parameters ϕ are jointly updated to reduce the cross-entropy in (8), ensuring that (z_Y, z_R) carry sufficient information about Y. Similarly, the classifier $f_s(z_S, z_R)$ predicts s:

$$\min_{\phi,\omega_s}\Big[\mathcal{L}_s(\omega_s,\phi)\Big]=\mathbb{E}_{(x,s)\sim\mathcal{D}}\Big[-\log f_s\big(\hat{s}\mid z_S,z_R\big)\Big]. \quad (9)$$

To measure the information leakage, we introduce:

$$\min_{\omega_{y\_op}}\Big[\mathcal{L}_{y\_op}(\omega_{y\_op};\phi)\Big]=\mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[-\log f_{y\_op}\big(\hat{y}\mid z_S\big)\Big], \quad (10)$$

where $z_S \sim q_\phi(z \mid x)$ is produced by the frozen encoder i.e ϕ is *not* updated during the minimization of (10); this network is trained to detect any Y-relevant information that may unintentionally exist in z_S. Analogously, the classifier $f_{s\_op}(z_Y)$ aims to predict s given z_Y:

$$\min_{\omega_{s\_op}}\Big[\mathcal{L}_{s\_op}(\omega_{s\_op};\phi)\Big]=\mathbb{E}_{(x,s)\sim\mathcal{D}}\Big[-\log f_{s\_op}\big(\hat{s}\mid z_Y\big)\Big]. \quad (11)$$

Likewise, ϕ is fixed, and only ω_{s_op} is updated when minimizing (11).

To achieve correlation-aware disentanglement learning, we propose directly minimizing the conditional mutual information between z_Y and z_S with respect to their corresponding opposite attributes S and Y, conditioned on z_R. This approach is complemented by an explicit relevance learning strategy that constrains z_R to effectively capture shared information between Y and S while avoiding redundant information. Detailed explanations of these strategies are provided in following section.

Conditional Independence for Disentanglement

Our fairness objective is to achieve independence between z_Y and z_S by enforcing conditional independence between their respective predictions, $\hat{Y}$ and $\hat{S}$, given the shared latent code z_R. The predictions are generated by classifiers: $\hat{Y} = f_y(z_Y, z_R)$ and $\hat{S} = f_s(z_S, z_R)$. This objective is formally expressed as $\hat{Y} \perp \hat{S} \mid z_R$. Following **Proposition 1**, this conditional independence is equivalent to the factorization of the conditional joint distribution:

$$p_\theta(\hat{Y},\hat{S}\mid z_R)=p_\theta(\hat{Y}\mid z_R)p_\theta(\hat{S}\mid z_R). \quad (12)$$

As noted in the Preliminary section, directly minimizing the divergence between the distributions in Eq. (12) is generally intractable. We therefore adopt an information-theoretic approach and minimize the Conditional Mutual Information (CMI), $I_\phi(\hat{Y};\hat{S}\mid z_R)$, as a tractable surrogate objective.

From **Definition 1**, the CMI is the expected KL divergence over z_R:

$$I_\phi(\hat{Y};\hat{S}\,|\,z_R)= \int D_{\mathrm{KL}}\Big(p_\theta(\hat{Y},\hat{S}\,|\,z_R)\big\|p_\theta(\hat{Y}\,|\,z_R)p_\theta(\hat{S}\,|\,z_R)\Big)dP_{z_R}. \quad (13)$$

According to **Lemma 1**, minimizing $I_\phi(\hat{Y};\hat{S}\mid z_R)$ to zero is equivalent to enforcing the conditional independence defined in Eq. (12). Furthermore, leveraging the symmetry property from **Lemma 2**, we note that $I_\phi(\hat{Y};\hat{S}\mid z_R) = I_\phi(\hat{S};\hat{Y}\mid z_R)$. Thus, minimizing this CMI ensures that any undesired dependence between $\hat{Y}$ and $\hat{S}$ not explained by z_R is removed.

Direct Minimization of Conditional Mutual Information

While FADES (Jang and Wang 2024) minimizing CMI (13) through approximation $I_\phi(\hat{Y};S\mid z_R)$:

$$\min_\phi\Big[I_\phi(\hat{Y};S\mid z_R)\Big]=\min_\phi\Big[H_\phi(\hat{Y}\mid z_R)-H_\phi(\hat{Y}\mid S,z_R)\Big]$$

by reducing CMI through ground truth-based sample grouping, its reliance on batch-level sampling introduces instability.

In contrast, our method directly minimizes CMI (13) via a principled information-theoretic approach, providing a more robust and stable disentanglement process by avoiding sampling variance and reducing dependency on batch-specific dynamics. To achieve CI as (12), we propose directly minimizing:

$$\min_\phi\Big[I_\phi(\hat{Y};\hat{S}\mid z_R)+I_\phi(\hat{S};\hat{Y}\mid z_R)\Big], \quad (14)$$

where:

$$I_\phi(\hat{Y};\hat{S}\mid z_R)=H_\phi(\hat{Y}\mid z_R)-H_\phi(\hat{Y}\mid\hat{S},z_R), \quad (15)$$

$H_\phi(*\mid *)$ stands for the conditional entropy. Incorporating $\hat{S}=f_s(z_S,z_R)$, we have:

$$\begin{aligned} I_\phi(\hat{Y};\hat{S}\mid z_R) &= H_\phi(\hat{Y}\mid z_R)-H_\phi(\hat{Y}\mid\hat{S},z_R)\\ &= H_\phi(\hat{Y}\mid z_R)-H_\phi(\hat{Y}\mid f_s(z_S,z_R),z_R)\\ &= H_\phi(\hat{Y}\mid z_R)-H_\phi(\hat{Y}\mid z_S,z_R)\\ &= I_\phi(\hat{Y};z_S\mid z_R), \end{aligned} \quad (16)$$

from the same transformation (see detailed derivation in *Appendix 2*):

$$\begin{aligned} I_\phi(\hat{S};\hat{Y}\mid z_R) &= I_\phi(\hat{S};z_Y\mid z_R)\\ &= H_\phi(\hat{S}\mid z_R)-H_\phi(\hat{S}\mid z_Y,z_R). \end{aligned} \quad (17)$$

Therefore, with the introduction of the correlated latent code z_R that captures all relevant information between $\hat{Y}$ and $\hat{S}$:

$$\begin{aligned}\min_{\phi}\Big[I_\phi(\hat{Y};\hat{S}\mid z_R)+I_\phi(\hat{S};\hat{Y}\mid z_R)\Big]\\ \equiv\min_{\phi}\Big[I_\phi(\hat{Y};z_S\mid z_R)+I_\phi(\hat{S};z_Y\mid z_R)\Big].\end{aligned}\tag{18}$$

For the minimization of $I_\phi(\hat{Y};z_S\mid z_R)$, as shown in (16) , since z_R is given as a condition, we can consider this CMI formula as a function where the independent variable is z_S and the dependent variable is $\hat{Y}$, as shown as:

$$\mathcal{L}_{\hat{Y}}(z_S)=H_\phi(\hat{Y}\mid z_R)-H_\phi(\hat{Y}\mid z_S,z_R),\tag{19}$$

where z_R is determined here, $H_\phi(\hat{Y}\mid z_R)$ is a constant, henceforth we need to minimize $-H_\phi(\hat{Y}\mid z_S,z_R)$. Empirically, we directly minimize the lower bound of it: $-H_\phi(\hat{Y}\mid z_S)$, since: $-H_\phi(\hat{Y}\mid z_S,z_R)\geq -H_\phi(\hat{Y}\mid z_S)$. Symmetrically, the minimization of $I_\phi(\hat{S};z_Y\mid z_R)$ shown in (17) is the same concept, see *Appendix 2* for detailed derivation. In this optimization process, z_R is responsible for containing any correlation information between target attribute Y and sensitive attribute S.

After these simplifications, we introduce the CMI loss to minimize (14):

$$\min_{\phi}\Big[\mathcal{L}_{\text{CMI}}(\omega_{y\_op};\omega_{s\_op};\phi)\Big]=-(H_\phi(\hat{Y}\mid z_S)+H_\phi(\hat{S}\mid z_Y)),\tag{20}$$

where only update encoder parameters ϕ. Utilizing opponent classifier $f_{y\_op}(z_S)$, the entropy term calculation are shown as below:

$$\begin{aligned}H_\phi(\hat{Y}\mid z_S)&=\mathbb{E}_{q_\phi(z_S|x)}\Big[-\sum_{\hat{y}\in\mathcal{Y}}p_\theta(\hat{y}\mid z_S)\log\,p_\theta(\hat{y}\mid z_S)\Big]\\&=\frac{1}{|B|}\sum_{i=1}^{|B|}\sum_{\hat{y}\in\mathcal{Y}}\Big[-p_\theta\big(\hat{y}\mid z_S^{(i)}\big)\log p_\theta\big(\hat{y}\mid z_S^{(i)}\big)\Big],\end{aligned}\tag{21}$$

where $p_\theta\big(\hat{y}\mid z_S^{(i)}\big)=f_{y\_op}\big(z_S^{(i)}\big)$. Here, $z_S^{(i)}$ denotes the z_S sample from the i-th element in a mini-batch of size $|B|$; the distribution $q_\phi(z_S\mid x)$ is given by the encoder. Similar calculation to $H_\phi(\hat{S}\mid z_Y)$, see *Appendix 3* for completed calculation formula. During this optimization process, the opponent classifier parameters ω_{y_op} and ω_{s_op} are frozen.

Learning Relevance Between Target And Sensitive Information

To encourage z_R capture and only capture the shared information relevant to target property and sensitive property, as well as z_Y, z_S capture main information of Y and S attributes respectively, we propose to maximize the conditional mutual information as Learning Relevance Information loss:

$$\min_{\phi}\Big[\mathcal{L}_{\text{LRI}}(\omega_y;\omega_s;\phi)\Big]=-(I_\phi(\hat{Y};Y\mid z_R)+I_\phi(\hat{S};S\mid z_R)),\tag{22}$$

where:

$$I_\phi(\hat{Y};Y\mid z_R)=H_\phi(\hat{Y}\mid z_R)-H_\phi(\hat{Y}\mid Y,z_R),\tag{23}$$

$$I_\phi(\hat{S};S\mid z_R)=H_\phi(\hat{S}\mid z_R)-H_\phi(\hat{S}\mid S,z_R),\tag{24}$$

$H_\phi(*\mid *)$ stands for the conditional entropy. Maximizing $H_\phi(\hat{Y}\mid z_R)$, $H_\phi(\hat{S}\mid z_R)$ avoid z_R capture all information of Y or S solely, which will lead to the Information Bottleneck phenomenon(Jang and Wang 2024; Creager et al. 2019; Kim and Mnih 2018) i.e z_R capture all the information about target attribute Y and sensitive attribute S: $Y\cup S$, degenerating disentanglement performance. On the other hand, minimizing $H_\phi(\hat{Y}\mid Y,z_R)$ and $H_\phi(\hat{S}\mid S,z_R)$ enforce z_R determine $\hat{Y}$ or $\hat{S}$ only within each Y or S subgroup, so that encourage z_R capture information both relevant to Y and S: $Y\cap S$.

In contrast to FADES(Jang and Wang 2024), which exhibits a conflict between the disentanglement term and the regularization term, our method achieves orthogonality between CMI (20) and LRI (22), ensuring fair disentanglement while preserving robust representations. See Appendix 5 for details.

For entropy calculation of $H_\phi(\hat{Y}\mid z_R)$, We approximate $p_\theta(\hat{y}\mid z_R)$ by marginalizing over z_Y:

$$\begin{aligned}p_\theta(\hat{y}\mid z_R^{(k)})&=\mathbb{E}_{p(x)}\Big[\mathbb{E}_{q_\phi(z_Y|x)}\Big[p_\theta(\hat{y}\mid z_Y,z_R^{(k)})\Big]\Big]\\&\approx\frac{1}{|B|}\sum_{i=1}^{|B|}p_\theta(\hat{y}\mid z_Y^{(i)},z_R^{(k)}),\end{aligned}\tag{25}$$

where $p_\theta(\hat{y}\mid z_Y^{(i)},z_R^{(k)})=f_y\big(z_Y^{(i)},z_R^{(k)}\big)$, then

$$\begin{aligned}H_\phi(\hat{Y}\mid z_R)&=\mathbb{E}_{q_\phi(z_R|x)}\left[-\sum_{\hat{y}\in\mathcal{Y}}p_\theta(\hat{y}\mid z_R)\log p_\theta(\hat{y}\mid z_R)\right]\\&=\frac{1}{|B|}\sum_{i=1}^{|B|}\sum_{\hat{y}\in\mathcal{Y}}\Big[-p_\theta(\hat{y}\mid z_R^{(i)})\log p_\theta(\hat{y}\mid z_R^{(i)})\Big].\end{aligned}\tag{26}$$

As for the calculation of conditional entropy term $H_\phi(\hat{Y}\mid Y,z_R)$, we regard known condition Y as attribute to grouping samples in a mini-batch of size $|B|$, and calculate the entropy term by marginalizing over z_Y within each group, $p_\theta(\hat{y}\mid z_R,y)$ can be computed for $z_R^{(k)}$ sampled from an instance $x^{(k)}\in B_y$ as:

$$\begin{aligned}p_\theta(\hat{y}|z_R^{(k)},y)&=\mathbb{E}_{p(x|Y=y)}\Big[\mathbb{E}_{q_\phi(z_Y|x)}\Big[p_\theta(\hat{y}\mid z_Y,z_R^{(k)})\Big]\Big]\\&\approx\frac{1}{|B_y|}\sum_{i=1}^{|B_y|}p_\theta(\hat{y}|z_Y^{(i)},z_R^{(k)}),\end{aligned}\tag{27}$$

where B_y denotes a subset of the batch with $Y=y$. Then the conditional entropy can be computed as:

$$\begin{aligned}&H_\phi(\hat{Y}\mid Y,z_R)=\\&\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{E}_{q_\phi(z_R|x)}\left[-\sum_{\hat{y}\in\mathcal{Y}}p_\theta(\hat{y}\,|\,z_R,y)\log p_\theta(\hat{y}\,|\,z_R,y)\right]\right]\\&=\frac{1}{|B|}\sum_{y\in\mathcal{Y}}\sum_{i=1}^{|B_y|}\sum_{\hat{y}\in\mathcal{Y}}\Big[-p_\theta(\hat{y}|z_R^{(i)},y)\log p_\theta(\hat{y}|z_R^{(i)},y)\Big].\end{aligned}\tag{28}$$

Methods	Downstream Classification Performance											
	CelebA (Liu et al. 2015)			UTKFace (Zhang, Song, and Qi 2017)			Dogs and Cats (Parkhi et al. 2012)			Color bias MNIST (Kim et al. 2019)		
	Acc ↑	EOD ↓	DP ↓	Acc ↑	EOD ↓	DP ↓	Acc ↑	EOD ↓	DP ↓	Acc ↑	EOD ↓	DP ↓
FADES (Jang and Wang 2024) [CVPR'24]	0.918	0.034	0.135	0.812	0.059	0.139	0.769	0.058	0.086	0.973	0.094	0.160
GVAE (Ding et al. 2020) [CVPR'20]	0.919	0.047	0.131	0.819	0.204	0.197	0.748	0.064	0.131	0.961	0.109	0.176
FFVAE (Creager et al. 2019) [PMLR'19]	0.892	0.076	0.072	0.766	0.269	0.201	0.729	0.059	0.110	0.952	0.081	**0.092**
ODVAE (Sarhan et al. 2020) [ECCV'20]	0.886	0.039	0.103	0.736	0.165	0.210	0.689	0.051	**0.038**	0.957	0.247	0.162
FairDisCo (Liu et al. 2022) [KDD'22]	0.839	0.074	**0.051**	0.766	0.266	0.200	0.680	0.115	0.111	0.949	0.129	0.136
FairFactorVAE (Liu, Sun, and Zhao 2023)	0.914	0.055	0.136	0.720	0.096	**0.134**	0.707	0.055	0.110	0.957	0.096	0.128
CAD-VAE (Ours)	**0.939**	**0.021**	0.065	**0.828**	**0.045**	0.137	**0.781**	**0.048**	0.069	**0.984**	**0.076**	0.108

Table 1: Evaluation of downstream classification on various datasets from learned representation. Best in **bold**, second in red.

The calculation of $H_\phi(\hat{S} \mid z_R)$ and $H_\phi(\hat{S} \mid S, z_R)$ are similar to $H_\phi(\hat{Y} \mid z_R)$, $H_\phi(\hat{Y} \mid Y, z_R)$ respectively, see *Appendix 3* for completed calculation formula. Plugging these estimates(26)(28) back into (22), shared feature between Y and S will be learned in z_R while getting rid of Information Bottleneck phenomenon. Note that the classifier parameters ω_y and ω_s remain frozen when optimizing (22).

Final Objective Function

To integrate the above components into a coherent training framework, we employ the two-step optimization strategy defined in (29) and (30).

$$\min_{\theta,\phi,\omega_y,\omega_s} \Big[\mathcal{L}_{\text{VAE}}(\theta,\phi) + \big(\mathcal{L}_y(\omega_y,\phi) + \mathcal{L}_s(\omega_s,\phi)\big)\Big] + \min_{\phi} \Big[\lambda_{\text{CMI}}\mathcal{L}_{\text{CMI}}(\omega_{y\_op},\omega_{s\_op};\phi) + \mathcal{L}_{\text{TC}}(\phi) + \lambda_{\text{LRI}}\,\mathcal{L}_{\text{LRI}}(\omega_y,\omega_s;\phi)\Big] \tag{29}$$

Specifically, in (29), we jointly update $(\theta, \phi, \omega_y, \omega_s)$ by minimizing the VAE loss (5) alongside the main classification losses (8) and (9), which together reformulate the ELBO. We further include the CMI loss (20) to reduce unwanted information leakage, the LRI loss (22) to capture shared patterns in z_R, and the TC penalty (6) to promote factorization among the latent codes (z_Y, z_R, z_S).

$$\min_{\omega_{y\_op},\omega_{s\_op}} \Big[\mathcal{L}_{y\_op}(\omega_{y\_op};\phi) + \mathcal{L}_{s\_op}(\omega_{s\_op};\phi)\Big] \tag{30}$$

In parallel, the second procedure (30) optimizes $(\omega_{y\_op}, \omega_{s\_op})$ by minimizing the opponent classification losses (10) and (11) while holding ϕ fixed.

The hyperparameters $\lambda_{CMI}, \lambda_{LRI} > 0$ control the relative importance of these terms, ensuring each network component learns its designated function while enforcing minimal information leakage, preserving shared information in z_R and maintaining the salient factors for Y and S in z_Y and z_S respectively. See hypermeter analysis in *Appendix 6*.

Experiment

To ensure a rigorous and comprehensive evaluation, we conduct experiments comparing our proposed method with a diverse set of state-of-the-art approaches across multiple categories of learning paradigms in various tasks. Specifically, we include FairFactorVAE (Liu, Sun, and Zhao 2023), FairDisCo (Liu et al. 2022), FFVAE (Creager et al. 2019), GVAE (Ding et al. 2020), ODVAE (Sarhan et al. 2020) and FADES (Jang and Wang 2024), shown in Section 2.1. As for traditional correlation-aware learning that is discussed in Section 2.2, since it require additional annotated data to build causal graph, we except them in our experiment.

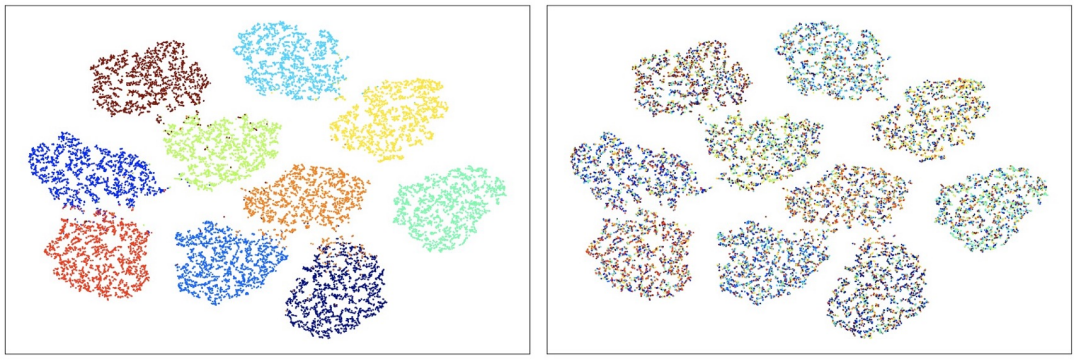
Metric	Ours	FADES	GVAE	FFVAE	ODVAE	FairFactorVAE
Accuracy ↑	**0.867**	0.782	0.771	0.744	0.721	0.807
EOD ↓	**0.141**	0.174	0.244	0.190	0.210	0.195
DP ↓	**0.167**	0.201	0.265	0.213	0.262	0.221

Table 2: Fair Classification on 95% Color Bias MNIST.

Fair Classification

The objective of fair classification is to achieve a balance between minimizing fairness violations and maintaining high predictive performance. To evaluate the effectiveness of our proposed method, we conduct experiments on a diverse set of benchmark fairness datasets. For facial attribute classification tasks, we utilize the CelebA (Liu et al. 2015) and UTKFace (Zhang, Song, and Qi 2017) datasets. Following prior works (Wang et al. 2022; Xu et al. 2020; Zeng et al. 2022; Jang and Wang 2024), we set the CelebA classification task to predict the "Smiling" attribute, while for UTKFace, the objective is to classify whether a person depicted in the image is over 35 years old, with gender serving as the sensitive attribute. Additionally, the Dogs and Cats dataset (Parkhi et al. 2012) is used to distinguish between dogs and cats, with fur color as the sensitive attribute. Furthermore, we assess fair classification performance using the Colored MNIST dataset (Kim, Lee, and Choo 2021; Kim et al. 2019; Nam et al. 2020), which incorporates a controlled color bias in the standard MNIST dataset to simulate spurious correlations. To assess fairness violations, we use standard metrics including Demographic Parity (DP) (Barocas and Selbst 2016) and Equalized Odds (EOD) (Hardt, Price, and Srebro 2016). See detailed experimental setup in *Appendix 6*. The result of fair classification can be seen in Table 1. Across all evaluated datasets, our method consistently achieves state-of-the-art classification accuracy and fairness, validating its effectiveness in robust disentanglement by preserving high-quality target-related information while minimizing sensitive attribute leakage.

To assess the robustness and effectiveness of these methods, we further conduct classification experiments under an extreme imbalance bias setting, which is commonly encountered in practical applications. In the MNIST experiment, we set the color bias rate to 95% to simulate a strong correlation between the target attribute and the sensitive attribute. The results, shown in Table 2, demonstrate that our method outperforms existing approaches in both disentanglement ability and robust representation preservation.



(a) Digit (b) Color

Figure 2: **t-SNE visualization of the target code from the test set for our method**. Left subfigure is colored by Digit; right subfigure is colored by Color.

t-SNE Visualization

To better understand the distribution and disentanglement of the learned representation, we present a t-SNE visualization analysis of the target latent code of each method. The visualizations are derived from experiments in Fair Classification, where the model is trained on a biased color MNIST dataset and tested on an unbiased color MNIST dataset. Each figure consists of two subfigures: the left subfigure is colored according to the Digit attribute (target attribute), and the right subfigure is colored according to the Color attribute (sensitive attribute). Clear and distinct clustering in the left subfigure indicates that the model has learned a robust and discriminative representation of the target attribute, thereby enhancing its recognizability. Conversely, if the right subfigure exhibits discernible color clusters, it suggests a correlation between the target and sensitive attributes, indicating weaker disentanglement performance. A uniform color distribution in the right subfigure, however, confirms that the sensitive information has been effectively filtered out.

This visualization in Figure 2 demonstrates that our proposed method effectively disentangles the learned representation. The target attribute (Digit) exhibits distinct, well-separated clusters with clear classification boundaries and a pure distribution, while the sensitive attribute (Color) is uniformly distributed and unrecognizable. This confirms that our method achieves superior separation of the target attribute without introducing unwanted bias from the sensitive attribute.

Fair Counterfactual Generation

We evaluate our approach on the CelebA dataset (Liu et al. 2015), a widely-used benchmark for facial attribute manipulation. We select *Smiling* as the target label Y and *Gender* as the sensitive attribute S. In our experiments, we substitute specific latent code of source images with reference images, including z_X, z_Y, z_S, and $[z_S, z_R]$. Figure 3 illustrates the generated counterfactuals, with the first row showing source and reference images and subsequent rows demonstrating the effects of substituting each latent code. The experimental results demonstrate the effectiveness of our method in generating fair counterfactuals. As shown in Figure 3, substituting $[z_S, z_R]$ (Row 5) leads to a natural adaptation of sensitive-relevant features without domain-specific knowledge. For instance, the model automatically adds makeup to female images and a mustache to male images, highlighting the semantic alignment of z_R with both the target and sensitive attributes. Compared to substituting only z_S, our approach achieves more interpretable translation, ensuring that fairness is maintained throughout the counterfactual generation. More fair counterfactual generation experiment results can be seen in *Appendix 7*.

To quantitatively assess the quality of the generated counterfactuals, we compare evaluation metrics between the direct reconstruction of the input image and the reconstructions obtained by randomly permuting z_Y and z_S within the evaluation set. Specifically, we use the FID (Heusel et al. 2017; Jang and Wang 2024) to assess reconstruction fidelity and the Inception Score (IS) (Chong and Forsyth 2020) to evaluate semantic and perceptual quality. Lower ΔFID values indicate minimal distortion and higher translation quality, while lower ΔIS values suggest that semantic and perceptual attributes are well preserved. Detailed experimental settings are provided in *Appendix 6*.

	CAD-VAE	FADES	GVAE	FFVAE	ODVAE	FairFactorVAE
$\Delta FID \downarrow$	**1.072**	1.167	3.710	1.409	14.647	6.239
$\Delta IS \downarrow$	**1.214**	2.379	3.148	3.829	6.113	5.378

Table 3: FID and IS difference between original reconstruction and perturbed target/sensitive codes' reconstruction.

Quantitative analysis in Table 3 further validates our approach. Our method achieves both lower ΔFID and ΔIS compared to other fair representation learning methods, demonstrating that our fair counterfactual generation approach renders counterfactuals with superior image quality and minimal distortion.

Fair Fine-Grained Image Editing

With the introduction of the correlated latent code z_R, fair fine-grained image editing—as a fundamental concept in counterfactual fairness—can be naturally achieved by align-

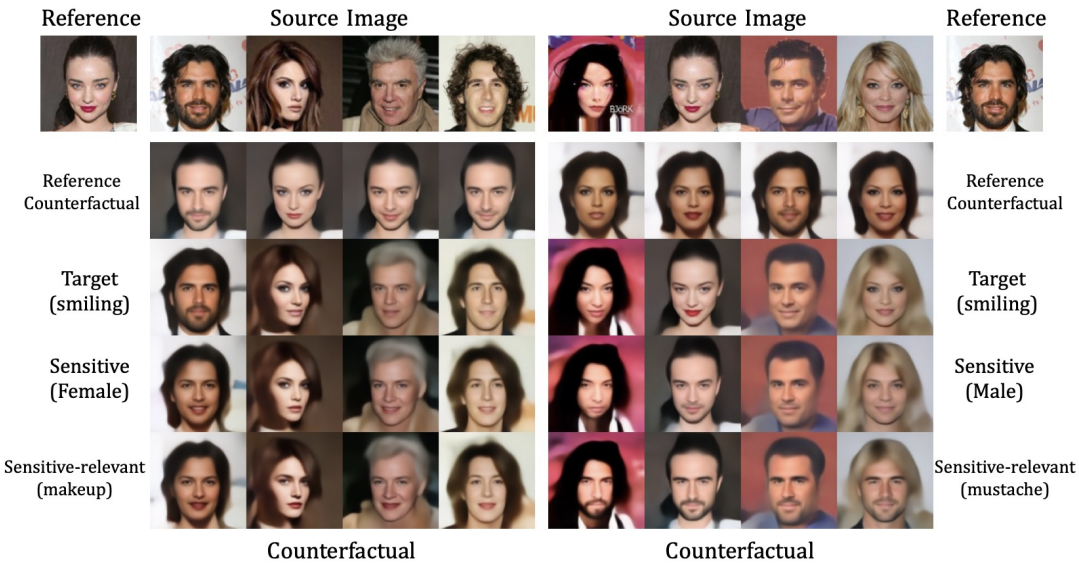


Figure 3: **Examples of Fair Counterfactual Generation**. Zoom in to check. The first row shows the source and reference images. Rows 2–5 display counterfactuals obtained by replacing latent subspaces z_X, z_Y, z_S, and $[z_S, z_R]$, respectively. Notably, the replacement with $[z_S, z_R]$ (row 5) naturally adapts sensitive features for different sensitive attributes without domain knowledge (mustache for men and makeup for women).

ing latent codes from different samples. We use linear interpolation to synthesize a latent code: $z' = (1 - \lambda)z_1 + \lambda z_2$, where z_1 is the latent code from the source image and z_2 is the corresponding code from the reference image. The synthesized z' replaces z_1, enabling a gradual transfer from the source to the reference latent code.

In our experiments, following the setup in the previous section where *Smiling* is the target label Y and *Gender* is the sensitive attribute S, we generate interpolated latent codes between source and reference images. In the blue-framed subfigure of Figure 4, images are generated by interpolating z_Y and z_S: the horizontal axis shows the transition of z_S from the source to the reference image, while the vertical axis shows the corresponding change in z_Y. During this interpolation, z_R and z_X remain unchanged, with the interpolation parameters for both z_Y and z_S set to $\lambda \in \{0, 0.33, 0.66, 1\}$.

Similarly, in the red-framed subfigure, images are generated by interpolating z_Y and z_R. Here, the horizontal axis corresponds to the transition of z_R from the source to the reference image, and the vertical axis corresponds to z_Y. Note that the source image's z_S is fully replaced by that of the reference image, and z_X remains constant. Initially, the interpolation parameters for z_R are set to $\lambda \in \{0.5, 1\}$, and when combined with the final column of the blue-framed subfigure, the range is extended to $\lambda \in \{0, 0.5, 1\}$.

Figure 4 demonstrates a smooth transformation of each attribute, with modifications in one latent code minimally affecting the others, a key characteristic of effective disentanglement. Specifically, as the correlated latent code z_R captures sensitive relevant information, we can explicitly control these properties: in the left subfigure, we gradually introduce makeup (such as enhanced lipstick and eyeshadow), while in the right subfigure, we progressively add a mustache. More fair fine-grained image editing experiment results can be seen in *Appendix 7*.

Similarly, we measure ΔFID and ΔIS to quantitatively assess the quality of fine-grained image editing. Unlike the evaluation setup in the previous section, we compute the differences in evaluation metrics between the direct reconstruction of the input image and the reconstructions obtained through latent code traversals for each λ combination. Detailed experimental settings are provided in *Appendix 6*. Table 4 summarizes the comparison results, showing that our method exhibits both lower ΔFID and ΔIS values compared to other fair generation methods. These results indicate that our fine-grained image editing approach not only ensures smoother attribute transformations and superior image fidelity, but also allows for more precise control of task-relevant features.

Fair Text-to-Image Editing

To further validate the capability and explore the applicability of our method, we integrated it as an adaptor on top of a pre-trained, frozen CLIP image encoder (Radford et al. 2021; Xiao et al. 2025c,a,b) and trained it on Facet dataset (Gustafson et al. 2023) to enhance fairness in vision-language tasks. Table 5 presents the experimental results. These results demonstrate that our approach significantly improves fairness without compromising performance compared to the linear probing baseline (ERM), underscoring its potential for a range of vision-language tasks with fairness considerations.

Furthermore, we applied our method in StyleCLIP (Patashnik et al. 2021) as a fair discriminator to address inherent fairness issues, such as career-gender biases, which persist even when an identity preservation loss is employed. As illustrated in Figure 5, StyleCLIP (Patashnik et al. 2021) exhibits

Figure 4: **Examples of Fair Fine-Grained Image Editing**. Zoom in to check. The leftmost column shows the source and reference images. The blue-framed section displays images generated by interpolating z_Y and z_S (with z_R and z_X fixed), where the horizontal axis varies z_S and the vertical axis varies z_Y. The red-framed section illustrates images produced by interpolating z_Y and z_R (with z_S fully replaced by the reference and z_X constant). Modification in one latent code minimally affecting others, harness z_R to edit sensitive relevant feature(makeup or mustache).

a bias by correlating the role of "dancer" with a specific gender. In contrast, our method effectively mitigates this bias while maintaining the efficacy of attribute modification. See *Appendix 7* for details.

	CAD-VAE	FADES	GVAE	FFVAE	ODVAE	FairFactorVAE
$\Delta FID \downarrow$	**1.642**	2.362	4.023	2.789	15.893	7.120
$\Delta IS \downarrow$	**1.849**	2.919	4.848	5.292	6.890	5.767

Table 4: FID and IS difference between original reconstruction and traversed target/sensitive codes' reconstruction.

Method	**Top-1 Acc. (%)**			**Top-3 Acc. (%)**		
	WG ↑	Avg ↑	Gap ↓	WG ↑	Avg ↑	Gap ↓
Zero-shot	2.79	53.45	50.66	15.31	76.79	61.48
Linear prob	1.17	65.46	64.29	1.79	85.34	83.55
CAD-VAE	**69.97**	**70.54**	**0.57**	**85.36**	**85.95**	**0.59**

Table 5: **Performance of CLIP(ViTB/32) on Facet dataset.** WG: Worst Group, Gap: Difference between WG and Avg.

Conclusion

Our method aims to solve fairness concerns in representation learning and deep generative models. By introducing a correlated latent code that captures shared information, sensitive information leakage can be eliminated directly and efficiently without conflicting with the prediction objective, which is a core issue in disentanglement, by minimizing the conditional mutual information between target latent code and sensitive latent code. Parallel with our explicit relevance learning strategy imposed on the correlated latent code, it is encouraged to capture the essential shared information that cannot be perfectly separated without additional domain knowledge. Various benchmark tasks further demonstrate the robustness and wide applicability of our method.

(a) Original (b) StyleCLIP (c) StyleCLIP+

Figure 5: **Style transfer using StyleCLIP and the CAD-VAE extension.** This example transforms the (a) into "a dancer with long blonde hair." "StyleCLIP+" means StyleCLIP + CAD-VAE.

Acknowledgments

This work was supported in part by the National Science Foundation under awards ECCS-2412484, ECCS-2442964 and GEO CI-2425748. This manuscript was co-authored by Oak Ridge National Laboratory (ORNL), operated by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. Any subjective views or opinions expressed in this paper do not necessarily represent those of the U.S. Department of Energy or the United States Government.

References

Bai, J.; Ye, T.; Chow, W.; Song, E.; Chen, Q.-G.; Li, X.; Dong, Z.; Zhu, L.; and Shuicheng, Y. 2024. Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis. In *The Thirteenth International Conference on Learning Representations*.

Barocas, S.; and Selbst, A. D. 2016. Big data's disparate impact. *Calif. L. Rev.*, 104: 671.

Chen, R. T.; Li, X.; Grosse, R. B.; and Duvenaud, D. K. 2018. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31.

Chiappa, S. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 7801–7808.

Chong, M. J.; and Forsyth, D. 2020. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6070–6079.

Creager, E.; Madras, D.; Jacobsen, J.-H.; Weis, M.; Swersky, K.; Pitassi, T.; and Zemel, R. 2019. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, 1436–1445. PMLR.

Ding, Z.; Xu, Y.; Xu, W.; Parmar, G.; Yang, Y.; Welling, M.; and Tu, Z. 2020. Guided variational autoencoder for disentanglement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7920–7929.

Dressel, J.; and Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1): eaao5580.

Gustafson, L.; Rolland, C.; Ravi, N.; Duval, Q.; Adcock, A.; Fu, C.-Y.; Hall, M.; and Ross, C. 2023. FACET: Fairness in Computer Vision Evaluation Benchmark. arXiv:2309.00035.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

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

Hwa, J.; Zhao, Q.; Lahiri, A.; Masood, A.; Salimi, B.; and Adeli, E. 2024. Enforcing Conditional Independence for Fair Representation Learning and Causal Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 103–112.

Jang, T.; and Wang, X. 2024. FADES: Fair Disentanglement with Sensitive Relevance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12067–12076.

Jung, S.; Yu, S.; Chun, S.; and Moon, T. 2025. Do Counterfactually Fair Image Classifiers Satisfy Group Fairness?–A Theoretical and Empirical Study. *Advances in Neural Information Processing Systems*, 37: 56041–56053.

Kim, B.; Kim, H.; Kim, K.; Kim, S.; and Kim, J. 2019. Learning Not to Learn: Training Deep Neural Networks With Biased Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kim, E.; Lee, J.; and Choo, J. 2021. Biaswap: Removing dataset bias with bias-tailored swapping augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14992–15001.

Kim, H.; and Mnih, A. 2018. Disentangling by factorising. In *International conference on machine learning*, 2649–2658. PMLR.

Kim, H.; Shin, S.; Jang, J.; Song, K.; Joo, W.; Kang, W.; and Moon, I.-C. 2021. Counterfactual fairness with disentangled causal effect variational autoencoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 8128–8136.

Kingma, D. P.; and Welling, M. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114.

Kohavi, R. 1996. Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, 202–207. AAAI Press.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

Lahoti, P.; Beutel, A.; Chen, J.; Lee, K.; Prost, F.; Thain, N.; Wang, X.; and Chi, E. H. 2020. Fairness without Demographics through Adversarially Reweighted Learning. arXiv:2006.13114.

Li, H.; Liu, Y.; Geng, Z.; and Zhang, K. 2025. A Local Method for Satisfying Interventional Fairness with Partially Known Causal Graphs. *Advances in Neural Information Processing Systems*, 37: 135415–135436.

Liu, E. Z.; Haghgoo, B.; Chen, A. S.; Raghunathan, A.; Koh, P. W.; Sagawa, S.; Liang, P.; and Finn, C. 2021. Just Train Twice: Improving Group Robustness without Training Group Information. arXiv:2107.09044.

Liu, J.; Li, Z.; Yao, Y.; Xu, F.; Ma, X.; Xu, M.; and Tong, H. 2022. Fair representation learning: An alternative to mutual information. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1088–1097.

Liu, S.; Sun, S.; and Zhao, J. 2023. Fair transfer learning with factor variational auto-encoder. *Neural Processing Letters*, 55(3): 2049–2061.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 3730–3738.

Ma, C.; Xiao, X.; Wang, T.; and Shen, Y. 2025a. Beyond Editing Pairs: Fine-Grained Instructional Image Editing via Multi-Scale Learnable Regions. *arXiv preprint arXiv:2505.19352*.

Ma, C.; Xiao, X.; Wang, T.; Wang, X.; and Shen, Y. 2025b. Learning Straight Flows: Variational Flow Matching for Efficient Generation. *arXiv preprint*.

Ma, C.; Xiao, X.; Wang, T.; Wang, X.; and Shen, Y. 2025c. Stochastic Interpolants via Conditional Dependent Coupling. *arXiv preprint arXiv:2509.23122*.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, 3384–3393. PMLR.

Mathieu, E.; Rainforth, T.; Siddharth, N.; and Teh, Y. W. 2019. Disentangling Disentanglement in Variational Autoencoders. arXiv:1812.02833.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Montanaro, A.; Aira, L. S.; Aiello, E.; Valsesia, D.; and Magli, E. 2024. MotionCraft: Physics-Based Zero-Shot Video Generation. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Nam, J.; Cha, H.; Ahn, S.; Lee, J.; and Shin, J. 2020. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33: 20673–20684.

Park, S.; Kim, D.; Hwang, S.; and Byun, H. 2020. README: REpresentation learning by fairness-Aware Disentangling MEthod. arXiv:2007.03775.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, 3498–3505. IEEE.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2085–2094.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Roy, P. C.; and Boddeti, V. N. 2019. Mitigating Information Leakage in Image Representations: A Maximum Entropy Approach. arXiv:1904.05514.

Sánchez-Martin, P.; Rateike, M.; and Valera, I. 2022. Vaca: Designing variational graph autoencoders for causal queries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 8159–8168.

Sarhan, M. H.; Navab, N.; Eslami, A.; and Albarqouni, S. 2020. Fairness by learning orthogonal disentangled representations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, 746–761. Springer.

Wang, X.; Chen, H.; Tang, S.; Wu, Z.; and Zhu, W. 2024. Disentangled Representation Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 9677–9696.

Wang, Z.; Dong, X.; Xue, H.; Zhang, Z.; Chiu, W.; Wei, T.; and Ren, K. 2022. Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10379–10388.

Wu, Y.; Zhang, L.; and Wu, X. 2019. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the twenty-eighth international joint conference on Artificial Intelligence*.

Xiao, X.; Zhang, Y.; Li, X.; Wang, T.; Wang, X.; Wei, Y.; Hamm, J.; and Xu, M. 2025a. Visual Instance-aware Prompt Tuning. In *Proceedings of the 33rd ACM International Conference on Multimedia*.

Xiao, X.; Zhang, Y.; Li, Y.; Li, X.; Wang, T.; Hamm, J.; Wang, X.; and Xu, M. 2025b. Visual Variational Autoencoder Prompt Tuning. *arXiv preprint arXiv:2503.17650*.

Xiao, X.; Zhang, Y.; Zhao, L.; Liu, Y.; Liao, X.; Mai, Z.; Li, X.; Wang, X.; Xu, H.; Hamm, J.; et al. 2025c. Prompt-based Adaptation in Large-scale Vision Models: A Survey. *arXiv preprint arXiv:2510.13219*.

Xu, D.; Yuan, S.; Zhang, L.; and Wu, X. 2018. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE international conference on big data (big data)*, 570–575. IEEE.

Xu, T.; White, J.; Kalkan, S.; and Gunes, H. 2020. Investigating bias and fairness in facial expression recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, 506–523. Springer.

Zeng, H.; Yue, Z.; Shang, L.; Zhang, Y.; and Wang, D. 2022. Boosting demographic fairness of face attribute classifiers via latent adversarial representations. In *2022 IEEE International Conference on Big Data (Big Data)*, 1588–1593. IEEE.

Zhang, Z.; Song, Y.; and Qi, H. 2017. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5810–5818.

Zhou, Z.; Liu, T.; Bai, R.; Gao, J.; Kocaoglu, M.; and Inouye, D. I. 2024. Counterfactual Fairness by Combining Factual and Counterfactual Predictions. *arXiv preprint arXiv:2409.01977*.

Zhu, H.; Dai, E.; Liu, H.; and Wang, S. 2023. Learning fair models without sensitive attributes: A generative approach. *Neurocomputing*, 561: 126841.