Towards Fairness with Limited Demographics via Disentangled Learning

Zichong Wang[1] , **Anqi Wu**[1] , **Nuno Moniz**[2] , **Shu Hu**[3] ,
Bart Knijnenburg[4] , **Xingquan Zhu**[5] and **Wenbin Zhang**[1*]
[1]Florida International University, FL, USA
[2]University of Notre Dame, IN, USA
[3]Purdue University, IN, USA
[4]Clemson University, SC, USA
[5]Florida Atlantic University, FL, USA

Abstract

Fairness in artificial intelligence has garnered increasing attention due to concerns about discriminatory AI-based decision-making, prompting the development of numerous mitigation approaches. However, most existing methods assume that demographic information is readily available, which may not align with real-world scenarios where such information is often incomplete. To this end, this paper tackles the pervasive yet overlooked challenge of developing fair machine learning algorithms with limited demographics. Specifically, we explore leveraging limited demographic information to accurately infer missing demographics while simultaneously evaluating and optimizing model fairness. We argue that this approach better aligns with common real-world socially sensitive scenarios involving limited demographics. Extensive experiments on three benchmark datasets highlight the effectiveness of the proposed method, surpassing state-of-the-art with significant gains in fairness while maintaining comparable utility.

1 Introduction

Artificial intelligence (AI)-driven decision-making systems are increasingly being applied in a wide range of high-risk decision-making scenarios, such as healthcare [Wang and Zhang, 2024], employment [Tiwari, 2023], credit assessment [Wang *et al.*, 2024d], and criminal justice [Travaini *et al.*, 2022]. Despite the significant success, there is growing concern that they may inadvertently discriminate against certain subgroups defined by demographics (*e.g.*, gender or race) [Vasudevan and Kenthapadi, 2020; Le Quy *et al.*, 2022; Zhang, 2024]. As a result, a substantial body of research has emerged focused on enhancing fairness in machine learning algorithms [Zhang *et al.*, 2025]. The majority of these works address the problem by adopting the statistical group fairness approach, which first identifies a small collection of high-level groups defined by demographics and then ensures similar outcome statistics of the predictor across these subgroups [Wang and Zhang, 2025]. The aim is to prevent situations where one socially salient group is collectively allocated a more favorable outcome compared to another (*e.g.*, determining which patients receive extra medical care or which customers receive promotional deals) [Zhang *et al.*, 2023b]. A common assumption in these prior works is that demographic information is readily and completely available as a precondition for defining fairness and for implementing debiasing algorithms that depend on these notions [Choi *et al.*, 2021; Zhang and Ntoutsi, 2019; Zhang and Weiss, 2022; Wang *et al.*, 2024c; Wang *et al.*, 2024b].

In real-world applications, however, demographics may not always be accessible due to privacy concerns [Coston *et al.*, 2019], or fear of discrimination [Krumpal, 2013], which poses challenges for the existing fairness approaches. For instance, consider a healthcare organization that utilizes an AI algorithm to optimize patient assignments based on factors like medical history and demographics. Patients from certain racial or ethnic backgrounds may withhold their race due to concerns about bias or prior experiences of discrimination in healthcare settings [Weber *et al.*, 2021]. In such scenarios, the majority of existing fairness approaches are rendered inapplicable, as they depend on access to complete demographics.

To this end, the research community has begun exploring methods to achieve fairness without complete demographics, either by inferring demographic proxies or using strategies like Max-Min fairness [Grari *et al.*, 2021; Hashimoto *et al.*, 2018]. However, a common limitation of these methods is that they assume an extreme case where demographic information is completely unavailable [Ashurst and Weller, 2023]. In reality, some demographics may still be available, even though certain users refuse to share their demographics for various reasons. For example, on the Facebook dataset commonly used to mitigate bias in friend recommendation models, 14% of teen users have made their full profiles public [Madden *et al.*, 2013], providing some available demographic information. Ignoring this rich available demographic information, as many existing methods do, can result in greater performance losses due to imposed fairness constraints [Kenfack *et al.*, 2024].

Furthermore, when inferring demographic proxies, many existing methods assume that demographics can be accurately inferred without adequately filtering out noisy or irrelevant

*Corresponding author.

information. For instance, if the demographic information of interest is gender, other attributes like race may introduce bias if not properly accounted for during the inference process. This can result in unrealistic proxies, undermining the effectiveness of fairness mitigation [Wang *et al.*, 2025c]. On the other hand, the Max-Min fairness strategy cannot guarantee that the uncovered groups are consistent with the demographics of interest [Yan *et al.*, 2020], as the identified groups may be influenced by irrelevant features rather than the targeted demographic information.

To this end, this paper aims to advance algorithmic fairness with limited demographics by collectively addressing three distinct challenges: **i) Efficiently Utilize Limited Demographic Information:** Existing methods often either ignore or simply use limited demographic information to predict missing demographics, resulting in inaccurate predictions that introduce additional bias. Ignoring limited demographics leads to unnecessary information loss, while simply using it can introduce bias due to distributional differences. Therefore, a more effective approach is needed to efficiently leverage limited demographic information to accurately infer missing demographics without introducing bias. **ii) Optimizing Fairness Amid Missing Demographics:** Conventional fairness methods depend on complete demographic information to measure and mitigate disparities across demographic groups. However, when such information is incomplete, evaluating model bias becomes difficult, as group membership cannot be determined for all samples. This presents a core challenge in defining and optimizing fairness objectives with limited demographics. **iii) Disentangling Demographic Inference and Label Prediction:** When a model is tasked with both inferring demographics and predicting labels, there is a risk that it may manipulate the inferred demographics to artificially improve fairness metrics. For example, the model could adjust demographic assignments to balance outcomes across groups, compromising both the integrity of demographic inference and predictive accuracy. It is crucial to ensure that demographic inference and label prediction remain independent and unbiased, preventing the model from exploiting this mechanism to game fairness evaluations.

To address these challenges, we propose the *Fairness Disentanglement Variational Autoencoder (FDVAE)*, a novel framework that *leverages disentangled learning and probabilistic imputation to infer missing demographic information while ensuring model fairness, to the best of our knowledge, making it the first approach of its kind*. Specifically, FDVAE uses limited available demographics to disentangle features into demographic-related and demographic-irrelevant components, enabling more accurate inference of missing demographic information and enhanced model fairness. This is achieved by decomposing the observed data into two latent variables that distinguish between demographic-related and unrelated information, allowing us to infer missing demographics through a Bayesian network by focusing on demographic-related information while excluding irrelevant information. Furthermore, we design a differentiable fairness loss that enables fairness measurement and optimization with limited demographic information. Our main contributions are as follows: **i)** We explore the effect of irrelevant information when inferring missing demographics and examine how this impacts the accuracy of demographic inference. **ii)** We propose a novel framework that employs a disentangling structure to exclude irrelevant demographic information, enabling more accurate inference of missing demographics. Additionally, we incorporate a probabilistic imputation method to enhance model fairness. **iii)** We conduct extensive experiments on three real-world benchmark datasets. The results demonstrate that our proposed method outperforms existing baselines across multiple fairness metrics while achieving comparable prediction performance in downstream tasks.

2 Related Work

Fairness is a significant challenge in machine learning systems [Wang *et al.*, 2023a; Wang *et al.*, 2023b; Wang *et al.*, 2024a; Wang *et al.*, 2025a; Zhang *et al.*, 2023a], driving extensive research into fairness-aware frameworks designed to reduce or eliminate the influence of demographic information and ensure decisions are independent of it; however, most approaches rely on constraints or regularizers that require complete demographic information, restricting their applicability in real-world scenarios where such accessibility is not guaranteed [Wang *et al.*, 2025b].

To this end, a few works have taken initial steps to address algorithmic bias without demographics. For example, ORD [Hashimoto *et al.*, 2018] optimizes for the worst-case loss over distributions close to the empirical one, improving fairness without requiring sensitive attributes. Similarly, ARL [Lahoti *et al.*, 2020] employs adversarial learning to reweight training samples, focusing on underrepresented groups and indirectly enhancing fairness. The core idea behind these approaches is rooted in Rawlsian Max-Min fairness, aiming to minimize the maximum risk across all groups [Wang *et al.*, 2023c]. Another strategy involves inferring surrogate group information from input features. For instance, Yan et al. [Yan *et al.*, 2020] use clustering to identify subgroups and balance class distributions within clusters to promote fairness. Additionally, some methods require prior knowledge of correlations between features and demographics. For example, FairRF [Zhao *et al.*, 2022] mitigates bias by controlling for features likely correlated with demographics.

Despite these advancements, existing models fall short in three key areas: i) They ignore available demographic information by assuming the complete absence of demographics, limiting their ability to achieve fairness effectively. ii) They either fail to accurately identify focused demographic information or overlook interference from demographic-irrelevant features during demographic inference, leading to unreliable fairness assessment. iii) These methods ignore the possibility of model manipulation when inferring demographics, which can lead to artificially enhanced fairness.

To this end, we propose an approach that achieves fairness through three key innovations. First, we infer missing demographics by identifying and leveraging demographic-related information through disentangled learning, without relying on pre-defined assumptions about feature relevance. Second, we proposed a novel fairness loss that enables bias measurement with limited demographic information to guide the fair-

ness optimization process. Third, we prevent potential manipulation during joint optimization by employing gradient isolation techniques, ensuring that fairness objectives do not compromise the accuracy of demographic inference.

3 Notations

Assume a set of labeled samples (x_i, y_i, s_i) drawn from the space $X \times Y \times S$. Here, $X \in \mathbb{R}^{n \times d}$ represents the feature matrix with n samples, where each sample x_i is characterized by a d-dimensional feature vector. The corresponding ground truth label vector is denoted as $Y \in \{0,1\}^n$, where $y_i = 1$ indicates a granted outcome for sample x_i, and $y_i = 0$ indicates a rejected outcome. Additionally, $S \in \{0,1\}^m$ represents the demographics, where s_i denotes the demographics for a sample x_i, and m represents the number of samples with known demographic information, where $m < n$ due to the limited demographics. We denote S_K and S_U as the set of samples with known and unknown demographic information, respectively. Last, we define the deprived group as $S_d = \{x_i \mid s_i = 0\}$ and the favored group as $S_f = \{x_i \mid s_i = 1\}$.

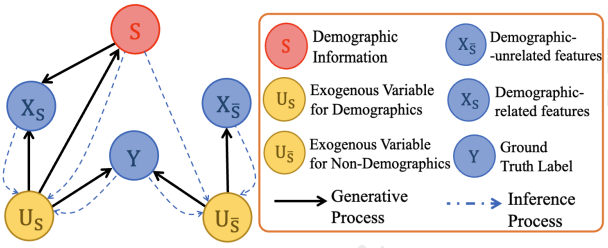

Figure 1: The causal model of the proposed method.

4 Methodology

4.1 Causal Model

This section first presents the proposed causal analysis, which is crucial for mitigating bias with limited demographics. Specifically, FDVAE infers missing demographics from observed data by leveraging the causal structure depicted in Figure 1, which considers four observed variables: demographic information (S), demographic-related features (X_S), demographic-unrelated features ($X_{\overline{S}}$), and the ground truth label (Y). Specifically, X_S and $X_{\overline{S}}$ are decomposed by model from the input features (X), where X_S are influenced by S, while $X_{\overline{S}}$ are independent of S. For instance, if the demographics S represents gender, then X_S includes features like beard or height, while $X_{\overline{S}}$ includes features unrelated to gender, such as race or weather. In other words, the demographic-related features should be independent of the demographic-unrelated features.

To implement this disentanglement, we introduce two exogenous variables, U_S and $U_{\overline{S}}$ as latent representations of X_S and $X_{\overline{S}}$, respectively. Through adversarial training with correlation measure, we ensure these latent variables are properly disentangled, where U_S captures demographic-related information and $U_{\overline{S}}$ captures demographic-irrelevant information. This design enables accurate demographic inference by leveraging demographic-related information in U_S while preventing interference from unrelated factors through the disentangling of U_S and $U_{\overline{S}}$, ensuring S is inferred solely from relevant information. For example, if the missing demographics to infer is gender, U_S captures latent information related to gender, while $U_{\overline{S}}$ captures latent information unrelated to gender, such as race, enabling focused demographic inference by using only demographic-related information from U_S. The practical implementation of this causal structure is achieved through a neural network architecture with min-max optimization, which we detail in Section 4.2. Additionally, both U_S and $U_{\overline{S}}$ contain information that contributes to accurately predicting the true label Y, while we prevent the influence of demographic information on predictions to achieve fairness.

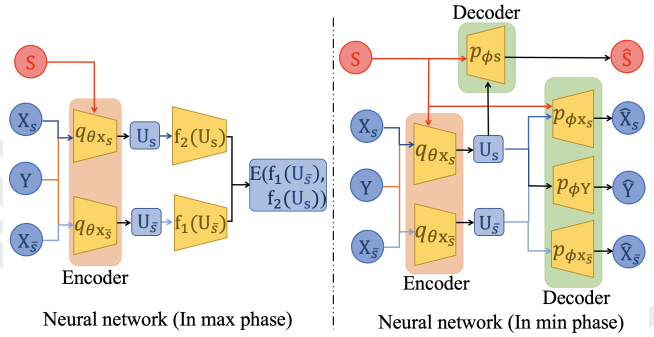

Figure 2: Neural network architecture of FDVAE.

4.2 Overview of Neural Network Architecture

Our framework consists of two main components that work together to achieve fair node classification with limited demographics through a neural network architecture that implements the Bayesian inference from our causal model, as illustrated in Figure 2. The first component employs variational inference to learn disentangled representations (U_S, $U_{\overline{S}}$) from fully observed features and partially observed demographics by maximizing the likelihood and incorporating adversarial correlation measures. Through min-max optimization, the encoder and decoder minimize reconstruction loss while maintaining latent variable independence (*i.e.*, $U_S \perp U_{\overline{S}}$), enabling reconstruction of fully observed X_S, $X_{\overline{S}}$ and Y, as well as partially observed demographics S. Second, as elaborated in Section 4.4, we achieve missing demographic inference and fair classification in a unified process through an integrated optimization framework. This dual-purpose design employs a specialized fairness loss and gradient isolation technique that ensures effective bias mitigation while maintaining accurate demographic inference.

4.3 Inferring Missing Demographics

Aligning with the causal analysis in Section 4.1, our causal model requires inference over the exogenous variables to identify the partial missing demographics S. A primary challenge in inferring S lies in accurately modeling the relationships between the observed variables and the latent vari-

ables. Previous work [Madras *et al.*, 2019] shows that unobserved variables can be recovered when the full joint distribution is successfully modeled, as causal effects are identifiable when exact inference is possible and observed covariates are sufficiently informative. Corresponding to our task, demographics can be recovered by modeling the joint distribution $P(S, X_S, X_{\overline{S}}, Y)$ among demographic-related features X_S, demographic-irrelevant features $X_{\overline{S}}$, demographics S, and ground truth label Y. Thus, we aim to learn two models: i) A decomposition model to separate the input features into those related to demographics and those that are unrelated, and ii) an inference model approximating the distribution of S given X_S, $X_{\overline{S}}$, and Y. Specifically, we employ variational inference, parameterized by deep neural networks, to jointly learn the parameters of both models. Our goal is to approximate the joint distribution $P(S, X_S, X_{\overline{S}}, Y)$ by maximizing the ELBO on the log-likelihood of the observed data, while disentangling U_S and $U_{\overline{S}}$ according to the Bayesian network structure depicted in Figure 2. As X_S and $U_{\overline{S}}$ are independent given U_S, and $X_{\overline{S}}$ and (S, U_S) are independent given $U_{\overline{S}}$, the joint probability $P(S, X_S, X_{\overline{S}}, Y)$ can be factorized as follows to learn a consistent generative model:

$$P(S, X_S, X_{\overline{S}}, Y) = \int_{U_S} \int_{U_{\overline{S}}} P(U_S)P(U_{\overline{S}})P(S|U_S) P(X_S|U_S, S)P(X_{\overline{S}}|U_{\overline{S}})P(Y|U_S, U_{\overline{S}})\, dU_S dU_{\overline{S}} \quad (1)$$

where $P(S|U_S)$ models the relationship between the exogenous variable U_S and S, effectively capturing U_S as a proxy for the demographics. The prior distributions $P(U_S)$ and $P(U_{\overline{S}})$ are typically modeled as standard normal distributions. The terms $P(X_S|U_S, S)$ and $P(X_{\overline{S}}|U_{\overline{S}})$ represent the decoders for the demographic-related and demographic-unrelated features, respectively. Finally, $P(Y|U_S, U_{\overline{S}})$ captures the dependency of the Y on both exogenous variables.

Building on this, the decoder distribution $p_\phi(X_S, X_{\overline{S}}, S, Y|U_S, U_{\overline{S}})$ can be factorized as below:

$$p_\phi(S, X_S, X_{\overline{S}}, Y) = p(U_S)p(U_{\overline{S}})p_{\phi_S}(S|U_S)p_{\phi_{X_S}}(X_S|U_S, S) p_{\phi_{X_{\overline{S}}}}(X_{\overline{S}}|U_{\overline{S}})p_{\phi_Y}(Y|U_S, U_{\overline{S}}) \quad (2)$$

We also assume the posterior $q_\phi(U_S, U_{\overline{S}}|X_S, X_{\overline{S}}, S, Y)$ can be factorized as:

$$q_\phi(U_S, U_{\overline{S}}|X_S, X_{\overline{S}}, S, Y) = q_\phi(U_S|X_S, S, Y)\, q_\phi(U_{\overline{S}}|X_{\overline{S}}, S, Y) \quad (3)$$

Given this approximate posterior, we define the Evidence Lower Bound (ELBO) [Kingma and Welling, 2013] as:

$$\log P(S, X_S, X_{\overline{S}}, Y|U_S, U_{\overline{S}}) \geq \mathbb{E}_{q_\phi(U_S, U_{\overline{S}}, S|X_S, X_{\overline{S}}, Y)} \left[\log \frac{P(X_S, X_{\overline{S}}, Y, S, U_S, U_{\overline{S}})}{q_\phi(U_S, U_{\overline{S}}|S, X_S, X_{\overline{S}}, Y)}\right] \quad (4)$$

The values of $\log P(S, X_S, X_{\overline{S}}, Y|U_S, U_{\overline{S}})$ correlates positively with the reality of observed data. The $P(X_S, X_{\overline{S}}, Y, S, U_S, U_{\overline{S}})$ represents the joint distribution between the observed data, while $q_\phi(U_S, U_{\overline{S}}|S, X_S, X_{\overline{S}}, Y)$ denote the posterior distribution of the demographics. Furthermore, to approximate the intractable posterior distribution of these latent variables, we introduce a variational distribution $Q(U_S|X_S, S, Y)$ and $Q(U_{\overline{S}}|X_{\overline{S}}, S, Y)$, which uses a parametric family of distributions to approximate the true posterior distribution $P(U_S|X_S, S, Y)$ and $P(U_{\overline{S}}|X_{\overline{S}}, S, Y)$.

Using the factorization of the variational distribution, the updated ELBO of our framework can be formally described as:

$$\begin{aligned}\log P(S, X_S, X_{\overline{S}}, Y|U_S, U_{\overline{S}}) \geq & \mathbb{E}_{q_\phi(U_S, U_{\overline{S}}|S, X_S, X_{\overline{S}}, Y)} \\ & \big[\log p_{\phi_S}(S|U_S) + \log p_{\phi_{X_S}}(X_S|U_S, S) \\ & + \log p_{\phi_{X_{\overline{S}}}}(X_{\overline{S}}|U_{\overline{S}}) + \log p_{\phi_Y}(Y|U_S, U_{\overline{S}}) \\ & - \mathrm{KL}(q_{\phi_{U_S}}(U_S|X_S, Y)\|p(U_S)) \\ & - \mathrm{KL}(q_{\phi_{U_{\overline{S}}}}(U_{\overline{S}}|X_{\overline{S}}, Y)\|p(U_{\overline{S}}))\big]\end{aligned} \quad (5)$$

where KL($\cdot$) denotes the Kullback-Leibler divergence of the posterior q_{θ_S}, $q_{\phi_{U_S}}$ and $q_{\phi_{U_{\overline{S}}}}$ from a prior $P(S)$, $P(U_S)$ and $P(U_{\overline{S}})$, as shown in follow:

$$\begin{cases} \mathrm{KL}\left(q_{\phi_{U_S}}(U_S|S, X_S, Y) \,\|\, P(U_S)\right) = \mathbb{E}_{q_{\phi_{U_S}}(U_S|S, X_S, Y)}\left[\log Q(U_S|S, X_S, Y) - \log P(U_S)\right] \\ \mathrm{KL}\left(q_{\phi_{U_{\overline{S}}}}(U_{\overline{S}}|X_{\overline{S}}, Y) \,\|\, P(U_{\overline{S}})\right) = \mathbb{E}_{q_{\phi_{U_{\overline{S}}}}(U_{\overline{S}}|X_{\overline{S}}, Y)}\left[\log Q(U_{\overline{S}}|X_{\overline{S}}, Y) - \log P(U_{\overline{S}})\right] \end{cases} \quad (6)$$

The maximization of the ELBO can be performed using stochastic gradient ascent and the reparameterization trick [Kingma and Welling, 2013]. However, merely optimizing the ELBO does not guarantee the independence between U_S and $U_{\overline{S}}$, which is crucial for accurate inference. To enforce independence through adversarial training, we first need a way to measure the dependence between latent variables. To this end, we adopt the Hirschfeld-Gebelein-Rényi (HGR) maximal correlation [Gebelein, 1941], which generalizes Pearson correlation to capture any non-linear relationship between random variables. Specifically, this HGR-based min-max optimization process involves two competing phases, as illustrated in Figure 2. In the minimization phase, the encoder and decoder parameters ϕ and θ are updated via gradient descent to minimize both the reconstruction loss and the dependence between U_S and $U_{\overline{S}}$. Simultaneously, in the maximization phase, two adversarial networks with parameters ω_{f_1} and ω_{f_2} take U_S and $U_{\overline{S}}$ as inputs, respectively, and are updated via gradient ascent to maximize their estimated dependence. This adversarial process measures the degree of disentanglement through iterative optimization. We incorporate this into our loss function through a penalization term $\mathcal{L}_D$, as shown in Equation 7. This alternating optimization between the main network and adversarial networks allows for increasingly accurate estimation of the HGR correlation at each step, leading to more stable disentanglement of the latent variables.

$$\mathcal{L}_D = \sup_{f_1, f_2} \frac{\mathbb{E}(f_1(U_S) f_2(U_{\overline{S}}))}{\sqrt{\mathbb{E}({f_1}^2(U_S))\mathbb{E}({f_2}^2(U_{\overline{S}}))}} \quad (7)$$

where f_1 and f_2 are measurable functions with positive and finite variance. Therefore, the ELBO can be reformulated as:

$$\mathcal{L} = -\mathcal{L}_{\text{ELBO}} + \lambda \mathcal{L}_D \tag{8}$$

where λ is a hyperparameter that balances the contribution of the disentanglement term. By minimizing this loss function, we ensure that the latent variables U_S and $U_{\overline{S}}$ are disentangled, enabling accurate inference of the missing demographics while mitigating bias.

4.4 Mitigation Bias with Limited Demographics

As discussed in Section 4.2, the second component is designed to make fair and accurate predictions. However, existing fairness-aware methods typically assume that complete demographics are available beforehand during this integrated process, which is unsuitable for simultaneous demographic inference and fairness optimization. Consequently, these methods cannot be directly applied. To address this issue, we introduce a differentiable fairness loss that leverages the probabilistic estimations of the demographic attributes during training. Specifically, we model the uncertainty in group membership X_U due to the missing S by considering the predicted probabilities $P(\hat{S}|U_S)$, where $\hat{S}$ is estimated from the latent variable U_S as training progresses.

The quality of these demographic predictions directly impacts our ability to ensure fairness, as inaccurate demographic inference could lead to incorrect fairness assessments, such as incorrectly labeling an individual from one demographic group as belonging to another, which could lead to incorrect assessments of fairness. To mitigate this risk, we impose a prior constraint on $P(\hat{S}|U_S)$ to minimize the probability of misclassification. This is achieved by employing a Dirichlet prior that assigns low probabilities to misclassifications of demographics (*e.g.*, $P(female|male)$ to a small value), thereby discouraging incorrect predictions that could adversely affect the fairness evaluation. Building upon this, we define the Estimated Group Disparity (EGD) as follows:

Definition 4.2 (Estimated Group Disparity). Given an input dataset, the Estimated Group Disparity is defined as:

$$\begin{aligned}\text{EGD} = &\left(\frac{\sum_i P(\hat{s}_i = 1 \,|\, U_S) \cdot p_{\phi_Y}(\hat{y}_i = 1 \,|\, x_i)}{\sum_i P(\hat{s}_i = 1 \,|\, U_S)}\right) \\ &- \left(\frac{\sum_i P(\hat{s}_i = 0 \,|\, U_S) \cdot p_{\phi_Y}(\hat{y}_i = 1 \,|\, x_i)}{\sum_i P(\hat{s}_i = 0 \,|\, U_S)}\right)\end{aligned} \tag{9}$$

where $P(\hat{s}_i = s \,|\, U_S)$ denotes the probability that sample x_i belongs to demographic group s as inferred from U_S, and $p_{\phi_Y}(\hat{y}_i = 1 \,|\, x_i)$ is the probability that the model predicts a granted label for sample x_i. In practice, EGD computes the estimated fairness loss by comparing the probability average predicted granted label between different demographic subgroups as identified by the inferred S.

However, while EGD enables fairness assessment with limited demographic information, directly integrating it into training poses challenges due to the concurrent optimization of demographic inference and prediction tasks. Specifically, when performing these tasks simultaneously, the model might manipulate demographic assignments to artificially reduce the fairness loss (*e.g.*, changing an individual from one group to another to achieve statistical parity). This issue arises when jointly optimizing the parameters of both the inference model q_θ and the predictor p_{ϕ_Y}. To prevent the fairness loss from influencing the learning of q_θ, we employ the stop-gradient technique. Specifically, during backpropagation, we halt the gradient of the EGD with respect to the parameters of q_θ, effectively treating the inferred demographic probabilities as constants in the computation of the fairness loss. This strategy maintains a separation between the learning of the demographic inference and the prediction model, preventing potential manipulation and simplifying the training process.

Furthermore, the prediction of Y not only affects predictive performance but also contributes to fairness metrics. If the model adjusts Y to optimize fairness, it may compromise accuracy. Essentially, since the fairness loss depends on the predictions of Y, applying a stop gradient prevents the simultaneous optimization of both fairness and performance objectives. To this end, we introduce a separate semi-supervised classifier $p_{\phi'_Y}(Y|x_i)$ that estimates Y without considering fairness constraints. This classifier is trained using available labeled data and semi-supervised techniques for unlabeled instances. We then use the predictions from $p_{\phi'_Y}$ to inform the fairness evaluation in the EGD. Specifically, for samples without observed labels, we use the estimated class probabilities from $p_{\phi'_Y}$ in the computation of the EGD. The updated EGD is defined as:

$$\mathcal{L}_{\text{EGD}} = \left(\frac{\sum_i P(\hat{s}_i = 1 \,|\, U_S) \cdot \hat{y}'_i}{\sum_i P(\hat{s}_i = 1 \,|\, U_S)}\right) - \left(\frac{\sum_i P(\hat{s}_i = 0 \,|\, U_S) \cdot \hat{y}'_i}{\sum_i P(\hat{s}_i = 0 \,|\, U_S)}\right) \tag{10}$$

where $\hat{y}'_i = p_{\phi'_Y}(Y = 1 \,|\, x_i)$ is the predicted probability of a granted label from the auxiliary classifier $p_{\phi'_Y}$. This modification ensures that the fairness loss is computed based on predictions that are not influenced by the fairness optimization of the main model, thereby avoiding unintended manipulation.

Finally, we incorporate the EGD into the ELBO of our variational framework. The extended ELBO becomes:

$$\mathcal{L} = -\mathcal{L}_{\text{ELBO}} + \lambda \mathcal{L}_D + \gamma \mathcal{L}_{EGD} \tag{11}$$

where γ denotes a hyperparameter that balances the contribution of fairness.

5 Experiment

5.1 Experimental Setup

Datasets. We evaluate the effectiveness of our proposed FD-VAE framework on three widely used datasets in the fairness domain: Adult [Ding *et al.*, 2021], COMPAS [Larson *et al.*, 2016], and CelebA [Liu *et al.*, 2015]. i) The **Adult** dataset consists of 49,531 samples with 14 attributes. The task is to predict whether an individual's income exceeds \$50K per year, with gender serving as the demographic for fairness evaluation. ii) The **COMPAS** dataset contains 6,150 samples (after selecting only Black and White defendants) with 11 attributes. The goal is to predict whether a defendant will reoffend within two years. Similar to the Adult dataset, we

Datasets	Methods / Metrics	ARL	DRO	FairKD	FairRF	FairDA	Reckoner	FDVAE
Adult	**Accuracy** (↑)	0.817 ± 0.016	0.817 ± 0.009	**0.847 ± 0.004**	0.829 ± 0.003	0.823 ± 0.004	0.811± 0.008	0.841 ± 0.012
	F1-Score (↑)	**0.671 ± 0.005**	0.643 ± 0.010	0.590±0.042	0.613 ± 0.012	0.597 ± 0.012	0.553±0.021	0.646 ± 0.018
	Δ DP(↓)	0.117 ± 0.025	0.123 ± 0.037	0.102 ± 0.016	0.084 ± 0.008	0.088 ± 0.005	0.097±0.006	**0.068 ± 0.008**
	ΔEO(↓)	0.108 ± 0.019	0.102 ± 0.024	0.105 ± 0.018	0.097 ± 0.023	0.071 ± 0.003	0.059±0.002	**0.054 ± 0.010**
COMPAS	**Accuracy** (↑)	0.633 ± 0.014	0.626 ± 0.008	**0.653 ± 0.011**	0.632 ± 0.007	0.612 ± 0.024	0.647±0.018	0.652 ± 0.031
	F1-Score (↑)	0.599 ± 0.036	0.637 ± 0.018	0.617 ± 0.023	0.605 ± 0.012	0.627 ± 0.009	**0.669±0.021**	0.667 ± 0.032
	ΔDP(↓)	0.148 ± 0.015	0.142 ± 0.021	0.115 ± 0.012	0.119±0.019	0.095 ± 0.008	0.142±0.010	**0.091 ± 0.007**
	ΔEO(↓)	0.141 ± 0.028	0.133 ± 0.037	0.108 ± 0.019	0.123±0.027	0.097 ± 0.016	0.148±0.011	**0.093 ± 0.013**
CelebA	**Accuracy** (↑)	0.802 ± 0.007	0.766 ± 0.003	0.808 ± 0.003	**0.843± 0.027**	0.841 ± 0.019	0.795±0.010	0.827 ± 0.024
	F1-Score (↑)	0.486 ± 0.009	**0.491 ± 0.012**	0.358 ± 0.029	0.407±0.033	0.403 ± 0.022	0.375±0.021	0.413 ± 0.029
	ΔDP(↓)	0.221 ± 0.015	0.232 ± 0.016	0.147 ± 0.022	0.128±0.017	0.115 ± 0.008	0.136±0.004	**0.107 ± 0.036**
	ΔEO(↓)	0.258 ± 0.023	0.246 ± 0.027	0.138 ± 0.020	0.211± 0.024	0.187 ± 0.016	0.118±0.011	**0.116 ± 0.005**

Table 1: Comparison results of FDVAE with baseline methods across real-world datasets. In each row, the best result is indicated in bold, while the runner-up result is marked with an underline.

use gender as the demographics. iii) The **CelebA** dataset includes 202,599 face images, each of size 178×218 pixels, annotated with 40 binary attributes. We conduct the binary classification tasks on this dataset: predicting attractiveness with gender as the demographics. For all datasets, we randomly split the data into 50% training data, 20% validation data, and 30% test data. All the methods evaluated are trained and tested on the same data partitions each time. To simulate missing demographics, we create two scenarios: i) Meager Level: We randomly mask the demographics of 20% of individuals who are both favored and granted or favored and rejected, as well as 10% of those who are deprived and granted or deprived and rejected. ii) Serried Level: We increase the masking proportions to 40% for the favored-granted and favored-rejected groups and 20% for the deprived-granted and deprived-rejected groups. The masking is applied to both the training and validation sets, and no demographics are used during testing.

Baselines. We evaluate the performance of our proposed FDVAE by comparing it with several baseline methods: i) **ARL** [Lahoti *et al.*, 2020]: Introduces adversarially reweighted learning to ensure fairness without using demographic data by reweighting training samples to mitigate biases. ii) **DRO** [Hashimoto *et al.*, 2018]: Achieves fairness in repeated loss minimization by controlling the worst-case performance over time without relying on demographics. iii) **FairKD** [Chai *et al.*, 2022]: Utilizes Rawlsian Max-Min fairness through knowledge distillation to achieve fairness across all subgroups. iv) **FairRF** [Zhao *et al.*, 2022]: Develops fair classifiers by exploring feature-related biases, eliminating the need for sensitive attribute data. v) **FairDA** [Liang *et al.*, 2023]: Employs a dual adversarial learning approach to ensure fair classification via domain adaptation without relying on demographic or sensitive attributes. vi) **Reckoner** [Ni *et al.*, 2024] achieves group fairness through learnable noise and knowledge-sharing in a dual-model architecture.

Evaluation Metrics. The evaluation involves two fairness metrics and two machine learning performance metrics. We use Accuracy and F1-score to assess the utility performance of the models. To evaluate fairness, we adopt two commonly used metrics: ΔDP [Dwork *et al.*, 2012] and ΔEO [Hardt *et al.*, 2016]. For both ΔDP and ΔEO, smaller values indicate fairer model predictions.

5.2 Experimental Results

Comparison Study. We compared the performance of FDVAE with six baseline methods, and Table 1 summarizes the results for the classification task. As one can see, FDVAE consistently outperforms all baseline methods across most evaluation metrics. Specifically, FDVAE demonstrates superior fairness performance (*i.e.*, ΔDP and ΔEO), as shown by the significant improvements in all baseline methods across various datasets. This enhanced fairness can be attributed to two main factors: i) FDVAE accurately infers missing demographics by filtering out demographic-irrelevant features, thereby laying the foundation for the fairness loss estimation in the model. ii) Our dual focus on inferring missing demographics and predicting sample labels prevents spurious fairness improvements, such as artificially enhancing fairness by manipulating sample demographics. Additionally, FDVAE achieves excellent utility performance, surpassing other methods in most cases. This outcome suggests that FDVAE's accurate demographic inference minimizes performance loss due to the enhancement of fairness associated with incorrect demographics. On the other hand, compared with Rawlsian Max-Min fairness approaches that may incur performance penalties when addressing fairness in not-focus subgroups, FDVAE maintains high utility. Overall, the experimental results validate FDVAE's effectiveness in achieving improved fairness while maintaining strong utility performance.

Quality of Proxy to Recovery Demographics. Table 2 shows the accuracy of FDVAE in recovering missing demographics across three datasets and two sparsity levels. At the Meager Level, FDVAE infers missing demographic attributes more accurately due to the availability of more labeled data. At the Serried Level, while there is a slight decrease in accuracy compared to the Meager Level, the results remain close, demonstrating FDVAE's robustness across different data sparsity conditions. We also introduced a variant, FDVAE-ND, which measures the accuracy of demographic recovery without filtering out demographic-irrelevant information. The results clearly show that failing to exclude irrelevant information significantly reduces accuracy, underscoring the importance of our approach in excluding demographic-

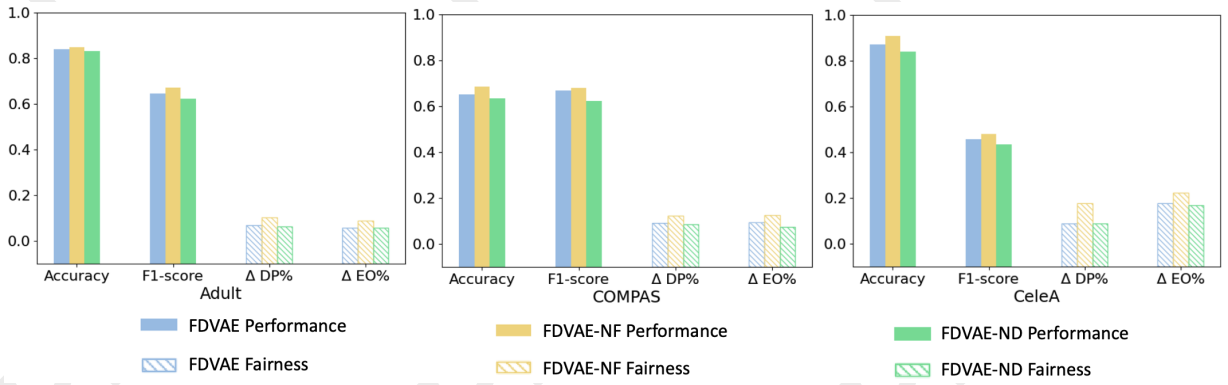

Figure 3: Ablation study results for FDVAE, FDVAE-NF, and FDVAE-ND.

irrelevant features during the inference process.

Level	Datasets	FDVAE (Accuracy)	FDVAE-ND (Accuracy)
Meager Level	Adult	0.675	0.626
	COMPAS	0.701	0.631
	CelebA	0.640	0.600
Serried Level	Adult	0.659	0.593
	COMPAS	0.673	0.614
	CelebA	0.616	0.578

Table 2: Recovered demographics results of FDVAE and FDVAE-ND (in terms of Accuracy).

Ablation Studies. We conducted ablation studies to understand the impact of each component of FDVAE on improving fairness. Specifically, we constructed two variants: FDVAE-ND and FDVAE-NF. FDVAE-ND does not exclude irrelevant information when inferring missing demographic information (*i.e.*, $\lambda = 0$), while FDVAE-NF prioritizes performance without considering fairness (*i.e.*, $\gamma = 0$). The results are presented in Figure 3. Compared to FDVAE, FDVAE-ND exhibits a decrease in both fairness and performance. This decline occurs because the reduced accuracy in inferring missing demographic information introduces additional bias, thereby compromising the effectiveness of subsequent bias mitigation. Similarly, FDVAE-NF shows a significant reduction in model fairness due to the absence of fairness considerations. These findings underscore the necessity of FDVAE's design, highlighting the importance of excluding information unrelated to demographics during the inference process to achieve effective bias mitigation.

Parameters Sensitivity. We examined the sensitivity of FDVAE with respect to two hyperparameters, λ and γ. In FairSAD, λ and γ control the balance between disentanglement and fairness. We varied λ and γ within the set $\{e^{-3}, e^{-2}, \ldots, e^{3}\}$, where e is the natural constant. Figure 4 shows the results of the parameter sensitivity analysis using the Adult dataset as an example. Our observations are as follows: i) The overall performance of FairSAD remains stable over a broad range of λ and γ values. ii) The model's fairness improvement is influenced by disentanglement, and utility performance degradation is more pronounced when disentanglement is poor. This finding underscores the equal importance of both disentanglement and fairness contributions.

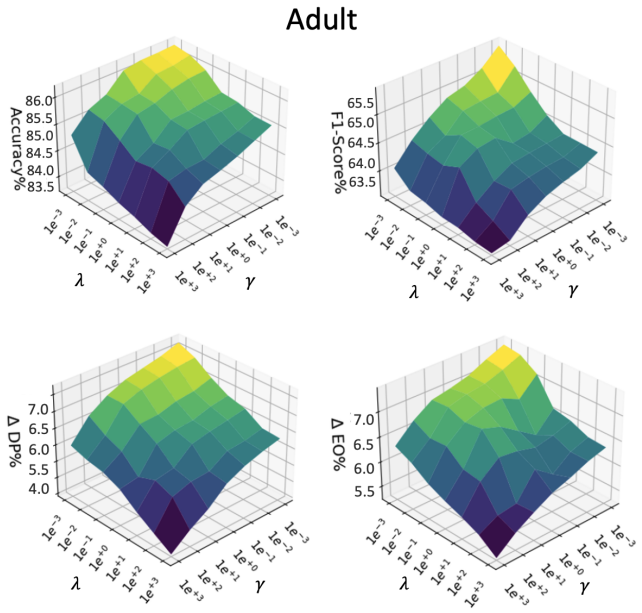

Figure 4: Parameters sensitivity analysis on Adult dataset.

6 Conclusion

Despite the growing attention to AI fairness, existing fairness studies typically assume that demographic information is either fully available or completely missing, thereby overlooking the real-world scenario of partial demographic availability. To this end, we propose a novel disentangled learning approach that effectively identifies demographic-related information from observed data to infer missing demographics while preventing the manipulation of inferred demographics during fairness optimization. Results on a real biased dataset show that our method effectively mitigates model bias by leveraging the available demographic information. This work opens a promising direction for developing fair ML algorithms that can operate with partially available demographic information.

Acknowledgements

This work was supported in part by the National Science Foundation (NSF) under Grant No. 2404039.

References

[Ashurst and Weller, 2023] Carolyn Ashurst and Adrian Weller. Fairness without demographic data: A survey of approaches. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–12, 2023.

[Chai *et al.*, 2022] Junyi Chai, Taeuk Jang, and Xiaoqian Wang. Fairness without demographics through knowledge distillation. *Advances in Neural Information Processing Systems*, 35:19152–19164, 2022.

[Choi *et al.*, 2021] YooJung Choi, Meihua Dang, and Guy Van den Broeck. Group fairness by probabilistic modeling with latent fair decisions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12051–12059, 2021.

[Coston *et al.*, 2019] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. Fair transfer learning with missing protected attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 91–98, 2019.

[Ding *et al.*, 2021] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

[Dwork *et al.*, 2012] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012.

[Gebelein, 1941] Hans Gebelein. Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379, 1941.

[Grari *et al.*, 2021] Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. Fairness without the sensitive attribute via causal variational autoencoder. *arXiv preprint arXiv:2109.04999*, 2021.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[Hashimoto *et al.*, 2018] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

[Kenfack *et al.*, 2024] Patrik Joslin Kenfack, Samira Ebrahimi Kahou, and Ulrich Aïvodji. A survey on fairness without demographics. *Transactions on Machine Learning Research*, 2024.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Krumpal, 2013] Ivar Krumpal. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & quantity*, 47(4):2025–2047, 2013.

[Lahoti *et al.*, 2020] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

[Larson *et al.*, 2016] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. Compas analysis. *GitHub. Available online: https://github. com/propublica/compas-analysis (accessed on 5 February 2022)*, 2016.

[Le Quy *et al.*, 2022] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3):e1452, 2022.

[Liang *et al.*, 2023] Yueqing Liang, Canyu Chen, Tian Tian, and Kai Shu. Fair classification via domain adaptation: A dual adversarial learning approach. *Frontiers in Big Data*, 5:1049565, 2023.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[Madden *et al.*, 2013] Mary Madden, Amanda Lenhart, Sandra Cortesi, Urs Gasser, Maeve Duggan, Aaron Smith, and Meredith Beaton. Teens, social media, and privacy. *Pew Research Center*, 21(1055):2–86, 2013.

[Madras *et al.*, 2019] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 349–358, 2019.

[Ni *et al.*, 2024] Hongliang Ni, Lei Han, Tong Chen, Shazia Sadiq, and Gianluca Demartini. Fairness without sensitive attributes via knowledge sharing. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1897–1906, 2024.

[Tiwari, 2023] Rudra Tiwari. The impact of ai and machine learning on job displacement and employment opportunities. *International Journal of Engineering Technologies and Management Research*, 7(1), 2023.

[Travaini *et al.*, 2022] Guido Vittorio Travaini, Federico Pacchioni, Silvia Bellumore, Marta Bosia, and Francesco De Micco. Machine learning and criminal justice: A systematic review of advanced methodology for recidivism risk prediction. *International journal of environmental research and public health*, 19(17):10594, 2022.

[Vasudevan and Kenthapadi, 2020] Sriram Vasudevan and Krishnaram Kenthapadi. Lift: A scalable framework for measuring fairness in ml applications. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 2773–2780, 2020.

[Wang and Zhang, 2024] Zichong Wang and Wenbin Zhang. Group fairness with individual and censorship constraints. In *27th European Conference on Artificial Intelligence*, 2024.

[Wang and Zhang, 2025] Zichong Wang and Wenbin Zhang. Fdgen: A fairness-aware graph generation model. In *Proceedings of the 42nd International Conference on Machine Learning*. PMLR, 2025.

[Wang *et al.*, 2023a] Zichong Wang, Giri Narasimhan, Xin Yao, and Wenbin Zhang. Mitigating multisource biases in graph neural networks via real counterfactual samples. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 638–647. IEEE, 2023.

[Wang *et al.*, 2023b] Zichong Wang, Nripsuta Saxena, Tongjia Yu, Sneha Karki, Tyler Zetty, Israat Haque, Shan Zhou, Dukka Kc, Ian Stockwell, Albert Bifet, et al. Preventing discriminatory decision-making in evolving data streams. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.

[Wang *et al.*, 2023c] Zichong Wang, Charles Wallace, Albert Bifet, Xin Yao, and Wenbin Zhang. Fg^2an: Fairness-aware graph generative adversarial networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 259–275. Springer Nature Switzerland, 2023.

[Wang *et al.*, 2024a] Zichong Wang, Zhibo Chu, Ronald Blanco, Zhong Chen, Shu-Ching Chen, and Wenbin Zhang. Advancing graph counterfactual fairness through fair disentangled representation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2024.

[Wang *et al.*, 2024b] Zichong Wang, Jocelyn Dzuong, Xiaoyong Yuan, Zhong Chen, Yanzhao Wu, Xin Yao, and Wenbin Zhang. Individual fairness with group awareness under uncertainty. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 89–106. Springer Nature Switzerland, 2024.

[Wang *et al.*, 2024c] Zichong Wang, Meikang Qiu, Min Chen, Malek Ben Salem, Xin Yao, and Wenbin Zhang. Toward fair graph neural networks via real counterfactual samples. *Knowledge and Information Systems*, pages 1–25, 2024.

[Wang *et al.*, 2024d] Zichong Wang, David Ulloa, Tongjia Yu, Raju Rangaswami, Roland Yap, and Wenbin Zhang. Individual fairness with group constraints in graph neural networks. In *27th European Conference on Artificial Intelligence*, 2024.

[Wang *et al.*, 2025a] Zichong Wang, Zhibo Chu, Thang Viet Doan, Shaowei Wang, Yongkai Wu, Vasile Palade, and Wenbin Zhang. Fair graph u-net: A fair graph learning framework integrating group and individual awareness. In *proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 28485–28493, 2025.

[Wang *et al.*, 2025b] Zichong Wang, Nhat Hoang, Xingyu Zhang, Kevin Bello, Xiangliang Zhang, Sundararaja Sitharama Iyengar, and Wenbin Zhang. Towards fair graph learning without demographic information. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

[Wang *et al.*, 2025c] Zichong Wang, Fang Liu, Shimei Pan, Jun Liu, Fahad Saeed, Meikang Qiu, and Wenbin Zhang. fairgnn-wod: Fair graph learning without complete demographics. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*, 2025.

[Weber *et al.*, 2021] Ann M Weber, Ribhav Gupta, Safa Abdalla, Beniamino Cislaghi, Valerie Meausoone, and Gary L Darmstadt. Gender-related data missingness, imbalance and bias in global health surveys. *BMJ global health*, 6(11):e007405, 2021.

[Yan *et al.*, 2020] Shen Yan, Hsien-te Kao, and Emilio Ferrara. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1715–1724, 2020.

[Zhang and Ntoutsi, 2019] Wenbin Zhang and Eirini Ntoutsi. Faht: an adaptive fairness-aware decision tree classifier. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 1480–1486, 2019.

[Zhang and Weiss, 2022] Wenbin Zhang and Jeremy C Weiss. Longitudinal fairness with censorship. In *proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 12235–12243, 2022.

[Zhang *et al.*, 2023a] Wenbin Zhang, Tina Hernandez-Boussard, and Jeremy Weiss. Censored fairness through awareness. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 2023.

[Zhang *et al.*, 2023b] Wenbin Zhang, Zichong Wang, Juyong Kim, Chen Cheng, Thomas Oommen, Pradeep Ravikumar, and Jeremy Weiss. Individual fairness under uncertainty. In *26th European Conference on Artificial Intelligence*, pages 3042–3049, 2023.

[Zhang *et al.*, 2025] Wenbin Zhang, Shuigeng Zhou, Toby Walsh, and Jeremy C Weiss. Fairness amidst non-iid graph data: A literature review. *AI Magazine*, 46(1):e12212, 2025.

[Zhang, 2024] Wenbin Zhang. Ai fairness in practice: Paradigm, challenges, and prospects. *Ai Magazine*, 45(3):386–395, 2024.

[Zhao *et al.*, 2022] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. Towards fair classifiers without sensitive attributes: Exploring biases in related features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1433–1442, 2022.