Fair Representation Learning with Controllable High Confidence Guarantees via Adversarial Inference

Yuhong Luo
Rutgers University
New Brunswick, NJ, USA
y.luo@rutgers.edu

Austin Hoag
Sony AI
New York, NY, USA
austin.hoag@sony.com

Xintong Wang
Rutgers University
New Brunswick, NJ, USA
xintong.wang@rutgers.edu

Philip S. Thomas
University of Massachusetts
Amherst, MA, USA
pthomas@cs.umass.edu

Przemyslaw A. Grabowicz [1,2]
[1]University College Dublin, Ireland
[2]University of Massachusetts, Amherst, MA, USA
przemek.grabowicz@ucd.ie

Abstract

Representation learning is increasingly applied to generate representations that generalize well across multiple downstream tasks. Ensuring fairness guarantees in representation learning is crucial to prevent unfairness toward specific demographic groups in downstream tasks. In this work, we formally introduce the task of learning representations that achieve high-confidence fairness. We aim to guarantee that demographic disparity in every downstream prediction remains bounded by a *user-defined* error threshold ε, with *controllable* high probability. To this end, we propose the ***F**air **R**epresentation learning with high-confidence **G**uarantees* (FRG) framework, which provides these high-confidence fairness guarantees by leveraging an optimized adversarial model. We empirically evaluate FRG on three real-world datasets, comparing its performance to six state-of-the-art fair representation learning methods. Our results demonstrate that FRG consistently bounds unfairness across a range of downstream models and tasks. The source code for FRG is available at: https://github.com/JamesLuoyh/FRG.

1 Introduction

In every prediction task, machine learning algorithms assume two distinct roles: the data producer and the data consumer [19, 48, 75]. The data consumer's role is to make accurate predictions using the data provided by the data producer. While the data producer may distribute raw data, it is common to generate new representations via *representation learning* for the input data that are used as predictors in downstream tasks. When multiple data consumers' prediction tasks involve inputs of the same type, such as natural language text or images, the data producer can generate *general representations* that are predictive to multiple subsequent tasks. This is an increasing trend with examples including the Variational Auto-Encoder (VAE) [38] or recent language models such as BERT [16] and GPT-4 [58], which are widely used as bases for downstream text classification tasks [13].

While representation learning can benefit various downstream predictions, it is also susceptible to the risk of producing unintended or undesirable behaviors in the downstream tasks, specifically, generating predictions that are unfair toward disadvantaged demographic groups. Especially in critical

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

domains, such as loan underwriting [9], hiring [53] and criminal sentencing [4], the consequences of algorithmic bias may severely impact individuals. To address these concerns, researchers have proposed fair representation learning (FRL), emphasizing that fairness should be the responsibility of the data producer who generates the representations [48, 75], rather than the data consumer who uses them. By ensuring fairness at the representation level, the data producer guarantees fairness across all downstream tasks, allowing the representations to be safely used by any data consumer.

Extensive prior work in FRL has shown effectiveness in promoting fairness for specific downstream tasks. Some methods [2, 26, 48, 51] *estimate* upper bounds for the unfairness across all downstream models and tasks based on the *training dataset*. However, there is *no guarantee* that these estimations give true upper bounds. These bounds can be *underestimated* because of overfitting to the training and validation sets. Thus, when their models are deployed on unseen test data, they can fail the desired fairness requirements. This calls for statistical guarantees such as high-confidence guarantees.

High-confidence guarantees are required to ensure that the unfairness across all downstream models and tasks will be *consistently* bounded with *high probabilities*. In many areas of supervised learning, providing high-confidence guarantees is considered essential for ensuring the fairness, privacy, and safety of the learning algorithm [1, 18, 43, 70]. This need becomes even more critical in the context of FRL as the absence of such guarantees can lead to undesired behaviors across multiple downstream applications. In FRL, a method called FARE [33] provides certificates that the downstream unfairness will be bound by some threshold with high probability. However, how to let users *explicitly control* the error thresholds and confidence levels jointly for the high-confidence guarantees is unexplored.

We propose ***F**air **R**epresentation learning with high-confidence **G**uarantees* (FRG), the first work to our knowledge that guarantees with high confidence that the output representation models are "fair" according to *user-specified* thresholds of unfairness and confidence levels. The fairness criterion we focus on is demographic parity (DP) [19], a group fairness notion (the extension to equal opportunity and equalized odds will be discussed in Appendix C). A fair representation model ensures that, across arbitrary downstream tasks and models, the largest difference in positive prediction rate – denoted as Δ_{DP} (defined in Sec. 3) – is bounded by an error threshold $\varepsilon \in [0, 1]$. Thus, once a user designates a threshold ε for Δ_{DP} and a confidence level, FRG guarantees with high *confidence* that any (including adversarial) downstream tasks and models using the representations generated by its learned representation model will not exceed the threshold ε. Equivalently, the representation models output by FRG will have a high probability of satisfying the desired fairness constraint even when applied to a worst-case adversarial downstream model.

FRG consists of three major components: (1) the *candidate selection* component proposes a representation model that will likely satisfy the fairness constraint with high probability; (2) the *adversarial inference* component aims to utilize the representations learned by the proposed model to adversarially predict the sensitive attributes to maximize Δ_{DP}; and (3) the *fairness test* component establishes a high-confidence upper bound on the worst-case downstream Δ_{DP} based on the "optimal" adversarial prediction to determine whether the proposed model passes the test and should be returned.

We provide theoretical justification for the high-confidence fairness guarantees based on the assumption that the fairness test has access to an optimal adversary (defined in Sec. 5) which is approximated with optimization. We find a direct mapping between Δ_{DP} and the absolute covariance between the sensitive attributes and the predictions, so the optimal adversary can be achieved by maximizing the absolute covariance. Alternatively, it is possible to find a high-confidence upper bound via the upper bounds on Δ_{DP} as derived by previous work [26, 68] to avoid relying on training an optimal adversary. However, our study shows that these upper bounds are typically loose and thus impractical for establishing guarantees while preserving utility (Appendix M). Specifically, there exists a non-trivial gap between Δ_{DP} and its theoretical upper bound (as demonstrated in Appendix Figure 12 and 13).

In experiments, we use three real-world datasets each with 2-3 tasks to verify that FRG can indeed be used to learn fair representation models that satisfy the fairness criteria with the desired high probability. Compared to FRG, six state-of-the-art (SOTA) FRLs either violate the fairness constraints with non-trivial probability (at least 0.1) or achieve lower predictive performance than FRG.

2 Related Work

Fair representation learning (FRL) has been studied for at least a decade [75]. While a stream of FRL studies optimizing the representations for a specific downstream task [10, 11, 24, 51, 62, 67, 74, 77]. numerous FRL methods learn general representations [2, 30, 52] that are fair, even when downstream

tasks are unknown or unlabeled. One category of these methods draws inspiration from information and probability theory [26, 31, 35, 37, 45, 54, 63, 65, 68, 72]. One work explores the use of distance covariance [44]. Some methods can limit downstream unfairness by constraining the total variation distance between the representation distributions of different groups [5, 48, 66, 76]. Other approaches promote independence from sensitive attributes through penalizing Maximum Mean Discrepancy [15, 47, 56] or statistical dependence [25, 61], meta-learning [57], PCA [39, 42], learning a shared feature space between groups [12], or disentanglement [14, 46, 55]. A stream of work uses adversarial training that limits the adversary's performance in predicting sensitive attributes [20, 22, 36, 60, 71]. Different from these methods, FRG constructs high-confidence guarantees based on a separately trained adversary without joint optimization with the primary objective under fairness constraints, which is considered more reliable than adversarial training.

Some FRLs provide theoretical analyses. Several works [26, 32, 48, 66, 76] prove upper bounds on the unfairness of all downstream models and tasks. These bounds can be estimated and verified with a training and validation set. However, these bounds may fail to generalize to an unseen test set. Some other works [5, 23, 33] provide statistical guarantees for test data. For example, FARE [33] provides practical certificates that serve as high-confidence upper bounds on downstream unfairness. Different from these methods, our framework provides an explicit way to *control both the confidence level* and *error threshold* for all downstream models, and yields tighter empirical bounds (Section 6).

Furthermore, in this study we focus on group fairness [19, 27], one of the most widely used fairness measures, including in legal setting, e.g., in the New York City Local Law 144 on Automated Employment Decision Tools and in the EEOC's rule of 80% hiring rates across sensitive groups. Some prior works [40, 41, 59, 64] focus on another important measure, i.e., individual fairness, without providing high confidence guarantees. Finally, some prior work [28, 43, 70] provides high-confidence guarantees for fair classification, but does not explore representation learning.

3 Preliminaries

We first introduce notations for representation learning and the unfairness measure we focus on. Let X be a random variable denoting the *feature vector*, and S a random variable denoting *sensitive attributes*. $D \coloneqq \{(X_i, S_i)\}_{i=1}^n$ denotes a dataset with i.i.d. data samples, where each (X_i, S_i) has the same joint distribution as (X, S). Let $\mathcal{D}$ be the set of all D's, $\phi \in \Phi$ be the *representation model parameters*, and q_ϕ be the *representation model* parameterized by ϕ. We define Z as the *representation* for (X, S) where $Z \sim q_\phi(\cdot|X, S)$ and $Z \in \mathbb{R}^l$.

The learned representation will be used for subsequent supervised learning *downstream* tasks. We denote the *label* for such a downstream task as the random variable Y. The objective in a downstream task is to predict Y given (X, S). It is common to use Z in place of (X, S) as input to a *downstream model* $\tau : \mathbb{R}^l \to \mathbb{R}$. Let $\hat{Y} \coloneqq \tau(Z)$ denote the prediction of Y by model τ. We call $\hat{Y}$ the *downstream prediction*. There can be multiple downstream tasks that make use of the same representation Z.

The goal is to learn a fair representation model that ensures a specified notion of fairness across downstream tasks and models. In this work, we focus on binary classification tasks[1] and a widely used group fairness objective called demographic parity (DP) [19]. The extension to Equal Oppertunity and Equalized Odds will be discussed in Appendix C. Below we formally define a measure of *demographic disparity* of how unfair a downstream model τ is under DP.

Definition 3.1 (Demographic disparity measure)**.** Let $\Delta_{\text{DP}}(\tau, \phi)$ represent the measure of unfairness in the downstream predictions $\hat{Y}$ produced by model τ when using representation parameters ϕ. Specifically, $\Delta_{\text{DP}}(\tau, \phi) \coloneqq |\Pr(\hat{Y} = 1|S = 1) - \Pr(\hat{Y} = 1|S = 0)|$.

For simplicity, we assume that Y and S are binary, and this definition can be generalized to non-binary settings. When S is non-binary, $\Delta_{\text{DP}}(\tau, \phi)$ is defined as the maximum absolute difference between the conditional probabilities, $\Pr(\hat{Y} = 1|S)$, with any pair of values of S [8] (Appendix A).

4 Problem Formulation

This section formulates the task of representation learning with high-confidence fairness guarantees.

[1] FRG can be easily extended to provide similar guarantees for non-binary classification and regression by limiting $\text{Cov}(S, Z)$. We focus on binary classification due to its prevalence in literature and legal systems.

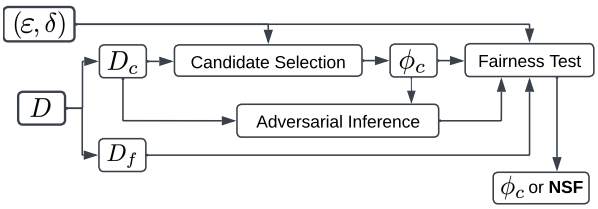

Figure 1: An overview of the FRG framework. Given a dataset D, with probability $1-\delta$, FRG generates an "ε-fair" representation model, or returns NSF if such a model cannot be found. See Section 5 for discussion.

A fair representation model should ensure with high confidence that the representations it generates will not lead to unfairness for downstream tasks. Specifically, a representation model is fair if and only if it results in fair predictions (as defined in Def. 3.1) for every possible downstream model and downstream task. That is, for all downstream tasks and all τ, $\Delta_{\text{DP}}(\tau, \phi)$ must be upper-bounded by a small constant, ε. We define an "ε-fair" representation model as follows.

Definition 4.1 ("ε-fair" representation model)**.** Representation model q_ϕ is ε-fair with parameter $\varepsilon \in [0, 1]$ if and only if $\Delta_{\text{DP}}(\tau, \phi) \leq \varepsilon$, for every downstream model τ and downstream task.

We define a representation learning algorithm $a : \mathcal{D} \to \Phi$ to be an algorithm that takes a data set as input and produces representation model parameters as output. In this paper, we aim to provide a representation learning algorithm such that any representation model it learns is guaranteed to be ε-fair under Def. 4.1, with high confidence. Such an algorithm has the following formal definition.

Definition 4.2 (A representation learning algorithm with high-confidence fairness guarantees)**.** Given $\varepsilon \in [0, 1], \delta \in (0, 1)$, and a dataset D, a representation learning algorithm a is said to provide a $1-\delta$ confidence ε-fairness guarantee if and only if $\Pr\left(g_\varepsilon(a(D)) \leq 0\right) \geq 1-\delta$, where $g_\varepsilon(\phi) \coloneqq \sup_\tau \Delta_{\text{DP}}(\tau, \phi) - \varepsilon$.

Observe that q_ϕ is an ε-fair representation model if and only if $g_\varepsilon(\phi) \leq 0$ (Def. 4.1). Therefore, any algorithm under Def. 4.2 guarantees that any representation model with parameters learned by this algorithm has at least $1-\delta$ probability to be an ε-fair representation model. Algorithms of this form can generally be categorized as *Seldonian* algorithms [70]. This guarantee implies that even in the worst case (when downstream models are adversarial), any resulting representation model should *not* fail the ε-fairness constraint with probability larger than δ.

Special case: unachievable ε-fair representation models. We note that in some scenarios, it may not be possible for any non-degenerate algorithm to ensure fairness with the specified confidence $1-\delta$, e.g., when ε, δ, and the amount of training data are all very small. In such cases, we allow the algorithm to output *No Solution Found* (NSF) as a way of indicating that it is unable to provide the required confidence that the learned representation will be fair given the amount of data it has been provided. To indicate that it is always fair for the algorithm to return NSF, we set $g_\varepsilon(\phi) = 0$. However, if an algorithm constantly returns NSF, it is of no value. We empirically evaluate the probability of returning a solution (i.e., not NSF) in Section 6.

5 Methodology

We now introduce our proposed framework, i.e., ***F**air **R**epresentation learning with high-confidence **G**uarantees* (FRG). It is the first representation learning algorithm that provides a *user-defined* high-confidence fairness guarantee. An overview of FRG is provided in Fig. 1.

FRG consists of three major components: *candidate selection*, *adversarial inference* and *fairness test*. First, we present a high-level summary of the algorithm before discussing each component in detail. FRG first splits the data D into disjoint sets, D_c and D_f. Candidate selection uses D_c to optimize and propose *candidate solution* ϕ_c. We train an adversarial model to predict sensitive attributes using the representations of D_c as learned by the model q_{ϕ_c}. The fairness test uses predictions made by the adversarial model on D_f to evaluate whether ϕ_c can satisfy $g_\varepsilon(\phi_c) \leq 0$ on future unseen data with sufficient confidence. Finally, FRG returns ϕ_c if the fairness test is passed and NSF otherwise. While it is the fairness test that ensures the high-confidence guarantees, an effective candidate selection that reasons about the fairness test procedure is needed to maximize the likelihood of passing the test.

In this section, we first define an optimal adversary and propose an effective optimization for the adversarial model to achieve an approximation of the optimum. We then introduce the fairness

test assuming the access to an oracle optimal adversary. Lastly, we provide details for a candidate selection that proposes candidates that aim to pass the fairness test and achieve high expressiveness.

5.1 Adversarial Inference

To learn an adversarial downstream model τ_{adv} that best predicts the sensitive attribute S, we generate the representations for D_c using a proposed candidate representation model q_{ϕ_c} as input to τ_{adv}. We define an optimal adversary τ^*_{adv} to be one that maximizes the Δ_{DP} (Def. 5.1), and prove in Theorem 5.2 that when both S and $\hat{Y}$ are binary, there exists a mapping between Δ_{DP} and the absolute covariance between $\hat{Y}$ and S (denoted as $|\text{Cov}(\hat{Y}, S)|$). The extension to non-binary sensitive attributes is provided in Appendix D.

Definition 5.1 (Optimal adversary)**.** Given a representation model q_ϕ, a downstream model τ^*_{adv} is an optimal adversary if and only if $\tau^*_{\text{adv}} \in \arg\max_\tau \Delta_{\text{DP}}(\tau, \phi)$.

Theorem 5.2. *Suppose $S, \hat{Y} \in \{0, 1\}$ are Bernoulli random variables. We have $\Delta_{\text{DP}}(\tau, \phi) = \frac{|\text{Cov}(\hat{Y},S)|}{\text{Var}(S)}$.* **Proof.** See Appendix B.

Following Theorem 5.2, we have $\tau^*_{\text{adv}} \in \arg\max_\tau |\text{Cov}(\hat{Y}, S)| = \arg\max_\tau \Delta_{\text{DP}}(\tau, \phi)$. Intuitively, the worst-case Δ_{DP} is achieved when the adversary is optimal either in predicting S or $1 - S$, and thus achieves either $\max_\tau \text{Cov}(\hat{Y}, S)$ or $\min_\tau \text{Cov}(\hat{Y}, S)$ (that are inversely correlated). In practice (Sec. 6), we find it sufficient to train an approximately optimal adversary to predict S based on Z using traditional gradient-based optimization strategies. We train τ_{adv} with representations and sensitive attributes from D_c, which will then be used by the fairness test to evaluate on D_f.

5.2 Fairness Test

The fairness test aims to evaluate whether a candidate solution ϕ_c induces a fair representation model with high confidence. In this section, we propose constructing a high-confidence upper bound on $g_\varepsilon(\phi)$, assuming access to an optimal adversary τ^*_{adv} (Def. 5.1), which in practice will be approximated by the adversarial model. If this high-confidence upper bound is at most zero, then we can conclude that $g_\varepsilon(\phi) \leq 0$ with confidence $1 - \delta$. We then detail the evaluation process for a candidate solution ϕ_c and show that it satisfies the $1 - \delta$ confidence ε-fairness guarantee.

5.2.1 $1 - \delta$ confidence upper bound on $g_\varepsilon(\phi)$

We define $U_\varepsilon : (\Phi, \mathcal{D}) \to \mathbb{R}$ to be such a function that produces a $1 - \delta$ confidence upper bound. Specifically, for $U_\varepsilon(\phi, D_f)$,

$$\Pr\Big(g_\varepsilon(\phi) \leq U_\varepsilon(\phi, D_f)\Big) \geq 1 - \delta. \tag{1}$$

The overall idea is to get unbiased estimates of $\Pr(\hat{Y} = y|S = s)$ of all combinations of y and s, construct confidence intervals on these probabilities, and then compose these intervals to form the confidence upper bound on g_ε. It takes two different approaches to compute U_ε when the sensitive attribute S is binary v.s. multiclass due to their different definitions of Δ_{DP} (Def. 3.1 and Def. A.1). We will provide the approach for multi-class S in Appendix F and focus on the binary case here.

We follow these steps. First, for each data point $(X_i, S_i) \in D_f$, we feed $Z_i = \phi(X_i, S_i)$ to the optimal adversary τ^*_{adv}, which aims to infer S, and get output $\hat{Y}_i$. We separate D_f into $D_{f,S=0}$ and $D_{f,S=1}$ such that all points in $D_{f,S=0}$ has $S = 0$ and all points in $D_{f,S=1}$ has $S = 1$. Each pair $((X_0^{(k)}, S_0^{(k)}), (X_1^{(k)}, S_1^{(k)}))$ can then be used to create a pair of unbiased estimates of $\Pr(\hat{Y} = 1|S = s)$ where $s \in \{0, 1\}$, denoted as $\hat{p}^{(k)}(1|s)$. Thus, m i.i.d. unbiased estimates of $\Pr(\hat{Y} = 1|S = s)$ can be obtained by sampling m pairs, i.e., $\mathbb{E}[\hat{p}^{(k)}(1|s)] = \Pr(\hat{Y} = 1|S = s)$ for any $k \in [1, ..., m]$. By linearity of expectation [70], they form m unbiased point estimates of $\Pr(\hat{Y} = 1|S = 0) - \Pr(\hat{Y} = 1|S = 1)$ which will be used to construct confidence intervals on $\Pr(\hat{Y} = 1|S = 0) - \Pr(\hat{Y} = 1|S = 1)$.

Second, we apply standard statistical tools such as Student's t-test [69], Hoeffding's inequality [29] etc., to construct a $1 - \delta$ confidence interval (CI) $[c_l, c_u]$ on $\Pr(\hat{Y} = 1|S = 0) - \Pr(\hat{Y} = 1|S = 1)$ using $\hat{p}^{(1)}(1|0) - \hat{p}^{(1)}(1|1), \ldots, \hat{p}^{(m)}(1|0) - \hat{p}^{(m)}(1|1)$. Finally, the $1 - \delta$ confidence upper bound of $g_\varepsilon = \sup_\tau |\Pr(\hat{Y} = 1|S = 0) - \Pr(\hat{Y} = 1|S = 1)| - \varepsilon$ can be obtained by taking $\max(|c_l|, |c_u|) - \varepsilon$.

Note that we use $\delta/2$ to obtain each of c_l and c_u to ensure that the bound on absolute value holds with probability at least $1 - \delta$ via the union bound.

While our framework is flexible to the techniques for achieving CIs, our experiments (Sec. 6) use the Student's t-test as it is well understood and used across the sciences for high-risk applications (e.g., biomedical research [50]). We include the procedure for constructing the confidence bounds with Student's t-test in Appendix G and more discussion on other variants of CIs in Appendix H.

5.2.2 Evaluation of candidate solutions

Suppose the fairness test gets a candidate solution ϕ_c and $U_\varepsilon(\phi_c, D_f) \leq 0$, it follows that there is at least confidence $1 - \delta$ that $g_\varepsilon(\phi_c) \leq 0$ (Inequality 1). Then, the fairness test concludes with at least $1 - \delta$ confidence that q_{ϕ_c} is an ε-fair representation model, and ϕ_c passes the test. If, however, $U_\varepsilon(\phi_c, D_f) > 0$, then the algorithm cannot conclude that $g_\varepsilon(\phi_c) \leq 0$ with high confidence. Therefore, the fairness test concludes that there is not sufficient confidence that q_{ϕ_c} is an ε-fair representation model, and ϕ_c fails the test.

Finally, if ϕ_c passes the fairness test, FRG outputs ϕ_c. Otherwise, it outputs `NSF`. When ϕ_c fails, we do not search for and test another representation model because this would result in the well-known "multiple comparisons problem." In this case, each run of the fairness test can be viewed as a hypothesis test for determining whether the representation is fair with sufficient confidence.

5.2.3 Theoretical Analysis

In this section, we prove that the fairness test with access to an optimal adversary (Def. 5.1) provides FRG with the desired high confidence ε-fairness guarantee, i.e., the probability that it produces a representation that is not ε-fair for every downstream task and model is at most δ.

Theorem 5.3. *Suppose fairness test finds $U_\varepsilon(\phi, D_f)$, a $1 - \delta$ confidence upper bound of $g_\varepsilon(\phi)$ for arbitrary ϕ, then FRG provides a $1 - \delta$ confidence ε-fairness guarantee.* **Proof.** See Appendix E.

5.3 Candidate Selection

Candidate selection searches for a representation model using D_c and proposes a candidate solution ϕ_c for the fairness test. Recall that the fairness test provides the desired $1 - \delta$ confidence ε-fairness guarantee (Def. 4.2) regardless of the choice of candidate selection (Theorem 5.3). However, candidate selection is considered ineffective if most of its proposed solutions fail the fairness test, which will lead to a high probability of returning `NSF`. In this section, we introduce an effective selection of candidates that are both likely to pass the fairness test and highly expressive.

5.3.1 Predicting Whether a Candidate Solution Will Pass the Fairness Test

Candidate selection proposes a candidate solution ϕ_c that it predicts will pass the fairness test. Such a prediction should leverage the knowledge of the exact form of the fairness test as much as possible, except using dataset D_c instead of D_f, i.e., checking whether $U_\varepsilon(\phi_c, D_c) \leq 0$. There are two differences in practice because the candidate selection repeatedly searches for the candidate solution.

First, for efficiency, we cannot fully optimize for an adversarial downstream model from scratch for every candidate searched. Thus, after initializing an adversary $\hat{\tau}_{\text{adv}}$, for each candidate searched, we take $t \in [1, 10]$ gradient steps (a hyperparameter) for optimizing $\hat{\tau}_{\text{adv}}$, without reinitialization.

Second, we repeatedly use the same dataset D_c to construct high confidence upper bounds and thus, may overfit to D_c, resulting in an overestimation of the confidence that the candidate solution will pass the fairness test and causing more `NSF`. One way to mitigate this issue is to inflate the confidence interval used in candidate selection. We multiply the confidence upper bound by α where $\alpha \geq 1$ is a hyperparameter, i.e., $\hat{U}_\varepsilon(\phi_c, D_c) \coloneqq \alpha U_\varepsilon(\phi_c, D_c)$ (Appendix L provides a case to show why such inflation could be critical in reducing the chance of getting `NSF`). We find $\hat{U}_\varepsilon(\phi_c, D_c)$, the inflated $1 - \delta$ confidence upper bound on $g_\varepsilon(\phi_c)$, following a similar procedure as Sec. 5.2.1, and use the constraint $\hat{U}_\varepsilon(\phi_c, D_c) \leq 0$ to find a candidate solution that reasons about the fairness test to increase the likelihood of passing.

5.3.2 Optimizing for a Candidate Solution With a Constrained Objective

In addition to candidate solutions that are likely to pass the fairness test, candidate selection also favors solutions that have high expressiveness, so that the representations they generate are effective for downstream tasks. We achieve this without being limited to a specific learning algorithm. We support most parameterized representation learning architectures proposed in previous work, including the

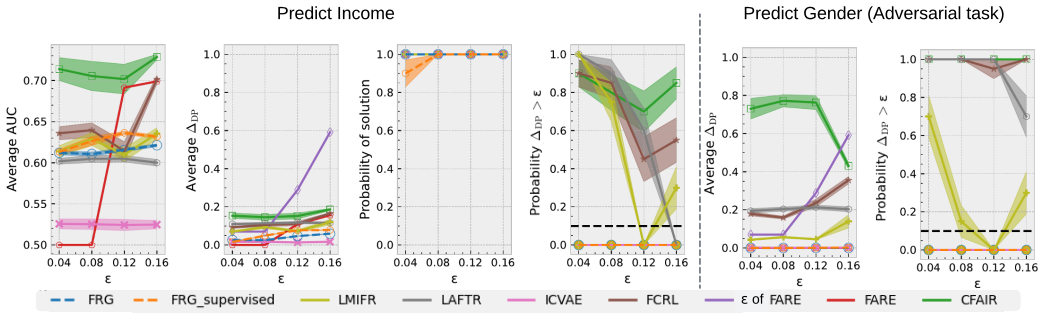

Figure 2: The evaluation on the **Adult** dataset. The target label is *income* and the sensitive attribute is *gender*. We vary $\varepsilon \in \{0.04, 0.08, 0.12, 0.16\}$. δ is fixed at 0.1. The four plots on the left are: (1) the average AUC; (2) the average Δ_{DP}; (3) the fraction of trials that returns a solution excluding `NSF`; (4) the fraction of trials that violates $\Delta_{\text{DP}}(\tau, \phi) \leq \varepsilon$ on the ground truth dataset. The AUC and Δ_{DP} on the adversarial task are on the right.

VAE-based methods [38, 47], contrastive learning methods [26, 55], etc. In our experiments, we focus on an adaptation of VAE [47] to construct the objective function that candidate selection optimizes. Specifically, we define $X \sim p_\theta(\cdot|Z, S)$ as the generative model for X with input (Z, S), parameterized by θ. Let $\mathbb{KL}$ denote KL-divergence, and $p(Z)$ be a standard isotropic Gaussian prior, i.e., $p(Z) = \mathcal{N}(0, \mathbf{I})$, where $\mathbf{I}$ is the identity matrix. Overall, we define the candidate selection process as approximating a solution to the constrained optimization problem:

$$\max_{\theta,\phi} \; \mathbb{E}_{q_\phi(Z|X,S)}\Big[\log p_\theta(X|Z,S)\Big] - \mathbb{KL}\Big(q_\phi(Z|X,S)\|p(Z)\Big) \qquad \text{s.t.} \quad \hat{U}_\varepsilon(\phi, D_c) \leq 0. \tag{2}$$

We propose using a gradient-based optimization to approximate an optimal solution (θ, ϕ). When gradient-based optimizers are used, the inequality constraint can be incorporated into the objective using the KKT conditions. That is, we find saddle-points of the following Lagrangian function:

$$\mathcal{L}(\theta, \phi; \lambda) \coloneqq -\mathbb{E}_{q_\phi(Z|X,S)}\Big[\log p_\theta(X|Z,S)\Big] + \mathbb{KL}\Big(q_\phi(Z|X,S)\|p(Z)\Big) + \lambda \hat{U}_\varepsilon(\phi, D_c), \tag{3}$$

where $\lambda \geq 0$ is a learnable Lagrange multiplier. Note that after each gradient step in the optimization, we need to update the adversary $\hat{\tau}_{\text{adv}}$ as mentioned above before evaluating $\hat{U}_\varepsilon(\phi, D_c)$.

This candidate selection procedure does not require any supervision. However, if a downstream task with labels is given, a supervised loss (e.g., binary cross-entropy) can be applied to $\mathcal{L}(\theta, \phi; \lambda)$ to improve the downstream predictions. We evaluate FRG both with and without supervision in Sec. 6.

6 Experiments

Here, we evaluate the performance and fairness of FRG, focusing on the following research questions. **RQ1:** Do the empirical results align with our expectation that FRG produces ε-fair representation models with high confidence? In other words, is Δ_{DP} of all downstream models and tasks upper-bounded by a desired ε with high probability? To address this question, we estimate the probabilities of violating the constraint $\Delta_{\text{DP}}(\tau, \phi) \leq \varepsilon$ using results from multiple runs of the algorithm with different training sets. **RQ2:** Can FRG learn expressive representations that are useful for downstream predictions? We evaluate the prediction performance on datasets with 2-3 downstream tasks using the area under the ROC curve (AUC), and compare its values across methods achieving similar demographic disparity bounds. **RQ3:** Would FRG frequently result in `NSF` to avoid unfairness even with sufficient data and reasonable values of ε and δ due to an ineffective candidate selection? To address this question, we evaluate the probability that FRG provides a solution other than `NSF`.

6.1 Experiment Setup

Datasets. We use three real-world datasets each with at least two downstream tasks, including the adversarial tasks that predict the sensitive attributes: UCI *Adult* [6] and *Income* (California only, commonly known as Retiring Adult) [17] both with *2* downstream tasks, and Heritage *Health* [34] with *3* downstream tasks. All downstream tasks and sensitive attributes are listed per dataset in Appendix Table 1, including their basic statistics (see details in Appendix I). For each dataset, we use the first downstream task for hyperparameter search and validation, and the last task is *adversarial task* predicting the sensitive attribute.

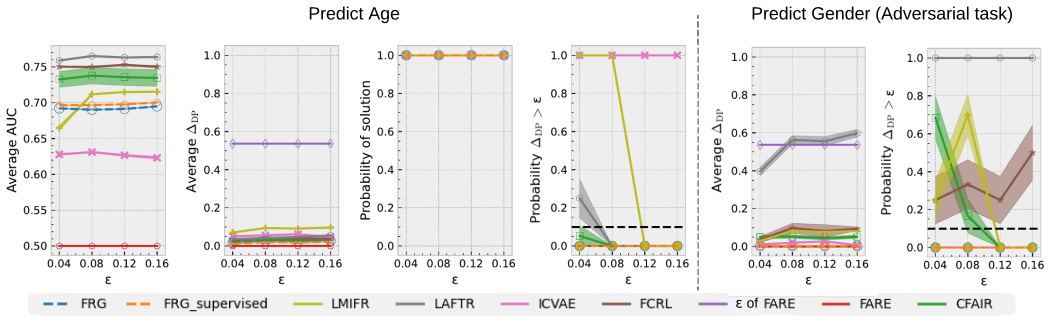

Figure 3: The evaluation on the **Health** dataset with sensitive attribute *gender*. The original target task predicting *Charlson Index* is in Appendix Figure 5. Here, we show the *transfer learning* capabilities on predicting *age*.

Baselines. We consider six competitive FRL baselines. *LAFTR* [48] proposes limiting the unfairness of arbitrary downstream classifiers with adversarial training. *ICVAE* [54] uses an upper-bound of the mutual information between S and $\hat{Y}$ as a regularizer. *LMIFR* [68] uses Lagrangian Multipliers to encourage a representation model to satisfy constraints that upper-bound the mutual information between S and $\hat{Y}$. *CFAIR* [76] adopts a balanced error rate (BER) and conditional learning to achieve parity. *FCRL* [26] proposes controlling parity via contrastive information estimators. *FARE* [33] provides high-confidence certificates with representations drawn from discrete distributions with finite support for downstream unfairness. All methods except FARE estimate upper bounds of Δ_{DP} on arbitrary downstream models using training data, which may not generalize to unknown test data.

For the Adult and Health datasets we train *LAFTR*, *FCRL*, *CFAIR*, and *FARE* in a supervised manner using the first downstream tasks (Appendix Table 1) because their model architectures rely on a supervised downstream task and avoiding it causes large performance decreases. *CVIB*, *LMIFR* are unsupervisedly trained without a labeled downstream task. For these two datasets, we train FRG with supervision (denoted as *FRG_supervised*) and without supervision (denoted as *FRG*). For the Income dataset, all models are trained with a supervised loss because the task is difficult for all models.

Evaluation process. For each dataset, we split the data into training (D_{train}), validation (D_{val}), and test (D_{test}) sets according to ratio 0.6:0.2:0.2. For FRG, we sample 10% of the training set to be D_f for fairness test and let candidate selection use the remaining 90% as D_c. We run experiments across 4-5 different ε's. We fix $\delta = 0.1$ for the main experiment, i.e., we want the probability of violating the constraint to be at most 0.1 and provide additional study on various δ's in Appendix Figure 8. In one experiment, we train all methods *20 times* with different randomly drawn training sets to get 20 representation models, which will then be applied to all downstream tasks. We report averages over the 20 trials for AUC, Δ_{DP}, the probability of returning a solution, and the probability of failing the constraint $\Delta_{\text{DP}}(\tau, \phi) \leq \varepsilon$. All figures plot the error bars evaluated with standard deviations.

Hyperparameter tuning. The goal of hyperparameter tuning is to find a set of parameters that achieve high downstream performance while satisfying $\Delta_{\text{DP}}(\tau, \phi) \leq \varepsilon$. Thus, we use validation sets of the *first* downstream tasks (Appendix Table 1) for tuning. We repeat the training for each parameter set at least 3 times. The first evaluation criterion is whether the constraint $\Delta_{\text{DP}}(\tau, \phi) \leq \varepsilon$ is satisfied. If finding a set of parameters that satisfies the constraint is impossible, we select the one that achieves the smallest Δ_{DP}. For FRG, we also prioritize the parameters that achieve the lowest probability of returning `NSF`. If there are ties, we choose the parameter set that achieves the highest average AUC. We note that the architectures and hyperparameters for the downstream models are consistent across methods for fair comparison. More details are provided in Appendix J.

6.2 Result and Discussion

The experiment results for the three datasets are provided in Figures 2, 3, and 4. For Health, the evaluation of the targeted downstream task is provided in Appendix Figure 5.

Overall, both FRG and FRG_supervised can maintain $\Delta_{\text{DP}} \leq \varepsilon$ with a sufficiently high probability (at least 0.9). In contrast, most baseline methods cannot consistently satisfy ε-fairness with high probability across all datasets. For baseline methods, we observe that a smaller ε causes a larger probability of failing the constraint. They also tend to fail on the adversarial tasks, in contrast to FRG (rightmost panels in Figures 2, 3, and 4). Thus, FRG provides high-confidence fairness guarantees across tasks for different ε's (**RQ1**). To answer **RQ2**, we highlight that FRG can match or outperform

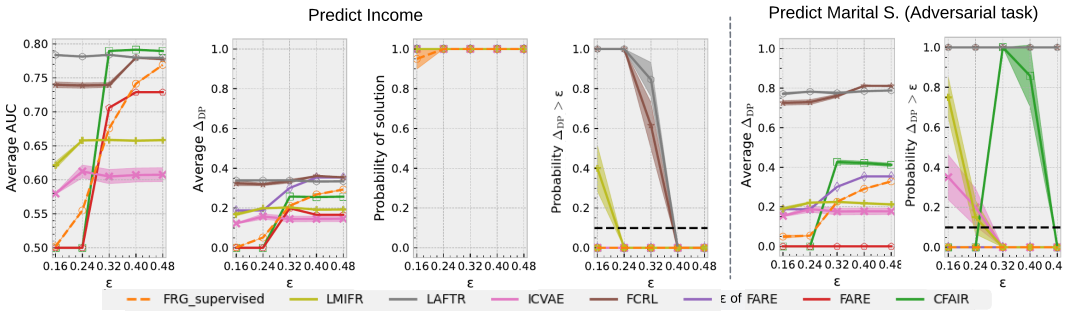

Figure 4: The evaluation on the **Income** dataset. The target label is *income* and the sensitive attribute is *marital status* which has 5 classes. We vary $\varepsilon \in \{0.16, 0.24, 0.32, 0.40, 0.48\}$.

baselines in terms of prediction performance (AUC). Compared with baselines that achieve $\Delta_{\mathbf{DP}} \leq \varepsilon$ with high probability (*ICVAE* for Adult and Health, and *FARE*), FRG tends to yield higher AUC, especially when ε is small. Further comparisons of the tradeoff between AUC and $\Delta_{\mathbf{DP}}$ in Appendix K confirm that FRG tends to achieve the highest AUC among the methods that achieve $\Delta_{\mathbf{DP}} \leq \varepsilon$ with high probability. Furthermore, FRG also keeps a high probability of returning a solution (at least 0.9) over all datasets, which demonstrates the effectiveness of our candidate selection and addresses **RQ3**.

Comparison between FARE and FRG. Similar to FRG, FARE can also achieve high probability of satisfying the fairness constraint across all datasets. While FARE does not support user-defined ε's, we use hyperparameter search (as discussed) to manipulate its high probability upper-bound of Δ_{DP}. However, these certificates are loose and often several times larger than the desired ε (compare the "ε of FARE" with the x-axes in Figures 2, 3, and 4). Additionally, even through this hyperparameter search, FARE does not certify fairness with enough granularity, i.e., the same certificates are given to multiple ε's. Perhaps FARE's use of representations drawn from a discrete distribution with a finite support limits the representations' variability. In contrast, FRG provides the high confidence guarantees and achieves accurate downstream models even for small user-specified ε's.

On the primary downstream tasks (left Figures 2, 4, and 5). All methods use the first tasks in Table 1 as the *target task* for hyperparameter search such that the models do not violate the fairness constraints on validation sets. Even though baseline methods including LMIFR, LAFTR, FCRL and CFAIR can satisfy the constraints on the validation set, they still fail with a large probability on the test set for the same task. In contrast, FRG and FARE can still keep $\Delta_{\mathbf{DP}}$ low with a high probability. This suggests that their theoretical upper bounds estimated using the training dataset overfit. Even with a hold-out validation set, it is still possible to underestimate the $\Delta_{\mathbf{DP}}$ for new test datasets. In comparison, FRG's high-confidence guarantees include statistical testing with held-out data, D_f, which automatically tests for and avoids unfair models resulting from overfitting.

On the adversarial tasks (right Figures 2, 3, and 4). Here, we check whether the learned representations can be used to adversarially infer sensitive attributes. Compared to other downstream tasks, the failure rates have increased for baseline methods (including LMIFR, LAFTR, FCRL, ICVAE, CFAIR), especially when ε is small. Even though these methods provide an estimated upper bound for Δ_{DP} in the worst case, the fairness constraints are still violated on these adversarial tasks due to the lack of high-confidence guarantees. However, FRG and FARE satisfy the constraints for unknown adversarial tasks, and FRG provides high-confidence *user-defined* fairness guarantees.

On transfer learning (left Figure 3). We first note that we use unsupervised learning for FRG, LMIFR and ICVAE on the Adult and Health datasets. So their performances on income prediction for Adult (left Fig. 2) and Charlson Index prediction (Appendix Fig. 5) for Health can also be used to demonstrate their transferability. Here, we focus on left Fig. 3 where task-specific labels are not exposed during training to all methods. Several baselines (LMIFR, ICVAE and LAFTR) increase their probability of violating the constraints compared to their performance on the target task that predicts Charlson Index. This may be the effect of overfitting the fairness constraint to a specific downstream task while not generalizing to all tasks. When the task is different, the sensitive information in the same representations can be exploited. Most supervised methods yield the lowest AUC scores (FARE, LAFTR, CFAIR). This may suggest that transferring the representations to a different task can hurt the prediction for supervised learning approaches.

Supervised v.s. unsupervised FRG. On the Adult and the Health datasets, although the supervised FRG performs slightly better than the unsupervised one, the improvement is insignificant. We hypothesize that while supervision helps, a more predictive candidate can also expose more sensitive information, making it easier to violate the fairness constraints. When the candidate selection aims to control $\hat{U}_\varepsilon(\phi, D_c) \leq 0$, the better-performing candidate may not be selected. In some cases when the candidate selection returns the better performing candidate, it can still fail the fairness test and be replaced by `NSF` if $U_\varepsilon(\phi_c, D_f) > 0$ (e.g., on the Adult dataset when ε is small).

We further study the effect of different confidence levels $\delta \in \{0.01, 0.05, 0.1, 0.15\}$ (Appendix Figure 8). The performance and the Δ_{DP} are similar when ε is small on both the target and adversarial tasks. On the target task, as ε gets large, e.g., $\varepsilon = 0.16$, the performance for the larger δ is marginally better than for the small δ while the Δ_{DP}'s are still similar. Overall, by increasing δ, one might gain a marginal improvement in the prediction accuracy or performance but at the cost of reducing the confidence in the fairness guarantees. Finally, we study varying α's in Appendix L. Other evaluation metrics including F1, Average Accuracy, Equal Opportunity Difference, Equalized Odds Difference are provided in Appendix Figures 9, 10, and 11.

7 Conclusion and Limitations

In this work, we introduced FRG, an FRL framework that provides high-confidence fairness guarantees, ensuring that demographic disparity for all downstream models and tasks is upper-bounded by a *user-defined* error threshold and confidence level. Our work is substantiated with theoretical analysis, and our empirical evaluation demonstrates FRG's effectiveness across various downstream tasks.

The theoretical guarantees of FRG make several assumptions. First, we assume all data samples are i.i.d. Second, the use of Student's t-test assumes the point estimates of g are normally distributed, which requires a large sample such that CLT holds. Third, we assume access to an optimal adversary (Def. 5.1) that uses representations as input to predict the sensitive attributes to maximize Δ_{DP}. We approximate it with an independently trained model.

In the future, FRG can be extended to provide guarantees related to measures of fairness, privacy [18], safety, robustness or concept erasure [7]. While FRG can account for label shift and concept drift because our guaranteed constraint on demographic disparity does not require any assumptions about the distribution of the downstream labels, future work could study guarantees under distributional shifts in features X and/or sensitive attributes S. Future work could also consider other approaches without assuming access to an optimal adversary (Appendix M).

8 Acknowledgments

This work is supported by the National Science Foundation under grant no. CCF-2018372, by a gift from the Berkeley Existential Risk Initiative, and by Rutgers SAS Research Grant in Academic Themes. Philip S. Thomas and Przemyslaw A. Grabowicz took similar advising roles on this project. The authors would also like to thank Linjun Zhang at Rutgers for the time to review and discuss this work.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, page 308–318, 2016. doi: 10.1145/2976749.2978318.

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. A reductions approach to fair classification. In *ICML*, 2018.

[3] T. W. Anderson. Confidence limits for the value of an arbitrary bounded random variable with a continuous distribution function. *Bulletin of The International and Statistical Institute*, 43: 249–251, 1969.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, May 2016. URL `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

[5] Mislav Balunović, Anian Ruoss, and Martin Vechev. Fair normalizing flows. In *ICLR*, 2022.

[6] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[7] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. In *NeurIPS*, 2023.

[8] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. Technical Report MSR-TR-2020-32, Microsoft, 2020. URL `https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/`.

[9] Nanette Byrnes. Artificial intolerance. *MIT Technology Review*, March 2016. URL `https://www.technologyreview.com/s/600996/artificial-intolerance/`.

[10] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, 2017.

[11] Flavio du Pin Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):1106–1119, 2018. doi: 10.1109/JSTSP.2018.2865887.

[12] Mattia Cerrato, Alexander Köppel, Marius Segner, and Stefan Kramer. Fair group-shared representations with normalizing flows. In *ICLR*, 2021.

[13] Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, and Mattia Samory. Semeval-2022 task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106, 2022.

[14] Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 1436–1445. PMLR, 2019.

[15] Namrata Deka and Danica J. Sutherland. Mmd-b-fair: Learning fair representations with statistical testing. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, 2023.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.

[17] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *NeurIPS*, 2021.

[18] Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming*, pages 1–12. Springer Berlin Heidelberg.

[19] 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*, ITCS '12, page 214–226, 2012. ISBN 9781450311151.

[20] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *International Conference in Learning Representations*, 2016.

[21] Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In *EMNLP*, 2018.

[22] Rui Feng, Yang Yang, Yuehan Lyu, Chenhao Tan, Yizhou Sun, and Chunping Wang. Learning fair representations via an adversarial framework. In *CoRR*, 2019.

[23] Xavier Gitiaux and Huzefa Rangwala. Learning smooth and fair representations. In *AISTATS*, 2021.

[24] Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2357–2365. PMLR, 2019.

[25] Vincent Grari, Oualid El Hajouji, Sylvain Lamprier, and Marcin Detyniecki. Learning unbiased representations via rényi minimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, page 749–764, 2021.

[26] Umang Gupta, Aaron M Ferber, Bistra Dilkina, and Greg Ver Steeg. Controllable guarantees for fair outcomes via contrastive information estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, page 7610–7619, 2021.

[27] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[28] Austin Hoag, James E. Kostas, Bruno Castro da Silva, Philip S. Thomas, and Yuriy Brun. Seldonian toolkit: Building software with safe and fair machine learning. In *2023 IEEE/ACM 45th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, pages 107–111, 2023. doi: 10.1109/ICSE-Companion58688.2023.00035.

[29] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.2307/2282952.

[30] Max Hort, Zhenpeng Chen, J Zhang, Federica Sarro, and Mark Harman. Bias mitigation for machine learning classifiers: A comprehensive survey. *ArXiv*, abs/2207.07068, 2022.

[31] Ayush Jaiswal, Daniel Moyer, Greg Ver Steeg, and Premkumar AbdAlmageed, Wael andNatarajan. Invariant representations through adversarial forgetting. In *AAAI Conference on Artificial Intelligence*, 2020.

[32] Taeuk Jang, Hongchang Gao, Pengyi Shi, and Xiaoqian Wang. Achieving fairness through separability: A unified framework for fair representation learning. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, 2024.

[33] Nikola Jovanović, Mislav Balunović, Dimitar I. Dimitrov, and Martin Vechev. Fare: Provably fair representation learning with practical certificates. In *ICML*, 2023.

[34] Kaggle. Health heritage prize, 2012. URL `https://www.kaggle.com/c/hhp`.

[35] Peter Kairouz, Jiachun Liao, Chong Huang, Maunil Vyas, Monica Welfert, and Lalitha Sankar. Generating fair universal representations using adversarial models. In *IEEE Transactions on Information Forensics and Security*, volume 17, pages 1970–1985, 2022.

[36] Dongha Kim, Kunwoong Kim, Insung Kong, Ilsang Ohn, and Yongdai Kim. Learning fair representation with a parametric integral probability metric. In *ICML*, 2022.

[37] Jin-Young Kim and Sung-Bae Cho. Fair representation for safe artificial intelligence via adversarial learning of unbiased information bottleneck. In *SafeAI@ AAAI*, page 105–112, 2020.

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

[39] Matthäus Kleindessner, Michele Donini, Chris Russell, and Muhammad Bilal Zafar. Efficient fair pca for fair representation learning. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, 2023.

[40] Preethi Lahoti, Krishna P. Gummadi, and Gerhard Weikum. Operationalizing individual fairness with pairwise fair representations. In *Proceedings of the VLDB Endowment*, volume 13, pages 506 – 518, 2019. doi: 10.14778/3372716.3372723.

[41] Preethi Lahoti, Krishna P. Gummadi, and Gerhard Weikum. ifair: Learning individually fair data representations for algorithmic decision making. In *International conference on data engineering (icde).*, page 1334–1345, 2019.

[42] Junghyun Lee, Gwangsu Kim, Matt Olfat, Mark Hasegawa-Johnson, and Chang D. Yoo. Fast and efficient mmd-based fair pca via optimization over stiefel manifold. In *AAAI*, 2022.

[43] Puheng Li, James Zou, and Linjun Zhang. Fairee: Fair classification with finite-sample and distribution-free guarantee. In *CoRR*, 2022.

[44] Ji Liu, Zenan Li, Yuan Yao, Feng Xu, Xiaoxing Ma, Miao Xu, and Hanghang Tong. Fair representation learning: An alternative to mutual information. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 1088–1097, 2022.

[45] Shaofan Liu, Shiliang Sun, and Jing Zhao. Fair transfer learning with factor variational auto-encoder. In *Neural Processing Letters*, page 1–13, 2022.

[46] Francesco Locatello, Gabriele Abbati, Tom Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. In *NeurIPS*, 2019.

[47] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. In *4th International Conference on Learning Representations (ICLR)*, 2016.

[48] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3384–3393. PMLR, 2018.

[49] A. Maurer and M. Pontil. Empirical bernstein bounds and sample variance penalization. In *In Proceedings of the Twenty-Second Annual Conference on Learning Theory*, pages 115–124, 2009.

[50] John H McDonald. *Handbook of biological statistics*, volume 2. sparky house publishing Baltimore, MD, 2009.

[51] Daniel McNamara, Cheng Soon Ong, and Robert C. Williamson. Provably fair representations. *arXiv preprint arXiv:1710.04394*, 2017.

[52] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54:1–35, 2021.

[53] Claire Cain Miller. Can an algorithm hire better than a human? *The New York Times*, June 2015. URL `https://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html`.

[54] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[55] Changdae Oh, Heeji Won, Junhyuk So, Taero Kim, Yewon Kim, Hosik Choi, and Kyungwoo Song. Learning fair representation via distributional contrastive disentanglement. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 1295–1305, 2022.

[56] Luca Oneto, Michele Donini, Giulia Luise, Carlo Ciliberto, Andreas Maurer, and Massimiliano Pontil. Exploiting mmd and sinkhorn divergences for fair and transferable representation learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15360–15370, 2020.

[57] Luca Oneto, Michele Donini, Massimiliano Pontil, and Andreas Maurer. Learning fair and transferable representations with theoretical guarantees. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 30–39, 2020.

[58] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[59] Momchil Peychev, Anian Ruoss, Mislav Balunović, Maximilian Baader, and Martin Vechev. Latent space smoothing for individually fair representations. In *ECCV*, 2022. doi: 10.1007/978-3-031-19778-9_31.

[60] Tao Qi, Fangzhao Wu, Chuhan Wu, Lingjuan Lyu, Tong Xu, Zhongliang Yang, Yongfeng Huang, and Xing Xie. Fairvfl: A fair vertical federated learning framework with contrastive adversarial learning. *arXiv preprint arXiv:2206.03200*, 2022.

[61] Novi Quadrianto and Oliver Sharmanska, Viktoriia Thomas. Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[62] Miriam Rateike, Ayan Majumdar, Olga Mineeva, Krishna P. Gummadi, and Isabel Valera. Don't throw it away! the utility of unlabeled data in fair decision making. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, page 1421–1433, 2022.

[63] Proteek Chandan Roy and Vishnu Naresh Boddeti. Mitigating information leakage in image representations: A maximum entropy approach. In *CVPR*, 2019.

[64] Anian Ruoss, Mislav Balunovic, Marc Fischer, and Martin Vechev. Learning certified individually fair representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 7584–7596, 2020.

[65] Mhd Hasan Sarhan, Abouzar Navab, Nassir Eslami, and Shadi Albarqouni. Fairness by learning orthogonal disentangled representations. In *ECCV*, 2020.

[66] Xudong Shen, Yongkang Wong, and Mohan Kankanhalli. Fair representation: Guaranteeing approximate multiple group fairness for unknown tasks. In *CoRR*. abs/2109.00545, 2021.

[67] Changjian Shui, Qi Chen, Jiaqi Li, Boyu Wang, and Christian Gagné. Fair representation learning through implicit path alignment. In *International Conference on Machine Learning*, volume 162, 2022.

[68] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 89:2164–2173, 2019.

[69] Student. The probable error of a mean. *Biometrika*, 6.1:1–25, 1908. doi: 10.2307/2331554.

[70] Philip S. Thomas, Bruno Castro da Silva, Andrew G. Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468): 999–1004, November 2019. doi: 10.1126/science.aag3311.

[71] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. Semi-fairvae: Semi-supervised fair representation learning with adversarial variational autoencoder. *arXiv preprint arXiv:2204.00536*, 2022.

[72] Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. In *Advances in Neural Information Processing Systems*, 2017.

[73] Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *ICLR*, 2021.

[74] Meike Zehlike, Philipp Hacker, and Emil Wiedemann. Matching code and law: Achieving algorithmic fairness with optimal transport. *Data Mining and Knowledge Discovery, Forthcoming*, 2019.

[75] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML*, 2013.

[76] Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J. Gordon. Conditional learning of fair representations. In *International Conference on Learning Representations*, 2020.

[77] Wei Zhu, Haitian Zheng, Haofu Liao, Weijian Li, and Jiebo Luo. Learning bias-invariant representation by cross-sample mutual information minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, page 15002–15012, 2021.

A Δ_{DP} for Multi-class Sensitive Attributes

Definition A.1 (Δ_{DP} for multiclass S)**.** We define Δ_{DP} for multi-class sensitive attributes $S \in \mathcal{S}$ where $|\mathcal{S}| > 2$ as follows.

$$\Delta_{\text{DP}}(\tau, \phi) \coloneqq \max_{i,j \in \mathcal{S}} \Big| \Pr(\hat{Y} = 1 | S = i) - \Pr(\hat{Y} = 1 | S = j) \Big|. \tag{4}$$

It follows from the implementation by Bird et al. [8].

B Proof of Theorem 5.2

To prove $\Delta_{\text{DP}}(\tau, \phi) = \frac{|\text{Cov}(S, \hat{Y})|}{\text{Var}(S)}$, we first prove the following lemma. To simplify the notations in the proofs, we define $p_{a,b} \coloneqq \Pr[\hat{Y} = a, S = b]$ where $a, b \in \{0, 1\}$.

Lemma B.1. *Suppose $S, \hat{Y} \in \{0, 1\}$ are Bernoulli random variables.*

$$\begin{aligned} \text{Cov}(\hat{Y}, S) &= \Pr[\hat{Y} = 1, S = 1] \Pr[\hat{Y} = 0, S = 0] - \Pr[\hat{Y} = 1, S = 0] \Pr[\hat{Y} = 0, S = 1] && (5) \\ &= p_{1,1} p_{0,0} - p_{1,0} p_{0,1}. && (6) \end{aligned}$$

Proof.

$$\begin{aligned} \text{Cov}(\hat{Y}, S) &= \mathbb{E}[\hat{Y}, S] - \mathbb{E}[\hat{Y}]\mathbb{E}[S] && \text{(by definition of covariance)} && (7) \\ &= \sum_{\hat{y}, s} p_{\hat{y}, s} \cdot (\hat{y} \cdot s) - \Pr[\hat{Y} = 1] \Pr[S = 1] && && (8) \\ &= p_{1,1} - \Pr[\hat{Y} = 1] \Pr[S = 1] && && (9) \\ &= p_{1,1} - (p_{1,1} + p_{1,0})(p_{1,1} + p_{0,1}) && && (10) \\ &= p_{1,1} - p_{1,1} p_{0,1} - p_{1,1} p_{1,0} - p_{1,1} p_{1,1} - p_{1,0} p_{0,1} && && (11) \\ &= p_{1,1}(1 - p_{0,1} - p_{1,0} - p_{1,1}) - p_{1,0} p_{0,1} && && (12) \\ &= p_{1,1} p_{0,0} - p_{1,0} p_{0,1} && && (13) \end{aligned}$$

Completing the proof.

Theorem B.2 (Theorem 5.2 restated)**.** *Suppose $S, \hat{Y} \in \{0, 1\}$ are Bernoulli random variables. Then* $\Delta_{\text{DP}}(\tau, \phi) = \frac{|\text{Cov}(\hat{Y}, S)|}{\text{Var}(S)}$.

Proof.

$$\begin{aligned} \Delta_{\text{DP}}(\tau, \phi) &= \Big| \Pr(\hat{Y} = 1 | S = 1) - \Pr(\hat{Y} = 1 | S = 0) \Big| && \text{(By Def. 3.1)} && (14) \\ &= \Big| \frac{p_{1,1}}{\Pr(S = 1)} - \frac{p_{1,0}}{\Pr(S = 0)} \Big| && && (15) \\ &= \Big| \frac{p_{1,1} \Pr(S = 0) - p_{1,0} \Pr(S = 1)}{\Pr(S = 1) \Pr(S = 0)} \Big| && && (16) \\ &= \frac{1}{\text{Var}(S)} \Big| p_{1,1} p_{0,0} + p_{1,1} p_{1,0} - p_{1,0} p_{0,1} - p_{1,0} p_{1,1} \Big| && && (17) \\ &= \frac{1}{\text{Var}(S)} \Big| p_{1,1} p_{0,0} - p_{1,0} p_{0,1} \Big| && && (18) \\ &= \frac{|\text{Cov}(\hat{Y}, S)|}{\text{Var}(S)} && \text{(By Lemma B.1)} && (19) \end{aligned}$$

Completing the proof.

C Extension to Equal Opportunity and Equalized Odds

In this section, we demonstrate that FRG can be extended to other group fairness metrics beyond demographic disparity (Δ_{DP}), specifically, Equal Opportunity Difference (Δ_{EOP}) and Equalized Odds Difference (Δ_{EOD}), with the assumption that the downstream task's labels are available. We start with their formal definitions.

Definition C.1 (Equal Opportunity Difference)**.**

$$\Delta_{\text{EOP}}(\tau, \phi) \coloneqq |\Pr(\hat{Y} = 1|S = 1, Y = 1) - \Pr(\hat{Y} = 1|S = 0, Y = 1)| \tag{20}$$
$$= \Delta_{\text{DP}}(\tau, \phi|Y = 1). \tag{21}$$

Definition C.2 (Equalized Odds Difference)**.**

$$\Delta_{\text{EOD}}(\tau, \phi) \coloneqq \max_{y\in\{0,1\}} \left(\left| \Pr(\hat{Y} = 1|S = 1, Y = y) - \Pr(\hat{Y} = 1|S = 0, Y = y) \right| \right) \tag{22}$$
$$= \max \Big(\Delta_{\text{DP}}(\tau, \phi|Y = 1), \Delta_{\text{DP}}(\tau, \phi|Y = 0) \Big). \tag{23}$$

Here $\Delta_{\text{DP}}(\tau, \phi|Y = y)$ for $y \in \{0, 1\}$ represents the Δ_{DP} under the conditional distribution $(X, S|Y = y)$. Empirically, it is the Δ_{DP} evaluated on the data samples whose downstream labels satisfy $Y = y$.

Thus, following the same procedure introduced in Sec. 5 on the data samples whose $Y = 1$, FRG can produce representation models that satisfy $\Delta_{\text{EOD}}(\tau, \phi) = \Delta_{\text{DP}}(\tau, \phi|Y = 1) \leq \varepsilon$ with probability at least $(1 - \delta)$.

By splitting δ in half, FRG (Sec. 5) can generate a representation model that satisfies $\Delta_{\text{DP}}(\tau, \phi|Y = 1) \leq \varepsilon$ and $\Delta_{\text{DP}}(\tau, \phi|Y = 0) \leq \varepsilon$, each with probability at least $(1 - \delta/2)$. By union bound, such a representation model satisfies $\Delta_{\text{EOD}}(\tau, \phi) = \max \Big(\Delta_{\text{DP}}(\tau, \phi|Y = 1), \Delta_{\text{DP}}(\tau, \phi|Y = 0) \Big) \leq \varepsilon$ with probability at least $(1 - \delta)$.

D The relationship between Δ_{DP} and $|\text{Cov}(\hat{Y}, S)|$ for non-binary sensitive attributes

The standard definition of covariance is not applicable to non-binary categorical random variables like the sensitive attributes. The reason is that the covariance takes the numerical values of the random variable into account but the numerical values of the sensitive attribute has no actual meaning. However, we can define auxiliary random variables for S for each pair of sensitive categories $i, j \in \mathcal{S}$, to represent S as a set of binary indicator variables, such that covariance can be applied. This approach enables the application of FRG to the setting of non-binary sensitive attributes S.

Suppose $S \in \mathcal{S}$ where $|\mathcal{S}| > 2$. Create one indicator variable $S'_{i,j}$ for each pair of $i, j \in \mathcal{S}$ where $i \neq j$ such that $S'_{i,j} = 0$ if $S = i$ and $S'_{i,j} = 1$ if $S = j$. We denote $p_{a,b} \coloneqq \Pr[\hat{Y} = a, S = b]$ where $a \in \{0, 1\}$ and $b \in \mathcal{S}$. We will first prove the lemma below before stating the main theorem.

Lemma D.1. $\text{Cov}(\hat{Y}, S'_{i,j}|S \in \{i, j\}) = \frac{p_{1,j}p_{0,i} - p_{1,i}p_{0,j}}{(\Pr(S=i)+\Pr(S=j))^2}$.

Proof. Following the same proof as Lemma B.1 in Appendix B, we have

$$\text{Cov}(\hat{Y}, S'_{i,j}|S \in \{i, j\}) \tag{24}$$
$$= \Pr[\hat{Y} = 1, S'_{i,j} = 1|S \in \{i, j\}] \Pr[\hat{Y} = 0, S'_{i,j} = 0|S \in \{i, j\}] \tag{25}$$
$$- \Pr[\hat{Y} = 1, S'_{i,j} = 0|S \in \{i, j\}] \Pr[\hat{Y} = 0, S'_{i,j} = 1|S \in \{i, j\}] \tag{26}$$
$$= \Pr[\hat{Y} = 1, S = j|S \in \{i, j\}] \Pr[\hat{Y} = 0, S = i|S \in \{i, j\}] \tag{27}$$
$$- \Pr[\hat{Y} = 1, S = i|S \in \{i, j\}] \Pr[\hat{Y} = 0, S = j|S \in \{i, j\}]. \tag{28}$$

Since $\Pr[S \in \{i, j\}] = \Pr(S = i) + \Pr(S = j)$, $\text{Cov}(\hat{Y}, S'_{i,j}|S \in \{i, j\}) = \frac{p_{1,j}p_{0,i} - p_{1,i}p_{0,j}}{(\Pr(S=i)+\Pr(S=j))^2}$. □

Demographic disparity can be defined separately for each pair of sensitive categories, $i, j \in \mathcal{S}$, as $\Delta_{\text{DP}}^{i,j} = \Big|\Pr(\hat{Y} = 1|S = i) - \Pr(\hat{Y} = 1|S = j)\Big|$. Then, we can limit $\Delta_{\text{DP}} = \max_{i,j} \Delta_{\text{DP}}^{i,j}$, to ensure demographic parity with respect to any pair of sensitive categories [8]. Next, we provide the relationship between Δ_{DP} and $\text{Cov}(\hat{Y}, S'_{i,j}|S \in \{i, j\})$.

Theorem D.2.

$$\Delta_{\text{DP}}(\tau, \phi) = \max_{i,j} \Big(2 + \frac{\Pr(S = i)}{\Pr(S = j)} + \frac{\Pr(S = j)}{\Pr(S = i)}\Big)\Big|\text{Cov}(\hat{Y}, S'_{i,j}|S \in \{i, j\})\Big|,$$

where $i, j \in \mathcal{S}$ and $i \neq j$.

Proof.

$$\Delta_{\text{DP}}(\tau, \phi) = \max_{i,j} \Big|\Pr(\hat{Y} = 1|S = i) - \Pr(\hat{Y} = 1|S = j)\Big| \tag{29}$$

$$= \max_{i,j} \Big|\frac{\Pr(\hat{Y} = 1, S = i)}{\Pr(S = i)} - \frac{\Pr(\hat{Y} = 1, S = j)}{\Pr(S = j)}\Big| \tag{30}$$

$$= \max_{i,j} \Big|\frac{\Pr(\hat{Y} = 1, S = i)\Pr(S = j) - \Pr(\hat{Y} = 1, S = j)\Pr(S = i)}{\Pr(S = i)\Pr(S = j)}\Big| \tag{31}$$

$$= \max_{i,j} \frac{1}{\Pr(S = i)\Pr(S = j)}\Big|p_{1,i}p_{0,j} + p_{1,i}p_{1,j} - p_{1,j}p_{0,i} - p_{1,j}p_{1,i}\Big| \tag{32}$$

$$= \max_{i,j} \frac{1}{\Pr(S = i)\Pr(S = j)}\Big|p_{1,i}p_{0,j} - p_{1,j}p_{0,i}\Big| \tag{33}$$

$$= \max_{i,j} \frac{(\Pr(S = i) + \Pr(S = j))^2}{\Pr(S = i)\Pr(S = j)}\Big|\frac{p_{1,i}p_{0,j} - p_{1,j}p_{0,i}}{(\Pr(S = i) + \Pr(S = j))^2}\Big| \tag{34}$$

$$= \max_{i,j} \Big(2 + \frac{\Pr(S = i)}{\Pr(S = j)} + \frac{\Pr(S = j)}{\Pr(S = i)}\Big)\Big|\text{Cov}(\hat{Y}, S'_{i,j}|S \in \{i, j\})\Big| \tag{35}$$

This completes the proof. Using this relationship, the optimal adversary can be approximated for non-binary sensitive features.

For instance, suppose that $\Pr(S = i) = \frac{1}{|\mathcal{S}|}$ for all i. Then, Δ_{DP} is minimized when the predicted label $\hat{Y}$ does not provide any information differentiating any pair of i and $j \in \mathcal{S}$. On the other hand, Δ_{DP} is maximized if there exists a pair of i and j such that $\hat{Y}$ is maximally correlated with S conditioning on $S \in \{i, j\}$.

E Proof of Theorem 5.3

Theorem E.1 (Theorem 5.3 restated)**.** *Suppose fairness test finds $U_\varepsilon(\phi, D_f)$, a $1 - \delta$ confidence upper bound of $g_\varepsilon(\phi)$ for arbitrary ϕ, then FRG provides a $1 - \delta$ confidence ε-fairness guarantee.*

Proof. By Def. 4.2, if FRG satisfies $\Pr\left(g_\varepsilon(a(D)) \leq 0\right) \geq 1 - \delta$, then FRG provides the desired $1 - \delta$ confidence ε-fairness guarantee. We prove by contradiction that $\Pr\left(g_\varepsilon(a(D)) \leq 0\right) \geq 1 - \delta$ if a corresponds to FRG and $a(D)$ corresponds to the representation model parameters returned by FRG when run on dataset D.

We begin by assuming the result is false and then derive a contradiction. The beginning assumption is that $\Pr\left(g_\varepsilon(a(D)) \leq 0\right) < 1 - \delta$. By contrapositive, we have $\Pr\left(g_\varepsilon(a(D)) > 0\right) \geq \delta$. By the construction of FRG, $a(D)$ is either `NSF` or the proposed candidate solution $\phi_c \in \Phi$. Notice that $g_\varepsilon(a(D)) > 0$ if and only if $a(D)$ does not return `NSF` but returns ϕ_c instead, i.e., $a(D) = \phi_c$. The fairness test in FRG returns ϕ_c if and only if $U_\varepsilon(\phi_c, D_f) \leq 0$ (Sec. 5.2). Therefore, the following events are equivalent ($\Pr(A, B)$ denotes the joint probability of A and B):

$$(g_\varepsilon(a(D)) > 0) \tag{36}$$

$$\iff (g_\varepsilon(a(D)) > 0, a(D) = \phi_c, U_\varepsilon(\phi_c, D_f) \leq 0) \tag{37}$$

$$\iff (g_\varepsilon(\phi_c) > U_\varepsilon(\phi_c, D_f), a(D) = \phi_c). \tag{38}$$

The joint event $(g_\varepsilon(\phi_c) > U_\varepsilon(\phi_c, D_f), a(D) = \phi_c)$ implies $(g_\varepsilon(\phi_c) > U_\varepsilon(\phi_c, D_f))$. Therefore,

$$\Pr\left(g_\varepsilon(\phi_c) \ > U_\varepsilon(\phi_c, D_f)\right) \geq \Pr\left(g_\varepsilon(\phi_c) > U_\varepsilon(\phi_c, D_f), a(D) = \phi_c\right) \geq \delta.$$

However, by construction of the fairness test and assuming the evaluated high confidence upper bound $U_\varepsilon(\phi_c, D_f)$ is correct, $\Pr\left(g_\varepsilon(\phi_c) \leq U_\varepsilon(\phi_c, D_f)\right) \geq 1-\delta$ (Inequality 1), which implies $\Pr\left(g_\varepsilon(\phi_c) > U_\varepsilon(\phi_c, D_f)\right) < \delta$. This gives a contradiction, completing the proof.

We note that this theorem is true for any choice of candidate selection, as the proof assumes the candidate solution ϕ_c is arbitrary.

F Computing $U_\varepsilon(\phi, D_f)$ for Multi-class Sensitive Attributes

Suppose $S \in \mathcal{S}$ where $|\mathcal{S}| > 2$. To evaluate an estimate of $U_\varepsilon(\phi, D_f)$, we need to find the worst-case Δ_{DP}.

We first feed $Z_i = \phi(X_i, S_i)$ for each data point in D_f to τ^*_{adv}. Since the adversarial predictor predicts the sensitive attribute, both the sensitive attribute and the label are multi-class. Thus, we obtain a predicted probability distribution of $\hat{Y}_i$ such that $\sum_{s\in\mathcal{S}} \Pr(\hat{Y}_i = s|Z_i) = 1$ (we apply *softmax* to the output of the multi-layer perceptron to get these probabilities). By splitting D_f into $|\mathcal{S}|$ groups where according to their sensitive attributes, we can get an unbiased estimate (denoted as $\hat{p}^{(i)}(s|j)$) of $\Pr(\hat{Y} = s|S = j)$ with a sample $(X_i, S_i = j, \hat{Y}_i = s)$.

Following a similar procedure in Sec. 5.2.1, we draw m unbiased estimates of $\Pr(\hat{Y} = s|S = j)$ for each $s, j \in \mathcal{S}$. Then we apply statistical tools (Student's t-test, for example) to construct a $1 - \delta/|\mathcal{S}|^2$ confidence interval (CI) $[c_l(s, j), c_u(s, j)]$ on $\Pr(\hat{Y} = s|S = j)$ with $\delta/(2|\mathcal{S}|^2)$ on both sides.

Finally, we set $U_\varepsilon(\phi, D_f) = \max_s(\max_j(c_u(s, j)) - \min_k(c_l(s, k))) - \varepsilon$.

We prove the correctness below.

Theorem F.1. *Suppose* $\Pr(\hat{Y} = s|S = j)$ *has* $1 - \delta/|\mathcal{S}|^2$ *confidence interval* $[c_l(s, j), c_u(s, j)]$ *with the confidence equally split on both sides for each* $s, j \in |\mathcal{S}|$*, and suppose* $\hat{Y}$ *is the optimal adversarial prediction that causes the maximum* Δ_{DP}*, then*

$$\Pr[g_\varepsilon(\phi) \leq \max_s(\max_j(c_u(s, j)) - \min_k(c_l(s, k))) - \varepsilon] \geq 1 - \delta.$$

Proof. Suppose $\Pr(\hat{Y} = s|S = j)$ has $1 - \delta/|\mathcal{S}|^2$ confidence interval $[c_l(s, j), c_u(s, j)]$ with the confidence equally split on both sides, then $\Pr[\Pr(\hat{Y} = s|S = j) \leq c_l(s, j)] \leq \delta/(2|\mathcal{S}|^2)$ and $\Pr[\Pr(\hat{Y} = s|S = j) \geq c_u(s, j)] \leq \delta/(2|\mathcal{S}|^2)$ for each $s, j \in |\mathcal{S}|$.

By the union bound, we have $\Pr[\min_k \Pr(\hat{Y} = s|S = k) \leq \min_k c_l(s, k)] \leq \delta/(2|\mathcal{S}|)$ and $\Pr[\max_j \Pr(\hat{Y} = s|S = j) \geq \max_j c_u(s, j)] \leq \delta/(2|\mathcal{S}|)$ for each $s \in \mathcal{S}$.

By the union bound, we have $\Pr[\max_j \Pr(\hat{Y} = s|S = j) - \min_k \Pr(\hat{Y} = s|S = k) \geq \max_j c_u(s, j) - \min_k c_l(s, k)] \leq \delta/|\mathcal{S}|$ for each $s \in \mathcal{S}$.

By the union bound again, we have

$$\Pr[\max_s(\max_j \Pr(\hat{Y} = s|S = j) - \min_k \Pr(\hat{Y} = s|S = k)) \geq \max_s(\max_j c_u(s, j) - \min_k c_l(s, k))] \leq \delta.$$

Assuming the adversary is optimal, we have

$$\begin{aligned}
g_\varepsilon(\phi) &= \sup_\tau \Delta_{\text{DP}}(\tau, \phi) - \varepsilon && (39)\\
&= \max_{j,k\in S} |\Pr(\hat{Y} = 1|S = j) - \Pr(\hat{Y} = 1|S = k)| - \varepsilon && (40)\\
&= \max_{j\in S} \Pr(\hat{Y} = 1|S = j) - \min_{k\in S} \Pr(\hat{Y} = 1|S = k) - \varepsilon && (41)\\
&\leq \max_{s\in S}(\max_{j\in S} \Pr(\hat{Y} = s|S = j) - \min_{k\in S} \Pr(\hat{Y} = s|S = k)) - \varepsilon, && (42)
\end{aligned}$$

where the predicted class s, the sensitive attribute classes j and k result in the worst-case differences over this group of samples.

Dataset	Sensitive (# group)	Downstream tasks	Size	Pr(S)
Adult	Gender (**2**)	Income, Gender	41K	**0.332**, 0.668
Health	Gender (**2**)	C.I., Age, Gender	55K	**0.553**, 0.447
Income	Marital Status (**5**)	Income, Marital Status	195K	**0.52**, 0.02, 0.09, 0.02, 0.35

Table 1: Summary of dataset statistics. For each dataset, the first task is used for hyperparameter search and validation, and the last task is an adversarial task that predicts the sensitive attribute. The **bold** in the last column indicates the fraction of positive labels for the adversarial tasks. C.I. stands for Charlson Index.

Then

$$\Pr[g_\varepsilon(\phi) \geq \max_s(\max_j c_u(s,j) - \min_k c_l(s,k)) - \varepsilon] \tag{43}$$

$$\leq \Pr[\max_{s\in S}(\max_{j\in S} \Pr(\hat{Y}=s|S=j) - \min_{k\in S} \Pr(\hat{Y}=s|S=k)) - \varepsilon \tag{44}$$

$$\geq \max_s(\max_j c_u(s,j) - \min_k c_l(s,k)) - \varepsilon] \tag{45}$$

$$\leq \delta. \tag{46}$$

Thus, $\Pr[g_\varepsilon(\phi) \leq \max_s(\max_j(c_u(s,j)) - \min_k(c_l(s,k))) - \varepsilon] \geq 1-\delta$. $\square$

G Using Student's T-test to Construct Confidence Intervals

We construct the $1-\delta$ confidence intervals of a random variable p using Student's t-test provided m samples $\hat{p}$ with the following steps: (1) Compute the sample mean $\bar{p} = \frac{1}{m}\sum_{k=1}^{m}\hat{p}^{(k)}$ where $k \in [1,\ldots,m]$; (2) Compute the sample standard deviation $\hat{\sigma} = \sqrt{\frac{1}{m-1}\sum_{k=1}^{m}(\hat{p}^{(k)}-\bar{p})^2}$; (3) Compute a $1-\delta/2$ confidence lower bound c_l and $1-\delta/2$ confidence upper bound c_u on $p_\varepsilon(\phi)$ using Student's t-test. That is $c_l = \bar{p} - \frac{\hat{\sigma}}{\sqrt{m}}t_{1-\delta/2,m-1}$ and $c_u = \bar{p} + \frac{\hat{\sigma}}{\sqrt{m}}t_{1-\delta/2,m-1}$ where $t_{1-\delta/2,m-1}$ is the $100(1-\delta/2)$ percentile of the Student's t-distribution with $m-1$ degrees of freedom. Student's t-test assumes the data to be normally distributed, and thus m needs to be sufficiently large for the guarantees to hold (following the central limit theorem (CLT)).

H Different Techniques for Obtaining Confidence Bounds

The statistical confidence interval method in our framework is modular such that researchers can substitute alternatives depending on their domain-specific needs.

We note that prior work, including the fairness experiments in Thomas et al. [70] employed the Student's t-test for similar guarantees and demonstrated that the t-test yielded sufficiently conservative failure rates in practice. In our work, we observed similar behavior across datasets. Other than the student's t-test, we have explored Hoeffding's inequality [29], but found its bounds to be overly conservative for our datasets, limiting its practical utility.

Different statistical tools for obtaining confidence bounds can have different tradeoffs. For instance: one could consider bounds based on the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality [3] which are less sensitive to distributional tails or empirical Bernstein bounds [49] which leverage sample variance. One could also explore bootstrap-based intervals, which empirically approximate sampling distributions.

We think that a rigorous comparison of confidence interval methods (e.g., parametric vs. non-parametric, bootstrap, etc.) is beyond the scope of this paper and merits its own study. Future work could systematically evaluate these alternatives to identify optimal bounds for specific applications.

I Datasets

The dataset statistics are listed in Table 1. The first dataset is the UCI *Adult* dataset [6],[2] which contains information of over 40,000 adults from the 1994 US Census. The sensitive attribute we consider is gender, and the targeted downstream task is to predict whether an individual earns more than \$50K/year.

The second dataset is Heritage Health [34].[3] The targeted downstream task is to predict Charlson Comorbidity Index, and we consider gender as a sensitive attribute. We include an additional

[2] `https://archive.ics.uci.edu/ml/datasets/Adult`

[3] `https://www.kaggle.com/c/hhp`

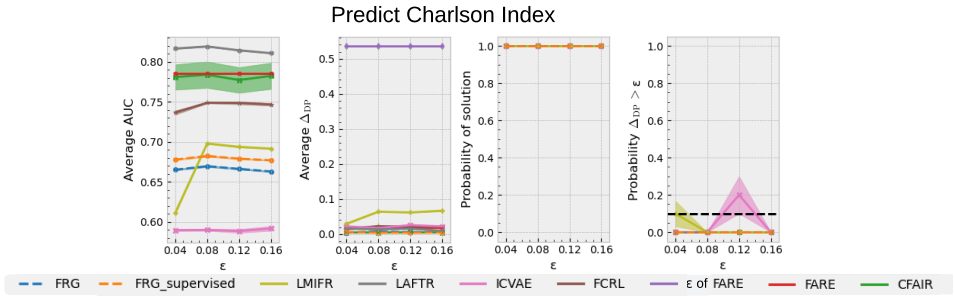

Figure 5: The evaluation on the Health dataset. The target label is Charlson Index. We vary $\varepsilon \in \{0.04, 0.08, 0.12, 0.16\}$. δ is fixed at 0.1.

downstream task that predicts age. Ages above 50 have positive labels, and ages below 50 have negative labels.

The third dataset we use is ACSIncome [17].[4] It includes data collected by the US Census across all states. We use the California dataset collected in 2018. The sensitive attribute is marital status, which has five classes: married, widowed, divorced, separated, and never married. The targeted downstream task is to predict whether an individual has income above $50,000.

J Hyperparameter Tuning

In our hyperparameter tuning process, we adjust various parameters, including the step sizes (for the primary objective, the Lagrange multipliers, and the adversarial predictor), the initial Lagrange multipliers, the weight of the regularizers, the number of epochs, etc. The primary objective of hyperparameter tuning is not only to find a set of hyperparameters for the algorithm that minimizes Δ_{DP}. Instead, our goal is to find hyperparameters that allow the algorithm to consistently provide a representation model that is ε-fair with as high expressiveness as possible. Thus, one should not tune hyperparameters separately for each of the training datasets we created. When we reuse the same training or validation set for hyperparameter search, we end up evaluating Δ_{DP} multiple times on the same training or validation set. As a result, Δ_{DP} evaluated on the model trained with the chosen hyperparameters may provide a biased estimation of Δ_{DP} on unseen future data. Consequently, the estimation of the probability $\Delta_{\text{DP}} \leq \varepsilon$ will also be biased. Therefore, we create additional datasets for hyperparameter tuning and adopt the same hyperparamters on different training datasets of the same size.

For baselines, we create validation sets by sampling 20% of the training data, while for FRG, we evaluate the models using the fairness test datasets (i.e., D_f in Sec. 5.2). We tune each algorithm with grid search according to the metrics evaluated on the validation sets (for baselines) or on the fairness test sets (for FRG). For the Health dataset, as there are multiple downstream tasks, we only assume the Charlson Index labels are available for hyperparameter tuning.

For FRG and FRG_supervised, we set the hyperparameter $\alpha = 2$ in the main experiment. We provide a study of various α's in Appendix L.

Note that we set the minimum allowed step size for the primary objective to 10^{-6} and the minimum number of epochs to 100. This choice is motivated by the fact that an algorithm with an excessively small step size may have minimal impact on optimizing the primary objective and could potentially produce random representations that lack utility for downstream predictions, despite being highly likely to be fair.

We also note that we use the same number of dimensions for representation Z ($Z = 50$ for Adult and Health and $Z = 100$ for Income) and the same hidden size for the downstream MLP for fair comparison. We use cross-entropy loss for all downstream models and Adam optimizer for all optimizations.

The detailed choices of hyperparameters for each of the datasets, the unfairness thresholds ε's, and the baselines are provided with config files in the source code.

[4] `https://github.com/socialfoundations/folktables`

K The Tradeoff Between the Prediction Performance and Fairness

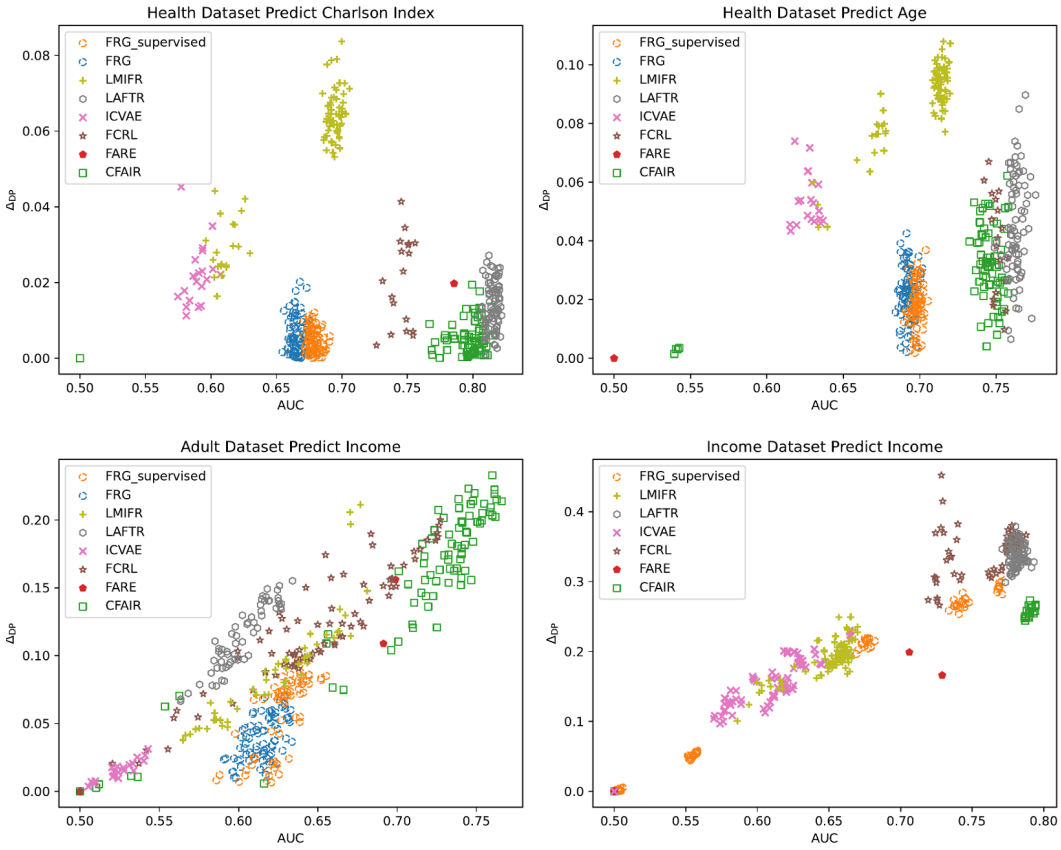

Figure 6: The tradeoff between AUC and Δ_{DP} on the Adult (bottom left), the Health (top left and top right) and the Income dataset (bottom right). FRG and FRG_supervised achieve the best tradeoff in Adult and are comparable to the best baselines in other datasets. Notice that even though some baselines achieve better tradeoff (e.g., FCRL, LMIFR, CFAIR in Health and CFAIR in income), they have high probabilities of violating the fairness constraints, especially in the adversarial downstream tasks (right Figures 3 and 4).

In Figure 6, the tradeoff between AUC and Δ_{DP} for each dataset is plotted. FRG and FRG_supervised achieve the best tradeoff in Adult and are comparable to the best baselines in other datasets. These results confirm that FRG and FRG_supervised are competitive in prediction performance while maintaining a low Δ_{DP}. Other than FRG and FRG_supervised, the baselines that achieve optimal tradeoff for one downstream task could achieve worse Δ_{DP} in the adversarial tasks. For example, FCRL, CFAIR and LAFTR in the Health dataset and CFAIR in the Income dataset achieve the best tradeoff in targeted downstream tasks, but violate the fairness constraints consistently in the adversarial task (right Figures 3 and 4). There are only 1-3 points for FARE because their use of discrete distribution with finite support lowers the variability of the representations, which is an issue discussed in Comparison between FARE and FRG in Section 6.2.

L Evaluating the impact of α

We highlight again that the choice of the confidence inflation hyperparameter α does not affect the validity of the high-confidence fairness guarantees. Here we evaluate different choices of α's in Figure 7 with $\delta = 0.1$. There are only minor differences in the performance and no impact on acceptance rates. Therefore, in this case, we may conclude that overfitting has not occurred. In our main experiment, we keep $\alpha = 2$ for all evaluations.

However, we think it is important to point out the potential overfitting if we do not inflate the confidence upper bound in candidate selection, especially when the constraint is restrictive. For example, when we set $\varepsilon = 0.035$ and $\delta = 0.01$, with results in Table 2, when α is closer to 1.0, it leads to a higher probability of returning `NSF`. This is the case when the candidate solution overfits

α	**Solution found**	**Avg. AUC**	**Avg. Δ_{DP}**
1.0	0.0	-	-
1.25	0.0	-	-
1.5	0.0	-	-
1.75	0.9	0.59	0.02
2.0	1.0	0.59	0.02
2.25	1.0	0.50	0.0
2.75	1.0	0.50	0.0
3.0	1.0	0.50	0.0

Table 2: On the Adult dataset with $\varepsilon = 0.035$ and $\delta = 0.01$.

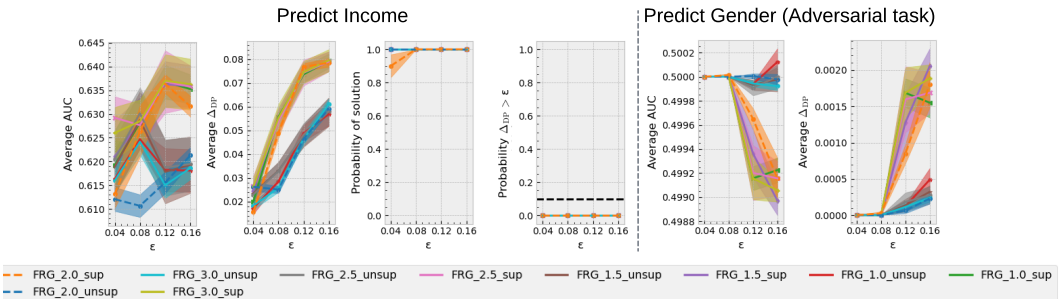

Figure 7: The study of the hyperparameter α on the Adult dataset with $\delta = 0.1$. We vary $\alpha \in 0.01, 0.05, 0.1, 0.15$. FRG_$\alpha$_sup and FRG_$\alpha$_unsup denotes FRG trained with and without supervision respectively.

the training data and overestimating the confidence that it can pass the fairness test. When α is larger it is more likely to return a solution. However, the performance can be affected because the candidate solution will be more conservative and may give a higher confidence than required to satisfy the fairness constraint.

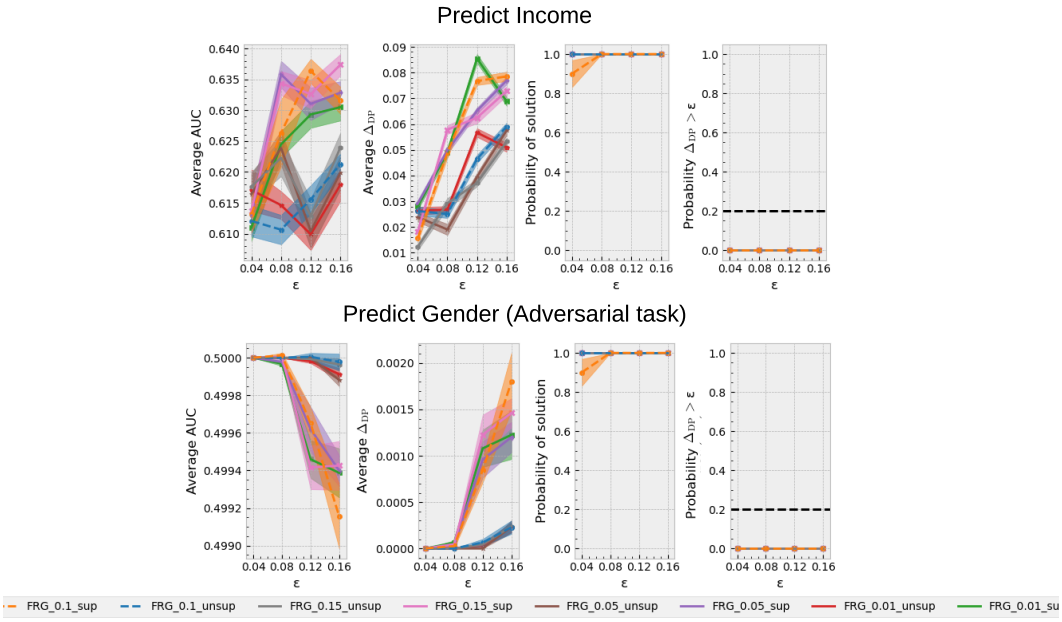

Figure 8: The study of the confidence level δ on the Adult dataset. We vary $\delta \in \{0.01, 0.05, 0.1, 0.15\}$. *FRG_δ_sup* and *FRG_δ_unsup* denotes FRG trained with and without supervision respectively.

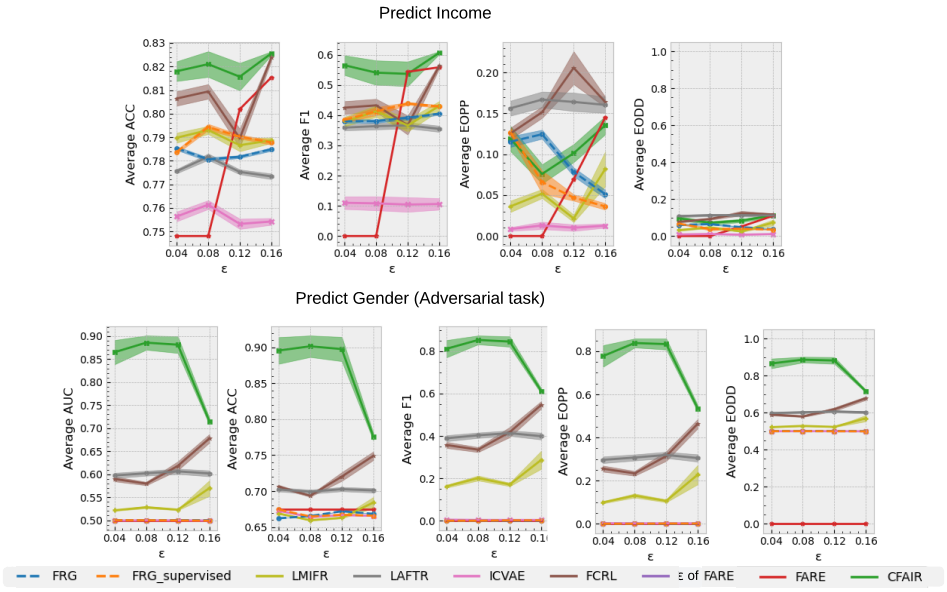

Figure 9: Additional metrics including F1, Average Accuracy (ACC), Equal Opportunity Difference (EOPP), Equalized Odds Difference (EODD) for the evaluation on the **Adult** dataset.

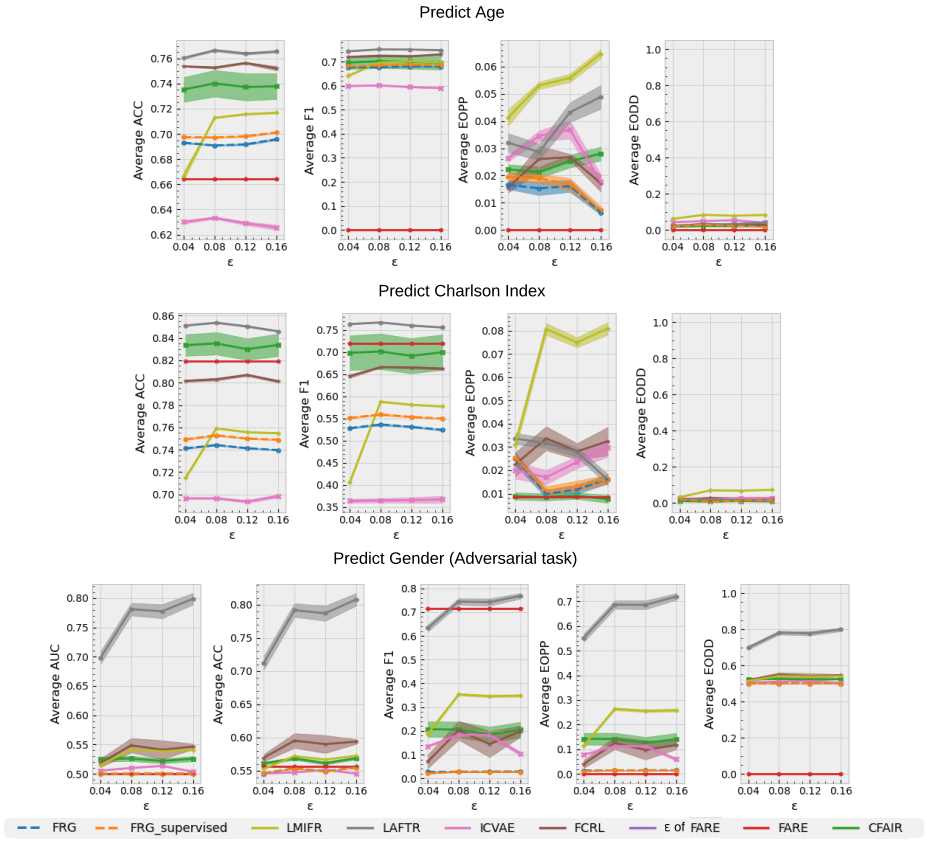

Figure 10: Additional metrics including F1, Average Accuracy (ACC), Equal Opportunity Difference (EOPP), Equalized Odds Difference (EODD) for the evaluation on the **Health** dataset.

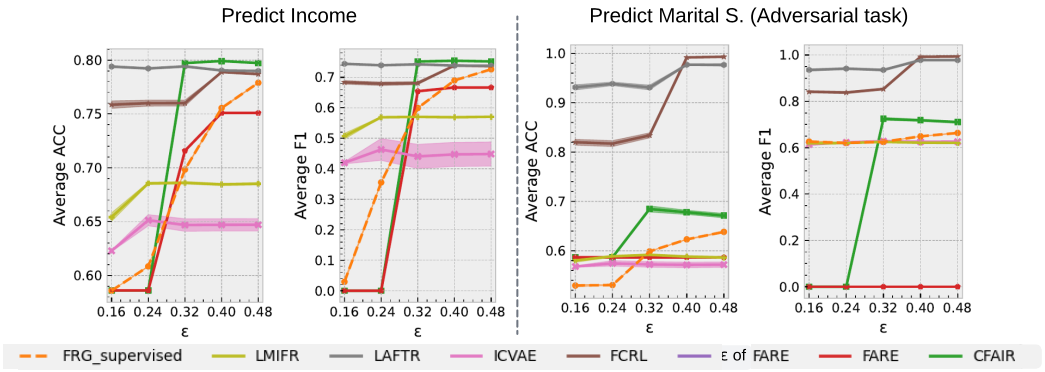

Figure 11: Additional metrics including F1, Average Accuracy (ACC) for the evaluation on the **Income** dataset.

M A Case Study: Using a Mutual Information-based Upper-bound for Constraining Δ_{DP} to Avoid the Assumption of Optimal Adversary

Different from the proposed method in the main text that uses the worst-case Δ_{DP} achieved by an oracle adversarial predictor to upper-bound Δ_{DP}, we provide an alternative method that uses a mutual information (MI)-based upper bound for constraining Δ_{DP}. Although the alternative method (named FRG-MI) does not rely on an oracle adversarial predictor for providing the guarantees, the upper-bound on Δ_{DP} is so loose that the method is shown impractical (the constraint is too conservative to provide good utility). In the following subsections, we will first introduce the MI-based bound on a strictly increasing convex function of Δ_{DP} as derived by [26] (Appendix M.1). As the alternative method shares similar components with FRG, including the candidate selection component and the fairness test component, we will document the changes to these components in Appendix M.2 and M.3 while referring to the main methods (Sec. 5) for the repeated details. In Appendix M.4, we prove that FRG-MI also provides a $1-\delta$ confidence ε fairness guarantees (Def. 4.2) as FRG does. In Appendix M.5, we evaluate FRG-MI empirically to show that it may not be suitable for practical use. Finally, we discuss several other alternatives one might consider for upper-bounding Δ_{DP} (Appendix M.6).

M.1 Mutual information bounds demographic parity

The demographic-parity-based measure (Def. 3.1) is specified for downstream models. Since want our representation model to guarantee fairness for every possible downstream model and downstream task, we consider using a mutual information-based upper-bound on Δ_{DP}. Gupta et al. [26] derived a bound for $\Delta_{\text{DP}}(\tau, \phi)$ that removes the dependency on the downstream model τ. Specifically, Gupta et al. [26] showed that the mutual information between the representation and the sensitive attributes, denoted by $I(Z;S)$, can be used to limit the demographic parity of downstream models.

Property M.1 (Relation of mutual information to $\Delta_{\text{DP}}(\tau, \phi)$). For all downstream models τ in all downstream tasks,

$$I(Z;S) \geq \psi(\Delta_{\text{DP}}(\tau, \phi)),$$

where ψ is a strictly increasing non-negative convex function derived by [26], and the details of which are in Appendix N.2. **Proof.** See the work of [26].

Notice that Property M.1 does not provide a direct upper bound on $\Delta_{\text{DP}}(\tau, \phi)$. Instead, it provides an upper bound on a strictly increasing non-negative convex function of $\Delta_{\text{DP}}(\tau, \phi)$. We use this property in the next section to guarantee fairness for representation models with high confidence.

M.2 The modified fairness test compared to Section 5.2

Different from Section 5.2, we now avoid the adversarial predictor but use Property M.4 to develop a high-confidence upper-bound for $g_\varepsilon(\phi)$. In this section, we first develop $\tilde{g}_\varepsilon(\phi)$ where $\tilde{g}_\varepsilon(\phi) \leq 0$ only if $g_\varepsilon(\phi) \leq 0$, and propose evaluating $\tilde{g}_\varepsilon(\phi) \leq 0$ to provably determine whether a representation model ϕ is ε-fair, i.e., $g_\varepsilon(\phi) \leq 0$. We then follow similar procedure as Sec. 5.2 to construct a high-confidence upper bound on $\tilde{g}_\varepsilon(\phi)$ instead of $g_\varepsilon(\phi)$. We follow Sec. 5.2 for the evaluation process for a candidate solution ϕ_c using this high-confidence upper bound.

A mutual-information-based evaluation. Our goal is to evaluate whether $g_\varepsilon(\phi) \leq 0$ with high confidence. However, estimating $g_\varepsilon(\phi) = \sup_\tau \Delta_{\text{DP}}(\tau, \phi) - \varepsilon$ is intractable because it requires knowledge of all downstream models and all downstream tasks. To remove the dependency on downstream models, we apply Property M.1, and evaluate whether $I(Z;S) - \psi(\varepsilon) \leq 0$ instead of $\sup_\tau \Delta_{\text{DP}}(\tau, \phi) - \varepsilon \leq 0$ (ψ is derived by Gupta et al. [26] and defined in Appendix N.2). Intuitively, when the mutual information between the representation and the sensitive attribute is low, it is hard for any model to predict S given Z with high accuracy. Therefore, any downstream model that does not explicitly aim to predict S is even less likely to take advantage of the sensitive attribute to produce unfair predictions. Theoretically, evaluating $I(Z;S) - \psi(\varepsilon) \leq 0$ can provably determine the ε-fairness of ϕ under Def. 4.1. We postpone the theoretical analysis to Appendix. M.4.

Unfortunately, computing $I(Z;S)$ is intractable because it requires marginalizing the joint distribution of (X, S, Z) over feature vector X, and so even this approach remains intractable. Multiple previous works have derived tractable upper bounds on $I(Z;S)$, which we discuss in detail in Appendix M.6.1. Let $\tilde{I}(Z;S)$ be one of these tractable upper bounds on $I(Z;S)$. Then, we define

$$\tilde{g}_\varepsilon(\phi) \coloneqq \tilde{I}(Z;S) - \psi(\varepsilon). \tag{47}$$

With this upper bound, we now evaluate the ε-fairness of ϕ by evaluating $\tilde{g}_\varepsilon(\phi) \leq 0$. In Lemma M.2, we prove if $\Pr\left(\tilde{g}_\varepsilon(a(D)) \leq 0\right) \geq 1-\delta$, then algorithm a provides the desired high-confidence fairness guarantee.

$1-\delta$ confidence upper bound on $\tilde{g}_\varepsilon(\phi)$. We follow two steps similar to Sec. 5.2 to compute a $1-\delta$ confidence upper bound on $\tilde{g}_\varepsilon(\phi)$. (1) Obtain m i.i.d. unbiased estimates $\hat{g}^{(1)}, \ldots, \hat{g}^{(m)}$ of $\tilde{g}_\varepsilon(\phi)$ using D_f, i.e., $\mathbb{E}[\hat{g}^{(j)}] = \tilde{g}_\varepsilon(\phi)$ for any $j \in [1, ..., m]$. (2) Apply standard statistical tools such as Student's t-test [69] or Hoeffding's inequality [29] to construct a $1-\delta$ confidence upper bound on $\tilde{g}_\varepsilon(\phi)$ using $\hat{g}^{(1)}, \ldots, \hat{g}^{(m)}$. We also use Student's t-test for our experiments (Appendix M.5).

Similar to Sec. 5.2, we define $U'_\varepsilon : (\Phi, \mathcal{D}) \to \mathbb{R}$ to be such a function that produces a $1-\delta$ confidence upper bound. Specifically, for $U'_\varepsilon(\phi, D_f)$, we have the following,

$$\Pr\Big(\tilde{g}_\varepsilon(\phi) \leq U'_\varepsilon(\phi, D_f)\Big) \geq 1-\delta. \tag{48}$$

The remaining steps for evaluating the candidate solution are equivalent to those of Sec. 5.2.

M.3 The modified candidate selection compared to Section 5.3

The candidate selection procedure is not changed from Section 5.3 except now it estimates the alternative high-confidence upper bound U'_ε (Def.48). We can also avoid the adversarial training process for estimating the upper bound as we can adopt the same procedure as in the fairness test.

M.4 Theoretical analysis

In this section we prove that FRG-MI is a representation learning algorithm that provides the desired high confidence ε-fairness guarantee, i.e., the probability that it produces a representation that is not ε-fair for every downstream task and model is at most δ.

We first prove in Lemma M.2 that if an algorithm a satisfies $\Pr\left(\tilde{g}_\varepsilon(a(D)) \leq 0\right) \geq 1-\delta$, then algorithm a provides the $1-\delta$ confidence ε-fairness guarantee described in Def. 4.2. We then prove in Theorem M.3 that FRG-MI indeed satisfies $\Pr\left(\tilde{g}_\varepsilon(a(D)) \leq 0\right) \geq 1-\delta$. Altogether, we can conclude that FRG guarantees with $1-\delta$ confidence that $\Delta_{\text{DP}}(\tau, a(D))$ is upper-bounded by ε for any τ (recall that here a corresponds to FRG-MI and $a(D)$ corresponds to the representation model parameters returned by FRG-MI when run on dataset D).

Lemma M.2. *If algorithm a satisfies $\Pr\left(\tilde{g}_\varepsilon(a(D)) \leq 0\right) \geq 1-\delta$, then algorithm a provides the $1-\delta$ confidence ε-fairness guarantee described in Def. 4.2.*

Proof. *Suppose $\Pr\left(\tilde{g}_\varepsilon(a(D)) \leq 0\right) \geq 1-\delta$. By Eq. 47, $\tilde{g}_\varepsilon(a(D)) = \tilde{I}(Z;S) - \psi(\varepsilon) \geq I(Z;S) - \psi(\varepsilon)$. By property M.1, $I(Z;S) \geq \sup_\tau \psi(\Delta_{DP}(\tau, a(D)))$. So, the event $(\tilde{g}_\varepsilon(a(D)) \leq 0)$ implies that$(I(Z;S) - \psi(\varepsilon) \leq 0)$, which further implies $(\sup_\tau \psi(\Delta_{DP}(\tau, a(D))) - \psi(\varepsilon) \leq 0)$. Using the fact that ψ is strictly increasing in $[0,1]$ (Appendix N.2), we have the following equivalent events:*

$$\left(\sup_\tau \psi(\Delta_{DP}(\tau, a(D))) - \psi(\varepsilon) \leq 0\right) \tag{49}$$

$$\iff \left(\psi(\sup_\tau \Delta_{DP}(\tau, a(D))) \leq \psi(\varepsilon)\right) \tag{50}$$

$$\iff \left(\sup_\tau \Delta_{DP}(\tau, a(D)) \leq \varepsilon\right) \tag{51}$$

$$\iff \left(\sup_\tau \Delta_{DP}(\tau, a(D)) - \varepsilon \leq 0\right) \tag{52}$$

$$\iff \left(g_\varepsilon(a(D)) \leq 0\right). \tag{53}$$

It follows that $\Pr\left(g_\varepsilon(a(D)) \leq 0\right) \geq \Pr\left(\tilde{g}_\varepsilon(a(D)) \leq 0\right) \geq 1-\delta$. So, FRG-MI (algorithm a) provides the desired $1-\delta$ confidence ε-fairness guarantee described in Def. 4.2, completing the proof.

Theorem M.3. *FRG-MI provides the $1-\delta$ confidence ε-fairness guarantee described in Def. 4.2.*

Proof. By Lemma M.2, if FRG-MI satisfies $\Pr\left(\tilde{g}_\varepsilon(a(D)) \leq 0\right) \geq 1-\delta$, then FRG-MI provides the desired $1-\delta$ confidence ε-fairness guarantee. Following the same proof for Theorem 5.3 in Appendix E, we can prove by contradiction that when a represents FRG-MI, $\Pr\left(\tilde{g}_\varepsilon(a(D)) \leq 0\right) \geq 1-\delta$.

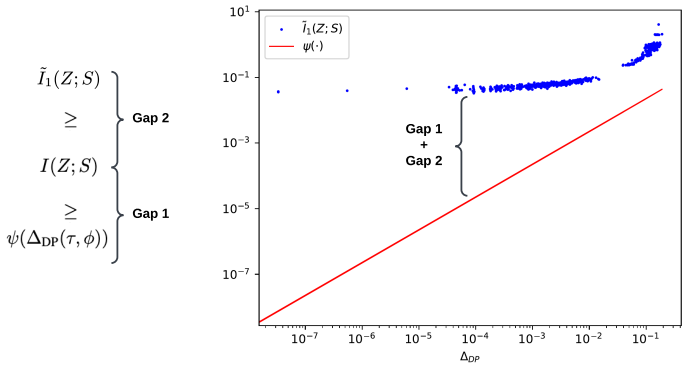

Figure 12: Using the *Adult* dataset (details in Appendix. I), we run L-MIFR [68] with different hyper-parameters to find representation models that achieve different $\Delta_{\text{DP}}(\tau, \phi)$. For each of the representation model, we record the corresponding tractable upper bound to $I(Z;S)$ by Song et al. [68, Section 2.2], denoted as $\tilde{I}_1(Z;S)$, and make the scatter plot in blue. We plot the function $\psi(\cdot)$ (Appendix N.2) in red. We highlight that there exists a gap between $\tilde{I}_1(Z;S)$ and $\psi(\Delta_{\text{DP}}(\tau,\phi))$, which consists of two gaps, $\tilde{I}_1(Z;S) - I(Z;S)$ and $I(Z;S) - \psi(\Delta_{\text{DP}}(\tau,\phi))$, and it can be observed empirically as shown by the plot. As Δ_{DP} decreases, the gap between $\tilde{I}_1(Z;S)$ and $\psi(\Delta_{\text{DP}}(\tau,\phi))$ increases.

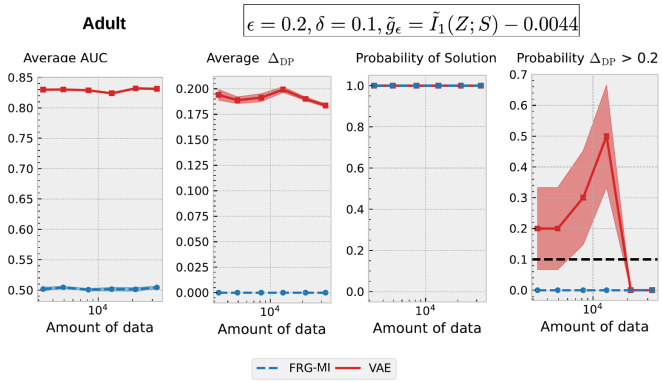

Figure 13: We give an example of employing FRG-MI to provide high-confidence fairness guarantees (Def. 4.2) on the *Adult* dataset, including VAE as a baseline.

M.5 Experiments

We first visualize in Fig. 12 and confirm that the gap between $\tilde{I}_1(Z;S)$ and $\psi(\Delta_{\text{DP}}(\tau,\phi))$ indeed exists empirically and the gap increases as Δ_{DP} approaches 0.

We then evaluate FRG-MI that provides high-confidence fairness guarantees (Def. 4.2) on the *Adult* dataset. For demonstration purposes, we select $\varepsilon = 0.2$ and $\delta = 0.1$, which means that FRG-MI guarantees with 90% confidence that downstream models do not violate $\Delta_{\text{DP}}(\tau,\phi) \leq 0.2$. It is worth noting that the selected $\varepsilon = 0.2$ is smaller than both the Δ_{DP} calculated with the true labels (0.26), and the upper bound on Δ_{DP} calculated with the prediction labels from a predictor that achieves equalized odds [76, Theorem 3.1]. We estimate $\Pr(S=1) \approx 0.668$ from the dataset, which yields $\psi(\varepsilon) \approx 0.0044$. We incorporate the constraint $\tilde{I}_1(Z;S) \leq \psi(\varepsilon)$ to guarantee ε-fairness with $1-\delta$ confidence ($\tilde{I}_1(Z;S)$ denotes the upper bound to $I(Z;S)$ as derived by Song et al. [68, Section 2.2]). We include a vanilla VAE without any fairness consideration as a baseline. The amount of training data used varies from 10%, 15%, 25%, 40%, 65% to 100% of the original data.

We show the result in Fig. 13. As demonstrated in the second and fourth plots, FRG-MI violates the constraint $\Delta_{\text{DP}}(\tau,\phi) \leq 0.2$ with a probability smaller than 0.1, whereas VAE violates the constraint with a probability larger than 0.1 when it uses less than 65% of the training data. According to the third plot, FRG-MI can also return solutions (i.e., not `NSF`) for all the trials.

Nonetheless, the constraint $\tilde{I}_1(Z;S) \leq \psi(\varepsilon)$ is overly conservative, which leads to relatively low AUC on average, as illustrated in the first plot. Additionally, the fourth plot demonstrates FRG-MI's ability to consistently keep Δ_{DP} near zero. Hence, applying an even stricter ε constraint on FRG-MI for high-confidence fairness guarantees is impractical and unnecessary.

We further analyze the gap between $I(Z;S)$ and $\psi(\sup_\tau \Delta_{\text{DP}}(\tau,\phi))$ in Appendix O, and the gap between $\tilde{I}_1(Z;S)$ and $I(Z;S)$ in Appendix P.

M.6 Other alternatives for upper-bounding Δ_{DP}

So far we have discussed using mutual information to upper bound Δ_{DP} (the violation of the demographic parity constraint), and ensure the ε-fairness of a representation model with high confidence (Sec. 5.2). Since $I(Z;S)$ is intractable, in Appendix M.6.1 we review four tractable upper bounds on $I(Z;S)$, and also discuss why in our experiments we adopt the first upper bound, $\tilde{I}_1(Z;S)$, derived by Song et al. [68, Section 2.2]. We then test whether $\tilde{I}_1(Z;S) \leq \psi(\varepsilon)$ to obtain the desired fairness guarantee (Eq. 47).

Because mutual information can be intractable, one might consider alternative methods for bounding Δ_{DP}. In Appendix M.6.2, we explore potential alternatives for upper bounding Δ_{DP} but find limitations that prevent the adoption of these methods in FRG.

M.6.1 The Tractable Upper Bounds on $I(Z;S)$

To our best knowledge, there are four tractable upper bounds on mutual information $I(Z;S)$ as derived by previous work [26, 54, 68]. Next, we discuss these approaches and their limitations. Although our general approach is compatible with any upper bound on mutual information, given the limitations of each method, we consider the first of the two approaches ($\tilde{I}_1(Z;S)$ below) by Song et al. [68] the most suitable in practice. Thus, we only adopt $\tilde{I}_1(Z;S)$ in our experiments.

Song et al. [68] proposed two upper bounds on $I(Z;S)$.

$\tilde{I}_1(Z;S)$: the first upper bound derived by [68, Section 2.2]. We denote the first upper bound as $\tilde{I}_1(Z;S)$ and $\tilde{I}_1(Z;S) \geq I(Z;X,S) \geq I(Z;S)$ [68, Section 2.2]. This is a theoretically guaranteed upper bound. We discuss the limitation of this upper bound in Appendix P that using this upper bound may diminish the expressiveness of the representations. However, we still find it effective for FRG to limit Δ_{DP} by ε in experiment (Sec. 6).

$\tilde{I}_2(Z;S)$: the second upper bound derived by [68, Section 2.3]. Song et al. [68] proposed a tighter upper bound compared to $\tilde{I}_1(Z;S)$, which we denote as $\tilde{I}_2(Z;S)$. However, it requires adversarial training, and the true upper bound can only be obtained when the adversarial model approaches global optimality. This is not ideal because if the adversarial model is under-performing, we may under-estimate the upper bound to $I(Z;S)$, and guaranteeing $\tilde{I}_2(Z;S) \leq \psi(\varepsilon)$ does not guarantee $I(Z;S) \leq \psi(\varepsilon)$ or ε-fairness. This result has also been confirmed by prior work including Elazar and Goldberg [21], Gupta et al. [26], Xu et al. [73] and Gitiaux and Rangwala [23].

$\tilde{I}_3(Z;S)$: the upper bound derived by Moyer et al. [54]. Moyer et al. [54] found $I(Z;S) = I(Z;X) - H(X|S) + H(X|Z,S)$ where H denotes entropy. They proposed ignoring the unknown positive constant term $H(X|S)$ and using the reconstruction error, i.e., $-\mathbb{E}_{q_\phi(Z|X,S)}\Big[\log p_\theta(X|Z,S)\Big]$ to be an upper bound of $H(X|Z,S)$ [54, Equations 2–7]. Let $\tilde{I}_3(Z;S) \coloneqq I(Z;X) - \mathbb{E}_{q_\phi(Z|X,S)}\Big[\log p_\theta(X|Z,S)\Big]$. Moreover, it can be difficult to estimate the gap $\tilde{I}_3(Z;S) - I(Z;S)$ because (1) $H(X|S)$ is hard to estimate; (2) $\tilde{I}_3(Z;S)$ is sensitive to the performance of the reconstruction model.

$\tilde{I}_4(Z;S)$: the upper bound derived by Gupta et al. [26]. Gupta et al. [26] observed that $I(Z;S) = I(Z;S|X) + I(Z;X) - I(Z;X|S)$. They then derived a lower bound for the term $I(Z;X|S)$ using constrative estimation so that $I(Z;S)$ can be upper-bounded. Specifically, they proved $I(Z;X|S) \geq \mathbb{E}_{p(X,Z,S)}\big[\log \frac{e^{f(X,Z,S)}}{\frac{1}{M}\sum_{m=1}^{M} e^{f(X_m,Z,S)}}\big]$, where $p(X,Z,S)$ is the joint distribution of (X,Z,S), $X_1,\cdots,X_M \sim p_{X|S}$, $p_{X|S}$ is the conditional distribution of X given S, and f is an arbitrary function [26, Proposition 5]. Since the distribution $p_{X|S}$ is unknown, the authors use the X,S pairs in the dataset as samples from this conditional distribution. When making a point estimate of the expectation, they use one sample from the dataset to evaluate the numerator, and use M samples from the same dataset to evaluate the denominator. This means that the estimation of

the expectation can be biased because the point estimates are not independent. Empirically, we also observe this issue and find that it tends to result in over-estimates of $I(Z;X|S)$ and under-estimates of $I(Z;S)$. Given how these terms are used in the expression for $I(Z;S)$, this results in bounds on mutual information that do not hold.

M.6.2 Alternative Methods for Upper-bounding Δ_{DP}

One might consider alternative methods for bounding Δ_{DP} because mutual information can be intractable and there can be a significant gap between mutual information and $\psi(\Delta_{\text{DP}})$ (that is, the upper bound can be loose). Several alternative methods have been proposed, which can provide bounds on Δ_{DP} using bounds on the total variation between the conditional distributions $p_{\tau,\phi}(\hat{Y}|S=0)$ and $p_{\tau,\phi}(\hat{Y}|S=1)$ [5, 48, 66, 76]. However, to our knowledge, there is not a known function such as ψ (Appendix N.2) that expresses the relation between the total variation and demographic parity, so total variation cannot be used to upper bound $\sup_\tau \Delta_{\text{DP}}(\tau,\phi)$ with a specific ε. In other work, Jovanović et al. [33, Section 5] proposed a practical certificate that upper bounds $\sup_\tau \Delta_{\text{DP}}(\tau,\phi)$. However, their method requires Z to be a discrete random variable, which is restrictive for general representation learning. Therefore, these methods are not suitable for our framework as they cannot be used to learn ε-fair representation models with a high-confidence guarantee.

N Details of Property M.1

Gupta et al. [26] has derived Property M.1 where $I(Z;S)$ is an upperbound for a strictly increasing non-negative convex function in Δ_{DP} of any τ, which we denote as ψ. Gupta et al. [26] has also found that when $I(Z;S)=0$, $\psi(\Delta_{\text{DP}}(\tau,\phi))=0$ and $\Delta_{\text{DP}}(\tau,\phi)=0$. We now define ψ in detail by first introducing a helper function f.

Definition N.1 (A helper function f)**.**

$$f(V)=\max\left(\log\left(\frac{2+V}{2-V}\right)-\frac{2V}{2+V},\frac{V^2}{2}+\frac{V^4}{36}+\frac{V^6}{288}\right).$$

with domain $V\in[0,2)$.

Definition N.2 (function ψ with parameter $\Delta_{\text{DP}}(\tau,\phi)$)**.** When S is binary, and f follows Def. N.1,

$$\psi(\Delta_{\text{DP}}(\tau,\phi))=(1-\pi)f(\pi\Delta_{\text{DP}}(\tau,\phi))+\pi f((1-\pi)\Delta_{\text{DP}}(\tau,\phi))$$

where $\pi=P_s(S=1)$ with P_s as the marginal distribution of $S\in\{0,1\}$.

When S is multinomial with K classes,

$$\psi(\Delta_{\text{DP}}(\tau,\phi))=f(\alpha\Delta_{\text{DP}}(\tau,\phi)),\alpha=\min_{k=1,\dots,K}\pi_k,$$

where $\pi_k=P_s(S=k)$ with P_s as the marginal distribution of $S\in\{1,\dots,K\}$.

O The Non-trivial Gap between $I(Z;S)$ and $\psi(\sup_\tau \Delta_{\text{DP}}(\tau,\phi))$

In this section, we analyze the non-trivial gap between $I(Z;S)$ and $\psi(\sup_\tau \Delta_{\text{DP}}(\tau,\phi))$ that makes it difficult for any algorithm to obtain ε-fairness.

As shown by Gupta et al. [26, Figure 6], there tends to be a significant gap between $I(Z;S)$ and $\psi(\sup_\tau \Delta_{\text{DP}}(\tau,\phi))$. Using their Figure 6 as an example, when $I(Z;S)\approx 0.035$, $\Delta_{\text{DP}}(\tau,\phi)\approx 0.15$ and $\psi(\Delta_{\text{DP}}(\tau,\phi))\approx 0.0025$. So, to ensure that $\Delta_{\text{DP}}(\tau,\phi)\leq 0.15$ with high confidence using the ψ-based bound on mutual information, one must ensure that $I(Z;S)\leq 0.0025$ with high confidence. However, in reality ensuring that $\Delta_{\text{DP}}(\tau,\phi)\leq 0.15$ only requires $I(Z;S)\leq 0.035$. Obtaining a solution that satisfies $I(Z;S)\leq 0.0025$ is far more difficult than obtaining a solution that satisfies $I(Z;S)\leq 0.035$, and hence using the ψ-based bound on mutual information results in exceedingly conservative bounds on Δ_{DP}.

P The Non-trivial Gap between $\tilde{I}_1(Z;S)$ and $I(Z;S)$

In this section we analyze the non-trival gap between $\tilde{I}_1(Z;S)$ and $I(Z;S)$ where $\tilde{I}_1(Z;S)$ (Appendix M.6.1) is one of the upper bounds to $I(Z;S)$ as derived by Song et al. [68, Section 2.2]. We begin by analyzing the gap between $I(Z;X,S)$ and $I(Z;S)$. $I(Z;X,S)-I(Z;S)=$

$H(Z|S) - H(Z|X,S) = H(X|S) - H(X|Z,S) = I(X;Z|S)$. This is the mutual information between X and Z given S, which is closely related to the primary objective we hope to maximize. Overall, we have the following:

$$\begin{aligned} I(Z;S) \leq & I(Z;X,S) && (54)\\ = & I(Z;S) + I(X;Z|S) && (55)\\ \leq & \tilde{I}_1(Z;S) && (56) \end{aligned}$$

Although using a constraint $\tilde{I}_1(Z;S) \leq \psi(\varepsilon)$ encourages both $I(Z;S)$ and $I(X;Z|S)$ to be small which seems to diminish the expressiveness of the representation model, we show empirically that it is effective for upper bounding mutual information and the Δ_{DP} of the downstream tasks with high probability in experiment (Sec. M.5).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We claimed that FRG provides high-confidence fairness guarantees for representation learning where both the threshold for unfairness and the confidence level are user-specified. This claims is justified with Sections 5 and 6.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We specified our limitations in Section 1, 5 and 7. Specifically The theoretical guarantees of FRG make several assumptions. First, we assume all data samples are i.i.d. Second, the use of Student's t-test assumes the point estimates of g are normally distributed, which requires a large sample such that CLT holds. Third, we assume access to an optimal adversary (Def. 5.1) that uses representations as input to predict the sensitive attributes to maximize Δ_{DP}. We approximate it with an independently trained model. Lastly, the guarantee is only applicable for controlling Δ_{DP} but not other fairness metrics such as Equal Opportunity or Equalized Odds.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide proofs for Theorems 5.2 and 5.3 in Appendix B and E. The assumptions are stated in the theorem statements.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the complete specification of our method in Section 5. Regarding the empirical evaluation, we provide detailed descriptions for the experiment setup, the baselines, the datasets, and the hyperparameter tuning in Section 6, Appendix I and J.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The source code is provided in an anonymous github repo with links provided in the abstract and in the supplementary material submitted. The datasets are open source. The instruction for downloading the data is provided in the README file in the source code.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All the training and test details (including data splits, hyper parameters, how they were chosen, type of optimizer, etc.) are clearly specified in Section 6 and Appendix J. The detailed choices of hyperparameters for each of the datasets, the unfairness thresholds ε's, and the baselines are provided with config files in the source code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We repeat each experiment at least 20 times. Each data point in the figures is accompanied by error bars. The error bars are calculated with the `.std()` function in the pandas library. We mentioned in Section 6 about how error bars are calculated.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The GPU used is one NVIDIA A16, and we use 128 CPUs with the model AMD EPYC 9354 32-Core Processor.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`.

   Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have sufficiently discussed both the potential positive societal impacts and negative societal impacts. The positive impacts are to provide high-confidence fairness guarantees for representation learning, which is critical for lots of applications (Section 1). However, there are limitations as we have discussed in Section 7, which may affect the effectiveness of the guarantees. For example, if the test data is under a distributional shift from the training data, the guarantees may no longer hold.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks because neither new data nor pretrained models are released.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The original papers that produced the code packages and datasets are properly cited. The URL to the datasets are included.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide our source code in an anonymized URL in our abstract and the supplementary material. The details about training, license, limitations, etc., are provided in the README file of the source code.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.