FairNVT: Dual-Level Fairness via Noise Injection in Vision Transformers

Anonymous authors
Paper under double-blind review

Abstract

This paper presents **FairNVT**, a lightweight debiasing framework for pretrained transformer-based encoders that improves both representation and prediction level fairness while preserving task accuracy. Unlike many existing debiasing approaches that address these notions separately, we argue they are inherently connected: suppressing sensitive information at the representation level can facilitate fairer predictions. Our approach learns task-relevant and sensitive embeddings via lightweight adapters, applies calibrated Gaussian noise to the sensitive embedding, and fuses it with the task representation. Together with orthogonality constraints and fairness regularization, these components jointly reduce sensitive-attribute leakage in the learned embeddings and encourage fairer downstream predictions. The framework is compatible with a wide range of pretrained transformer encoders. Across three datasets spanning vision and language, FairNVT reduces sensitive-attribute attacker accuracy, improves demographic-parity and equalized-odds metrics, and maintains high task performance.

1 Introduction

Modern machine learning systems largely follow a pretrain–transfer paradigm, where large foundation models are trained on massive, noisy datasets and then adapted to downstream tasks using their transferable representations. Yet these models often encode *social and demographic biases* present in their training data, leading to systematic unfairness across attributes such as gender, race, and age. When deployed in sensitive domains, such as recruitment, credit scoring, or facial recognition, these biases may propagate to downstream tasks, resulting in inequitable treatment of individuals and undermine the reliability of deploying machine learning systems (Gallegos et al., 2024; Li et al., 2024).

Fairness in machine learning is commonly studied at the *prediction-level*, where the goal is to ensure that model outputs satisfy group-based criteria such as demographic parity, equal opportunity, or equalized odds. These metrics quantify disparities in model predictions across demographic groups and are essential for evaluating the societal impact of deployed models. In the typical downstream pipeline, a classifier is trained on top of pretrained embeddings to perform the task while satisfying these fairness criteria. However, prediction-level fairness provides only a partial view of the problem. Even when a downstream classifier is trained to satisfy fairness metrics, the internal embeddings may still remain strongly correlated with sensitive attributes. In such cases, sensitive information can often be easily recovered from embeddings, enabling downstream misuse, model inversion, or fairness degradation when the representation is used in new tasks (Feng et al., 2023; Gallegos et al., 2024). These observations motivate *representation-level* fairness, which aims to learn embeddings that remain predictive for downstream tasks, while being invariant, or at least weakly informative, with respect to the sensitive attributes.

A growing body of work has explored both fairness notions, but typically treats them in isolation. Prediction-level methods directly impose fairness constraints on classifier outputs (Kang et al., 2022; Wang et al., 2023; Xie et al., 2024), while representation-level methods attempt to remove sensitive signals from learned embeddings using adversarial learning (Zhang et al., 2018; Götte, 2023), contrastive learning (Park et al., 2022), or projection-based techniques (Islam et al., 2024; Shi et al., 2024). Recent studies show that these

two notions of fairness are not trivially aligned and highlight the challenge of achieving both forms of fairness simultaneously (Shen et al., 2022).

In this work, we show that *enforcing representation-level fairness prior to downstream classifier training can lead to improved prediction-level fairness*. When the latent space is demonstrably purged of sensitive information, the downstream prediction head is forced to rely on task-relevant features, leading to a more robust and naturally fair decision-making process. Motivated by this view, we revisit representation-level fairness from a robust learning perspective and propose **FairNVT**, a lightweight debiasing framework for pretrained transformer-based encoders built on noise injection. By explicitly disentangling and "noising" the sensitive embedding subspace, we obfuscate the underlying signals that the classifier uses to make biased decisions while preserving task-relevant cues, effectively connecting robustness and fairness. Empirically, we demonstrate that FairNVT achieves a favorable fairness–utility balance: sensitive-attribute predictability from debiased embeddings is substantially reduced, while task accuracy remains high and prediction-level fairness is consistently improved. This dual-level fairness makes the learned representations more reliable for downstream use.

In summary, this paper makes the following three key contributions:

- We introduce a dual-level debiasing mechanism that injects calibrated noise into sensitive subspaces to suppress sensitive information at the representation level, leading to more fair downstream predictions.
- The proposed method is lightweight and integrates with pretrained frozen transformer encoders in an end-to-end, single-stage setup, disentangling task-relevant and sensitive components without requiring adversarial objectives or updates to the frozen backbone parameters.
- Extensive experiments across vision and language benchmarks demonstrate that our method improves both representation and prediction level fairness while maintaining competitive task performance.

2 Preliminaries: Fairness in Classification

We address fairness at two complementary levels: *prediction* and *representation*. At the **prediction level**, fairness requires making model outputs independent of the sensitive attribute S. For data-label pair (X, Y), a classifier $f : X \to \hat{Y}$ is considered fair if $P(\hat{Y} \mid S = s) = P(\hat{Y} \mid S = s')$, i.e., changes in S should not affect the predicted label. This notion of fairness has been widely studied in prior work, including Park et al. (2022); Tian et al. (2024); Park & Byun (2024), which analyze prediction-level fairness criteria in vision models. At the **representation level**, fairness enforces that learned embeddings $Z \coloneqq e(X)$ do not encode sensitive information: $I(Z; S) = 0$, where $I(\cdot\,;\cdot)$ denotes mutual information. This ensures that sensitive attributes cannot be inferred from internal representations. Representative approaches targeting this form of fairness include Ravfogel et al. (2020); Kumar et al. (2023). We discuss the detailed related works in Appendix A.

Intuitively, we could see the common condition for both notions of fairness to hold: if the encoded embedding is independent of the sensitive attribute, i.e. $Z \perp\!\!\!\perp S$, then $I(Z; S) = 0$, and $P(\hat{Y} \mid S = s) = P(\hat{Y} \mid S = s') = P(\hat{Y})$. Therefore, suppressing sensitive attribute signals in the embedding space encourages model to avoid encoding sensitive information and making biased predictions. We formally discuss such intuition in Appendix B. Next, we outline how our model supports these objectives.

3 FairNVT Framework

FairNVT mitigates bias in pre-trained embeddings and promotes fair predictions by injecting calibrated noise and optimizing with fairness-aware objectives. Figure 1 shows an overview of the proposed framework. Specifically, we attach light-weight adapters to extract task-relevant and sensitive information from the potentially biased embedding. Then, the sensitive embedding is perturbed with random, calibrated Gaussian noise and fused with the task embedding for downstream classification. The model is jointly optimized with classification, orthogonality, and fairness losses to balance multiple objectives. Section 3.1 discusses the key components of the model. Section 3.2 introduces the optimization objectives, and 3.3 describes the training and inference procedures.

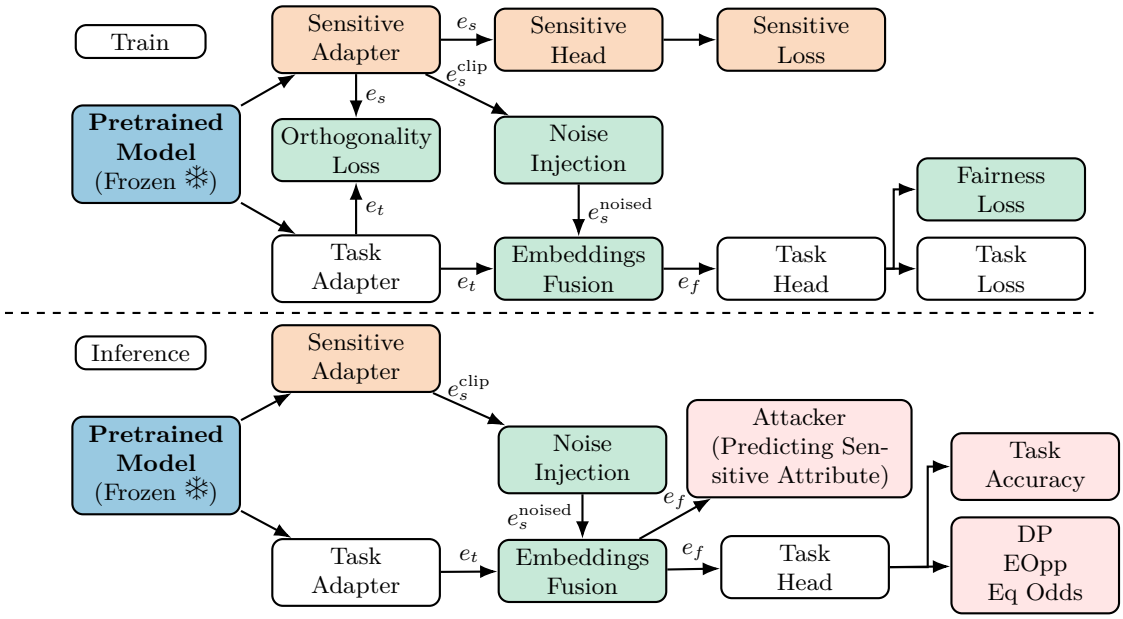


Figure 1: **Overview of the proposed FairNVT framework.** During *training*, a frozen ViT backbone is attached with lightweight task and sensitive adapters. The adapters yield task (e_t) and sensitive (e_s) embeddings. The sensitive path inputs e_s for the sensitive head and a clipped and noised embedding e_s^{noised} (e_s^{clip} injected with noise), that is concatenated with e_t to get the fused embedding e_f for task prediction. We jointly optimize a weighted sum of task and sensitive classification losses, orthogonality and fairness losses. During *inference*, the sensitive and task adapters remain frozen at trained weights. A random noise is drawn from the same distribution during training to obfuscate the sensitive embedding and fused with the task embedding before entering the task classification head. Fairness metrics are calculated from the fused embedding e_f. Blue marks the frozen backbone; Orange extracts sensitive information; Green performs debiasing, fusion, and prediction; Red marks the evaluation metrics.

3.1 Model Components

Adapters and classification heads. We use the Adapter modules to extract task-relevant and sensitive information from the frozen pre-trained models, with supervision from the task and sensitive labels. The Adapters are model-agnostic, lightweight blocks of trainable parameters attached to various blocks of the frozen pre-trained model. For example, for the image classification tasks which we use the Vision Transformer (ViT, Dosovitskiy et al. (2021)) as frozen pre-trained models, the Adapters [1] are bottleneck feed-forward layers attached to each Transformer block, consisting of down-projection matrix to project the hidden states into a lower dimension layer, and an up-projection matrix to project back into the original hidden dimension (Poth et al., 2023). To encourage disentangling task and sensitive embeddings, we attach separate Adapters and classification heads to learn the task and sensitive labels respectively. We use the class token representation (`[CLS]` token embedding) from the adapted ViT model as the task (e_t) and sensitive (e_s) embeddings, with only the task and sensitive Adapters activated respectively. The classification heads are simple Multi-layer Perceptrons (MLPs) that takes in the adapted embeddings and predict the task and sensitive labels respectively.

Noise injection. To introduce perturbation to the sensitive information, we clip the sensitive embedding and add random noise sampled from a Gaussian distribution. Specifically, let e_s be the sensitive embedding vector, we clip the sensitive embedding to upper-bound its $L2$-norm to C, $e_s^{\text{clip}} = e_s / \max(1, \frac{||e_s||_2}{C})$, where C is a hyperparameter to control the embedding scale. The clipping procedure ensures that we calibrate the noise level to the scale of the embedding, thus better controlling the strength of perturbations. The noise z is

[1] https://docs.adapterhub.ml/methods.html#bottleneck-adapters

then randomly drawn from a Normal distribution with mean equals zero, such that the perturbed embedding is unbiased, and with the variance scales with C , i.e., $z \sim \mathcal{N}(0, C^2\sigma^2\mathbb{I}^d)$, where σ controls the noise level and d is the dimension of e_s. Finally, the perturbed embedding is obtained by adding noise z isotopically to the sensitive embedding, $e_s^{\text{noised}} = e_s^{\text{clip}} + z$. Injecting calibrated noise has two effects in terms of making fair task predictions. First, it perturbs the sensitive information extracted from the sensitive adapter, such that the model learns to depend less on irrelevant sensitive information when predicting the task label. Second, the task information extracted from the task adapter is preserved but perturbed with noisy sensitive information, to encourage more robust predictions of the task labels from the noisy embedding.

Embedding fusion. After obtaining the perturbed sensitive embedding, we concatenate it with the task embedding, and use it as the embedding for downstream classification tasks. We add two steps to ensure that the task and sensitive Adapter capturing the correct information such that the fused embedding works as expected. The noised sensitive embedding only enters the task label classification head and not the sensitive head, such that the sensitive adapter captures the clean sensitive information. The gradient is blocked from the task loss to the noised sensitive embedding, such that the learned task information does not interfere with the sensitive Adapter.

3.2 Optimization Objectives

The proposed framework is trained jointly with classification, orthogonality and fairness losses to balance between making accurate task predictions and maintaining fairness with respect to the sensitive information. We describe each loss and their objectives in this section.

Classification loss. The cross-entropy loss is used for each classification head to evaluate the predictive performances. Let i, k be the sample and class index, θ be the model parameters, (x_i, y_i) be each data-label pair, $\hat{y}$ be the predicted label, then the cross-entropy loss for predicting task (t) and sensitive (s) labels are,

$$L_{\text{ce}}^{\alpha}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} y_{i,\alpha}^{k} \log p_\theta(\hat{y}_{i,\alpha} = k|x_i), \alpha \in \{s, t\}. \tag{1}$$

Orthogonality loss. It is common that sensitive information might help predicting the task label as they share common features. For example, when the task is predicting 'wearing glasses' and the sensitive label is 'Age', the features might be correlated as it might coincide that elder people often wear glasses. In such cases, we want to penalize similar patterns in the task and sensitive embeddings and encourage finding distinct features in predicting the task label. We use mean cosine similarity to quantify the similarities between task and sensitive embeddings, and penalize for higher scores to encourage de-correlating these embeddings. Let $e_{t,i}, e_{s,i}$ denote the per-sample task and sensitive embedding that depend on θ,

$$L_{\text{orth}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{e_{t,i}^{\top} e_{s,i}}{||e_{t,i}||_2 ||e_{s,i}||_2}\right)^2. \tag{2}$$

Fairness loss. While the orthogonality loss and noise injection help disentangling the sensitive information, to ensure making fair predictions, we add a fairness loss to encourage learning similar logits among different sensitive groups. Following the definition of demographic parity difference (Agarwal et al., 2019), let n_0, n_1 be the number of samples in a batch belonging to sensitive group $0, 1$ respectively, $p = p_\theta(\hat{y} = 1|x)$ be the probability of predicting the positive class of label y, and $\mathbf{1}[\cdot]$ be the indicator function then,

$$L_{\text{dp}}(\theta) = \Big|\frac{1}{n_0}\sum_{i=1}^{n_0}\mathbf{1}[s_i = 0]p_i - \frac{1}{n_1}\sum_{j=1}^{n_1}\mathbf{1}[s_j = 1]p_j\Big|. \tag{3}$$

We optimize demographic parity during training because it provides a simple and stable surrogate objective for suppressing sensitive signals in the classifier embedding. Metrics such as equalized odds require extra conditioning on the task label y, which leads to more complex and less stable optimization. Since our goal is

to reduce sensitive information in the embedding itself, enforcing demographic parity encourages the model to minimize the dependence between predictions and sensitive attributes. Empirically, we find that reducing such loss also improves equalized odds and equal opportunity, which we report as evaluation metrics.

The overall loss is weighted to adjust the scale differences between the three losses and to allow flexibility of viewing different importance of the optimization targets, $L = L_{\text{ce}}^{\text{t}} + \beta_1 L_{\text{ce}}^{\text{s}} + \beta_2 L_{\text{orth}} + \beta_3 L_{\text{dp}}$, where βs are hyperparameters representing weights on each loss.

3.3 Training and Inference Procedures

The arrows in Figure 1 shows the forward pass direction. During backpropagation, only the Adapters and classification heads parameters are updated with loss L, while the pre-trained model remains frozen. The noise injection and embedding concatenation steps do not induce learnable parameters. In the training stage, lightweight task and sensitive adapters are attached to the frozen backbone to produce the task embedding e_t and sensitive embedding e_s. The sensitive embedding is clipped and noised to obtain the embedding e_s^{noised}, which is then concatenated with e_t to form the fused embedding $e_f = [e_t, e_s^{\text{noised}}]$. The fused embedding is used as the input to the task classifier head. The adapters and classifier parameters are jointly optimized with loss L to learn a fair predictor on the task attribute. In the inference stage, the adapters and classifier head are fixed at the learned parameters and applied directly to the input data.

We make several key remarks for the training and inference pipeline. First, noise is sampled randomly from a fixed distribution and injected during both training and inference stages. This ensures consistency between the training and the inference procedure, allowing the model to adapt to the perturbations it will encounter at test time. As a result, the model learns a smoothed classifier that is robust to variations in the sensitive embedding and makes stable predictions that depend less on the sensitive attribute. While we also explore sampling multiple random noise at inference time with majority voting in Appendix E, we observe that a single draw of noise is sufficient in practice to effectively obfuscate sensitive signals in the fused embedding. Second, fairness evaluations are conducted on the fused embedding e_f, since it is the representation used by the trained classifier and therefore determines the model's predictions and reflects the information accessible to the decision-making process. Finally, the sensitive attribute labels are only used during training for computing the fairness loss. They are not accessed during inference, as the classifiers remain fixed at the trained parameters.

4 Experiments

We examine the performance of FairNVT on image classification tasks using the CelebA and UTKFace datasets, with a pretrained ViT-B/16 model as the frozen backbone. Across multiple task and sensitive attribute pairs, FairNVT demonstrate strong performance, achieving high task accuracy while improving prediction fairness with respect to the sensitive attribute (§4.1). We further validate our hypothesis that suppressing sensitive information through controlled noise can enhance prediction-level fairness while preserving task-relevant signals (§4.2). Implementation details are discussed in Appendix C. In appendix D.2, we extend FairNVT to the text domain and present results on the BIOS dataset, where a pre-trained Bert-Base model is used as the frozen backbone.

Datasets[2] and tasks. We use publicly available datasets CelebA (Liu et al., 2015) and UTKFace (Zhang et al., 2017) for facial attribute classification. CelebA contains roughly 200K images with attribute annotations. Following prior work (Tian et al., 2024; Park et al., 2022), we consider perceived gender or age as sensitive attributes, and we study target attributes such as expression(smiling), big nose and wavy hair. We use the official train/validation/test splits. UTKFace contains approximately 20K images with annotations including gender and age. To follow a binary fairness formulation (Park et al., 2022), we group age into < 35 vs. ≥ 35 and use age as the sensitive attribute and gender as the target attribute [3].

[2]The datasets are publicly available and include perceived annotations provided by the dataset creators. We use these labels solely for modeling and fairness evaluation to compare with previously published results on these benchmarks.

[3]Though there is no official train/validation/test splits, thee dataset has three subsets, where we use subsets 1, 2, and 3 for training, validation, and testing, respectively.

Metrics. We evaluate task performance, prediction-level and representation-level fairness performances with standard metrics[4] used by baseline methods. All reported values are scaled by $\times 10^2$.

(1) Task performance is measured using *accuracy* (Acc). We additionally report *balanced accuracy* (BAcc), as the task attribute is often imbalanced, and relying solely on accuracy may lead to misleading evaluations of task performance.

$$Acc \coloneqq \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{P}+\mathrm{N}},\ BAcc \coloneqq \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}+\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right), \tag{4}$$

where P, N denote real positive and negatives, TP, TN, FP, and FN denote true/false positives/negatives, respectively. To assess prediction-level fairness, we employ three widely used group fairness metrics. Given true labels $Y \in \{0,1\}$ and predictions $\hat{Y}$, let $S \in \{0,1\}$ denote the binary sensitive attribute:

(2) *Demographic Parity (DP)* computes the difference between the largest and smallest rates across all groups:

$$DP \coloneqq \max_s \mathbb{E}[\hat{Y} \mid S] - \min_s \mathbb{E}[\hat{Y} \mid S], \tag{5}$$

which simplifies to $|\mathbb{E}[\hat{Y}=1 \mid S=0] - \mathbb{E}[\hat{Y}=1 \mid S=1]|$ in the binary case.

(3) *Equalized Odds (EO)* adds conditioning on the task label compared to *DP*. EO evaluates group fairness by averaging disparities in both true positive rates and false positive rates across sensitive groups:

$$EO \coloneqq \frac{1}{2}\sum\nolimits_{y\in\{0,1\}} \big|\mathbb{E}[\hat{Y}=1 \mid Y=y, S=0] - \mathbb{E}[\hat{Y}=1 \mid Y=y, S=1]\big|, \tag{6}$$

(4) *Equal Opportunity (EOpp)* is a relaxed version of *EO* that only considers conditional expectations with respect to positive task labels. EOpp considers the disparities in true positive rates only:

$$EOpp \coloneqq |\mathbb{E}[\hat{Y}=1 \mid Y=1, S=0] - \mathbb{E}[\hat{Y}=1 \mid Y=1, S=1]|. \tag{7}$$

We report absolute DP/EO/EOpp gaps throughout, in all three metrics, lower values indicate higher fairness level. To assess representation-level fairness, we examine the prediction accuracy of the sensitive attribute from an attacker. Lower values indicate higher fairness level.

(5) *Attacker accuracy (Att.Acc)* measures sensitive-information leakage using a post-hoc attacker: a MLP trained to predict S from embeddings (e_f) at a saved checkpoint (encoder frozen). Lower attacker accuracy indicates less recoverable sensitive information and thus stronger representation-level fairness. Architecture and training details of the attacker model are provided in Appendix C.

Baselines. We compare our approach with the vanilla setup, and several image-based fair classification baselines under a unified evaluation protocol.

- **Vanilla (ViT)** (Dosovitskiy et al., 2021): ViT with a task adapter and classification head trained, with no fairness intervention.
- **ViT-FSCL** (Park et al., 2022): Representation-level contrastive debiasing; we re-implement it on a ViT backbone for consistent comparison.
- **FairViT** (Tian et al., 2024): Architecture-level debiasing via adaptive masking on ViT attention maps.
- **FairVPT** (Park & Byun, 2024): Prediction-level debiasing using Visual Prompt Tuning that adapts pre-trained ViT model to downstream classification tasks with eliminated biased information [5].

[4] https://fairlearn.org/

[5] FairVPT does not have official code release and is implemented by the authors based on the descriptions in the paper.

4.1 Main Results

Table 1 shows the results that compare FairNVT with the baselines on CelebA and UTKFace datasets. We tune the hyperparameters for all methods based on the highest task accuracy and report the mean values and standard deviations for all metrics over 3 runs. Best results are shown in **bold**, and the second-best results are underlined. Overall, we observe that *FairNVT achieves significantly lower attacker accuracy and fairer downstream predictions, while preserving strong task performance.*

Table 1: **Image-Based Classification task:** Comparing our method with baselines on CelebA (a-c) and UTKFace (d) dataset. FairNVT demonstrates strong performance in higher task performance while achieving fairer predictions.

Method	Acc(↑)	BAcc(↑)	DP(↓)	EOpp(↓)	EO(↓)	Att.Acc(↓)
Vanilla	89.6±0.1	89.0±0.1	16.9±0.2	8.4±1.2	6.6±1.2	98.7±0.0
ViT-FSCL	89.9±1.0	87.1±1.0	14.5±2.2	6.9±2.2	5.1±2.1	97.7±0.1
FairViT	92.7±0.2	92.0±0.3	16.0±0.3	4.3±0.4	2.7±0.6	97.0±0.1
FairVPT	91.6±0.2	91.4±0.2	13.9±0.3	2.4±0.3	1.8±0.6	98.6±0.2
FairNVT(Ours)	**93.1**±0.2	**93.0**±0.3	**9.9**±0.3	**0.8**±0.3	**1.5**±0.5	**51.6**±0.4

(a) Task: Expression (Smiling), Sensitive Attribute: Gender (Male)

Method	Acc(↑)	BAcc(↑)	DP(↓)	EOpp(↓)	EO(↓)	Att.Acc(↓)
Vanilla	80.2±0.2	63.4±1.2	25.0±1.2	19.2±0.2	25.2±1.2	88.3±0.0
ViT-FSCL	81.5±0.8	68.1±1.6	25.5±4.4	23.2±4.2	18.3±3.8	87.0±0.3
FairViT	**84.6**±0.2	**69.9**±0.2	22.7±0.8	22.9±1.4	16.7±1.2	86.0±0.3
FairVPT	83.1±0.2	64.0±0.3	17.8±1.0	23.0±1.0	15.1±0.8	87.5±0.2
FairNVT(Ours)	82.1±0.2	69.2±0.5	**10.9**±1.5	**2.3**±0.8	**1.9**±0.2	**67.6**±0.6

(b) Task: Big Nose, Sensitive Attribute: Age (Young)

Method	Acc(↑)	BAcc(↑)	DP(↓)	EOpp(↓)	EO(↓)	Att.Acc(↓)
Vanilla	84.4±0.5	76.1±0.1	31.1±2.7	31.3±2.5	34.9±2.7	98.6±0.1
VIT-FSCL	83.5±0.5	69.8±2.7	31.0±1.1	28.9±4.1	36.7±11.6	97.6±0.1
FairViT	**86.4**±0.4	79.9±0.3	38.0±0.8	30.0±1.1	20.9±0.9	94.2±0.2
FairVPT	84.6±0.4	76.5±0.4	31.8±0.8	28.5±0.8	17.5±1.0	98.4±0.1
FairNVT (Ours)	84.7±0.3	**82.3**±0.2	**18.9**±0.9	**5.6**±0.6	**6.3**±0.7	**62.8**±0.4

(c) Task: Wavy Hair, Sensitive Attribute: Gender (Male)

Method	Acc(↑)	BAcc(↑)	DP(↓)	EOpp(↓)	EO(↓)	Att.Acc(↓)
Vanilla	97.3±0.1	96.0±0.5	19.5±0.3	1.3±0.2	3.1±0.2	82.7±0.5
ViT-FSCL	97.4±0.2	96.7±0.5	19.2±0.1	2.2±0.3	**1.1**±0.1	82.3±0.2
FairViT	97.5±0.0	97.1±0.1	21.0±0.9	1.8±0.3	1.1±0.4	81.0±0.2
FairVPT	95.3±0.1	93.9±0.2	19.4±0.2	2.0±0.4	2.0±0.5	74.1±0.2
FairNVT(Ours)	**97.7**±0.5	**97.4**±0.5	**18.4**±0.7	**0.6**±0.2	1.5±0.7	**50.2**±1.0

(d) Task: Gender, Sensitive Attribute: Age

Comparison on CelebA. As shown in Table 1 (a-c), FairNVT maintains a balanced trade-off between fairness and task performance across all task–sensitive pairs. For example, FairNVT has the highest task accuracy of 93.1% (1% improvement from the best baseline), with the best fairness scores (lower DP, EO and EOpp) for the *expression-gender* attribute pair. Across other pairs of attributes, our method achieves comparable or higher task accuracies while consistently improving multiple fairness metrics. Notably, FairNVT reduces sensitive information leakage as indicated by attacker accuracy dropping by an average of 35% across three different attribute pairs, bringing attacker accuracy closer to random, and demonstrating the effectiveness of the proposed noise-based regularization.

Comparison on UTKFace. Table 1 (d) presents the results on the UTKFace dataset where *Gender* and *Age* are treated as task and Sensitive attribute respectively. FairNVT consistently shows strong fairness improvements while maintaining a high task accuracy of 97%, achieving a balanced outcome with high task performance and reduced fairness disparities.

Qualitative results. Figure 3 visualizes model attributions on *CelebA*, where task is *expression (smiling)* and sensitive attribute is *gender (male)*. We observe that the Vanilla model frequently relies on irrelevant background or on gender-correlated regions (e.g., hair/beard), while ViT-FSCL, FairViT and FairVPT partially down-weight such cues. Our method consistently attends more to expression-relevant regions such as the mouth, cheeks, and eyes, while effectively suppressing gender-related cues. This behavior aligns with the observed quantitative improvements in DP and EOpp, as well as the significant reduction in attacker accuracy.

4.2 Ablation Studies

Except otherwise noted, we perform the ablation studies on the CelebA dataset with *expression (smiling)* and *gender (male)* as the task and sensitive attributes respectively. By ablating and adjusting the strength of model components, we show that adding a controlled level of noise can suppress sensitive signals in the embedding prior to the classifier, and producing fairer predictions for the classifier.

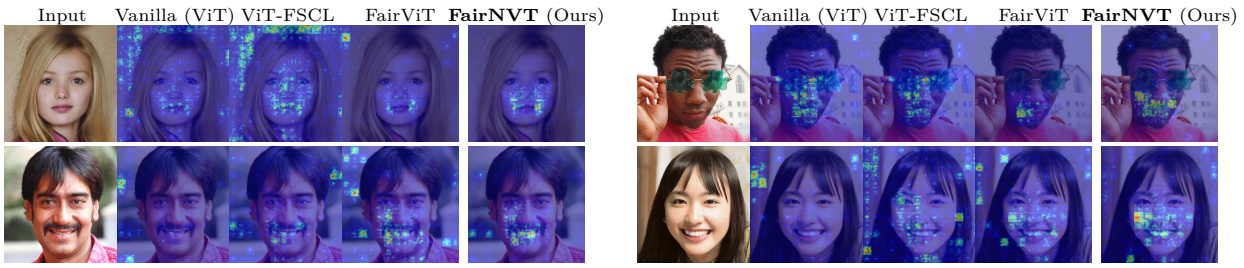


Figure 2: **Gradient-based saliency map** for the *Expression (smiling)* as main task and *Gender (male)* as sensitive attribute. Warmer regions indicate stronger contribution to the output logit. FairNVT primarily attends to expression-relevant areas (mouth/cheeks), demonstrating reduced reliance on gender-correlated cues.

Effect of different model components. Table 2 presents the ablation study evaluating the contribution of different components in the proposed method. Across settings, task accuracy and balanced accuracy remain broadly stable, indicating that these fairness components do not degrade utility. Splitting into task and sensitive adapters already improves fairness metrics from the Vanilla finetuning baseline, suggesting that disentangling information with supervision from both task and sensitive labels is more effective than learning from the task label directly. The orthogonality loss encourages the task and sensitive adapters to capture distinct features; without it, the model can mix task-relevant and sensitive signals, and may inadvertently discard information that is also predictive for the task during debiasing. The fairness loss consistently strengthens fairness, by explicitly reducing dependence between predictions and the sensitive attribute. Finally, noise is critical for leakage reduction: removing it weakens all fairness measures, with the clearest impact on attacker accuracy, confirming that perturbing the sensitive adapter embedding effectively conceals sensitive attributes.

Table 2: **Ablation of FairNVT components on CelebA.** We toggle Fairness loss (Fair), Orthogonality loss (Orth), and Noise injection (Noise) for *expression (smiling)* as main task and *gender (male)* as sensitive attribute. ✓and ✗means the component is present and absent respectively. Fairness loss consistently drives fairness metrics, Noise suppresses sensitive attribute leakage, and Orth further improves fairness, with minimal utility change.

Fair Loss	Orth Loss	Noise	Acc(↑)	BAcc(↑)	DP(↓)	EOpp(↓)	EO(↓)	Att Acc(↓)
✗	✗	✗	92.7±0.1	92.2±0.1	13.8±0.6	4.8±0.4	2.8±0.5	98.4±0.0
✗	✓	✓	**93.2±0.1**	92.8±0.1	14.6±0.5	4.9±0.5	2.2±0.6	52.8±0.5
✓	✗	✓	92.6±0.1	92.8±0.1	**9.9±0.4**	1.1±0.4	2.4±0.5	53.0±0.3
✓	✓	✗	92.9±0.3	92.9±0.2	10.1±0.4	2.4±0.4	3.0±0.4	98.5±0.1
✓	✓	✓	93.1±0.2	**93.0±0.3**	**9.9±0.3**	**0.8±0.3**	**1.5±0.5**	**51.6±0.4**

Analysis of noise strength (σ). We analyze how the noise strength σ influences both representation and prediction level fairness. As shown in Table 3, moderate noise substantially lowers attacker accuracy, indicating that the injected perturbation effectively hides sensitive information without disturbing the task signal. When the noise becomes very large, the model shows improvements in several fairness metrics (DP, EO, Att.Acc) but shows a slight decline in predictive accuracy. In practice, a moderate noise level provides a stable trade-off between privacy and utility. Additional experiments related to component isolation (noise only, fairness loss only, orthogonality loss only) and sensitivity analyses of the corresponding loss weights are included in Appendix E.

Representation-level fairness results with stronger attacker models. Table 4 summarizes results obtained with stronger attacker models. We additionally report balanced attacker accuracy (Balanced Att. Acc.) in case the sensitive attribute is imbalanced. We observe that increasing the number of hidden layers in the MLP attacker does not substantially affect its accuracy (Att. Acc.) in predicting the sensitive attribute from

Table 3: **Sensitivity to noise.** We ablate on noise levels for *expression (smiling)* as main task and *gender (male)* as sensitive attribute. Moderate noise levels balance utility (Acc/BAcc), fairness gaps (DP/EOpp/EO), and sensitive information leakage (Att. Acc). Very large noise further improves most fairness metrics but begins to slightly reduce accuracy, reflecting a utility-fairness trade-off at higher noise levels.

Noise Level (σ)	**Acc**(↑)	**BAcc**(↑)	**DP**(↓)	**EOpp**(↓)	**EO**(↓)	**Att.Acc**(↓)
1	93.0±0.2	93.1±0.2	9.4±0.4	1.0±0.2	2.0±0.4	67.4±0.2
5	93.1±0.2	93.0±0.3	9.9±0.3	0.8±0.3	1.5±0.5	51.6±0.4
100	91.0±0.3	91.2±0.2	9.2±0.5	0.9±0.5	1.1±0.4	50.5±0.3

the de-biased embeddings, indicating that clipping and noise injection effectively mitigate sensitive-attribute leakage. In experiments where *age (young)* serves as an imbalanced sensitive attribute, the balanced attacker accuracies confirm that FairNVT consistently reduces the attacker's success rate to near-random levels. The results show that the information relevant to the sensitive attribute is effectively suppressed in the debiased embedding e_f prior to the classifier.

Table 4: **Performance with stronger attackers.** We evaluate the representation-level fairness result of FairNVT using stronger attacker models with increasing model depth. The attacker accuracies remain low with deeper attack models, indicating the sensitive information is obfuscated in the embedding before downstream classifiers.

Task	**# Hidden Layers**	**Att. Acc**(↓)	**Balanced Att. Acc**(↓)
Task: Expression (Smiling) **Sens.: Gender (Male)**	1	51.6±0.4	50.9±0.8
	3	52.2±0.4	50.8±1.0
	10	52.4±0.6	51.1±1.0
Task: Big Nose **Sens.: Age (Young)**	1	67.6±0.6	53.2±0.9
	3	67.9±0.5	53.3±1.2
	10	68.3±0.6	53.6±1.1

5 Conclusion

We introduced FairNVT, a plug-in framework that injects calibrated Gaussian noise in a learned sensitive subspace to improve both representation- and prediction-level fairness, while keeping backbone weights frozen. Across multiple datasets with image and text-based tasks, it consistently reduces sensitive-attribute leakage, and matches or improves prediction fairness while maintaining high task performance. While our study focuses on image and text modalities, the same recipe naturally extends to additional modalities and a wide range of transformer-encoder–based architectures. We are excited about these directions and expect the approach to scale with little engineering overhead, motivating researchers to broaden applicability beyond text and image to new modalities and transformer-based models.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, 2018.

Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, 2019.

Eric Aubinais, Elisabeth Gassiat, and Pablo Piantanida. Fundamental limits of membership inference attacks on machine learning models. *arXiv preprint arXiv:2310.13786*, 2023.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *nnual Conference on Neural Information Processing Systems*, 2017.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, 2019.

Jun Dan, Yang Liu, Haoyu Xie, Jiankang Deng, Haoran Xie, Xuansong Xie, and Baigui Sun. Transface: Calibrating transformer training for face recognition from a data-centric perspective. In *Proceedings of the IEEE international conference on computer vision*, 2023.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

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

Zahra Fatemi, Chen Xing, Wenhao Liu, and Caimming Xiong. Improving gender fairness of pre-trained language models without catastrophic forgetting. In *Annual Meeting of the Association for Computational Linguistics*, 2023.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Annual Meeting of the Association for Computational Linguistics*, 2023.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3), 2024.

Somayeh Ghanbarzadeh, Yan Huang, Hamid Palangi, Radames Cruz Moreno, and Hamed Khanpour. Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models. In *Findings of the Association for Computational Linguistics*, 2023.

Gesa Götte. The effect of adversarial debiasing on model performance. In *INFORMATIK*, 2023.

Karina Halevy, Karly Hou, and Charumathi Badrinath. Who's the (multi-) fairest of them all: Rethinking interpolation-based data augmentation through the lens of multicalibration. In *Association for the Advancement of Artificial Intelligence*, 2025.

Lukas Hauzenberger, Shahed Masoudian, Deepak Kumar, Markus Schedl, and Navid Rekabsaz. Modular and on-demand bias mitigation with attribute-removal subnetworks. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Annual Meeting of the Association for Computational Linguistics*, 2023.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021.

Rashidul Islam, Huiyuan Chen, and Yiwei Cai. Fairness without demographics through shared latent space-based debiasing. In *Association for the Advancement of Artificial Intelligence*, 2024.

Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.

Jiayin Jin, Zeru Zhang, Yang Zhou, and Lingfei Wu. Input-agnostic certified group fairness via gaussian parameter smoothing. In *International conference on machine learning*, 2022.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

Jian Kang, Tiankai Xie, Xintao Wu, Ross Maciejewski, and Hanghang Tong. Infofair: Information-theoretic intersectional fairness. In *IEEE international conference on big data (big data)*, 2022.

Adam Karvonen, Can Rager, Samuel Marks, and Neel Nanda. Evaluating sparse autoencoders on targeted concept removal tasks. In *Second NeurIPS Workshop on Attributing Model Behavior at Scale*, 2024. URL `https://openreview.net/forum?id=H9DhZTb19S`.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys*, 54:1–41, 2022.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *International Conference on Learning Representations*, 2015.

Deepak Kumar, Oleg Lesota, George Zerveas, Daniel Cohen, Carsten Eickhoff, Markus Schedl, and Navid Rekabsaz. Parameter-efficient modularised bias mitigation via AdapterFusion. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia, 2023.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. Sustainable modular debiasing of language models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMLP*, 2021.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE symposium on security and privacy (SP)*, pp. 656–672. IEEE, 2019.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. A survey on fairness in large language models, 2024. URL `https://arxiv.org/abs/2308.10149`.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pp. 3384–3393. PMLR, 2018.

Shahed Masoudian, Cornelia Volaucnik, Markus Schedl, and Navid Rekabsaz. Effective controllable bias mitigation for classification and retrieval using gate adapters. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

Kartik Narayan, Vibashan VS, Rama Chellappa, and Vishal M Patel. Facexformer: A unified transformer for facial analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11369–11382, 2025.

Sungho Park and Hyeran Byun. Fair-vpt: Fair visual prompt tuning for image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12268–12278, June 2024.

Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Fair Contrastive Learning for Facial Attribute Classification . In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Alban Desmaison, Andreas Kopf, Edward Fiścher, Yuandong Tian, Vincent Hoffman, Nachiket Dalal, Siddharth Narang, Soumith Chintala, and Gregory P. Chanan. Pytorch: An imperative style, high-performance deep learning library. *ArXiv*, abs/1912.01703, 2019.

Momchil Peychev, Anian Ruoss, Mislav Balunović, Maximilian Baader, and Martin Vechev. Latent space smoothing for individually fair representations. In *European Conference on Computer Vision*, pp. 535–554. Springer, 2022.

Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. Adapters: A unified library for parameter-efficient and modular transfer learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, December 2023.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647/.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *nnual Conference on Neural Information Processing Systems*, 34, 2021.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. Contrastive learning for fair representations. *arXiv preprint arXiv:2109.10645*, 2021.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. Does representational fairness imply empirical fairness? In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pp. 81–95, 2022.

Enze Shi, Lei Ding, Linglong Kong, and Bei Jiang. Debiasing with sufficient projection: A general theoretical framework for vector representations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5960–5975, 2024.

Rui Sun, Fengwei Zhou, Zhenhua Dong, Chuanlong Xie, Lanqing Hong, Jiawei Li, Rui Zhang, Zhen Li, and Zhenguo Li. Fair-cda: continuous and directional augmentation for group fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9918–9926, 2023.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20730–20740, 2022.

Bowei Tian, Ruijie Du, and Yanning Shen. Fairvit: Fair vision transformer via adaptive masking. In *European Conference on Computer Vision*, pp. 451–466. Springer, 2024.

Rui Wang, Pengyu Cheng, and Ricardo Henao. Toward fairness in text generation via mutual information minimization based on importance sampling. In *International conference on artificial intelligence and statistics*, 2023.

Yaoli Wang, Yaojun Deng, Yuanjin Zheng, Pratik Chattopadhyay, and Lipo Wang. Vision transformers for image classification: A comparative survey. *Technologies*, 13(1):32, 2025.

Junsong Xie, Yonghui Yang, Zihan Wang, and Le Wu. Learning Fair Representations for Recommendation via Information Bottleneck Principle. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

Ke Yang, Charles Yu, Yi R. Fung, Manling Li, and Heng Ji. Adept: A debiasing prompt framework. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):10780–10788, Jun. 2023.

Samuel Yeom and Matt Fredrikson. Individual fairness revisited: transferring techniques from adversarial robustness. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021. ISBN 9780999241165.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2017.

Hanzhang Zhou, Zijian Feng, Zixiao Zhu, Junlang Qian, and Kezhi Mao. Unibias: Unveiling and mitigating llm bias through internal attention and ffn manipulation. *Advances in Neural Information Processing Systems*, 2024.

A Related Works

Transformers for Classification. Transformers have seen broad adoption for classification in both vision and text. Image-level transformer models such as ViT Dosovitskiy et al. (2021) have been widely used across domains including face analysis Dan et al. (2023); Narayan et al. (2025); Jacob & Stenger (2021), medical imaging Shao et al. (2021); Tang et al. (2022), and general object recognition Khan et al. (2022); Wang et al. (2025). These backbones match or surpass strong CNN models while offering flexible transfer to new datasets.

Similarly, transformer-based language models (e.g., BERT Devlin et al. (2019), RoBERTa Liu et al. (2019), DeBERTa He et al. (2021)) have become the dominant choice for text classification, often outperforming CNN/RNN architectures and transferring effectively via pretrain–adapt pipelines.

Given their importance, understanding and mitigating their fairness challenges is crucial. We focus on image classification with frozen vision transformers and show that the proposed framework also transfers effectively to text.

Fairness Approaches. Many approaches mitigate group disparities by modifying the training distribution itself. Classical methods include reweighing Kamiran & Calders (2012), disparate impact removal Feldman et al. (2015), and optimized preprocessing Calmon et al. (2017), which explicitly adjust sample weights or features to balance sensitive groups. More recent strategies alter the training data more subtly through curated fine-tuning Ghanbarzadeh et al. (2023), group rebalancing, or fairness-oriented augmentation Sun et al. (2023); Halevy et al. (2025), aiming to reduce distributional bias without modifying model parameters.

Unlike data modification approaches, we make no data level changes, group labels are used only at training to optimize demographic parity, and not required at inference.

Beyond data manipulation, fairness has also been pursued through changing the learning objective Agarwal et al. (2018); Zhang et al. (2018). More recently, transformer-based methods adjust attention Zhou et al. (2024), mask bias-correlated ViT regions Tian et al. (2024), or apply fairness-aware prompting Park & Byun (2024). These recent approaches typically adjust attention or prompting, whereas we operate in a learned sensitive latent subspace and apply randomized smoothing without architectural changes or retraining the frozen model.

Another direction introduces parameter-efficient modules for debiasing, such as adapters Fatemi et al. (2023); Hauzenberger et al. (2023); Yang et al. (2023); Lauscher et al. (2021); Kumar et al. (2023); Masoudian et al. (2024). For example, DAM Kumar et al. (2023) adds debiasing adapters alongside task adapters to handle multiple sensitive attributes, while ConGater Masoudian et al. (2024) introduces controllable gates that balance fairness and utility at inference time. Although these approaches lower training cost, they typically act indirectly on representations without explicitly identifying or perturbing a sensitive subspace. In contrast, our method keeps the transformer backbone frozen and directly manipulates a learned sensitive subspace through noise injection.

Recent methods improve fairness by directly altering latent representations, with approaches based on latent factorization or variational modeling Zemel et al. (2013); Louizos et al. (2015) and adversarially aligned representations Madras et al. (2018); Zhang et al. (2018); Götte (2023) that aim to reduce sensitive information in learned features through min–max optimization, can be unstable and often requiring multi-stage training.

Projection-based methods such as INLP Ravfogel et al. (2020), sufficient projection (SUP) Shi et al. (2024), and SLSD Islam et al. (2024) remove subspaces predictive of sensitive attributes; however, linear removal can discard task-relevant information when sensitive and semantic directions overlap. Information-theoretic approaches Kang et al. (2022); Wang et al. (2023); Xie et al. (2024) estimate mutual information to regularize fairness, while contrastive debiasing Park et al. (2022); Shen et al. (2021) often relies on group-balanced sampling and two-stage optimization. Recent concept-editing methods Karvonen et al. (2024) learn sparse subspaces aligned with sensitive concepts and suppress them to reduce probe recoverability.

Fairness via Smoothing Models. Randomized smoothing Lecuyer et al. (2019); Cohen et al. (2019) is primarily studied as a robustness technique, where prediction stability under noise yields certified guarantees for robust predictions. Although not originally developed for fairness, the resulting invariance suggests that smoothing could help reducing reliance on sensitive factors.

Individual fairness is formalized via task-relevant similarity metrics Dwork et al. (2012). Empirical work connecting smoothing to fairness remains limited. For example, Jin et al. (2022) trains group-specific models and averages their parameters to certify group fairness in low-dimensional tabular settings, while Yeom & Fredrikson (2021); Peychev et al. (2022) encourage individual fairness via smoothing in input or latent spaces. These approaches, however, require isolated sensitive attributes in tabular data style, or operate only when input perturbations are well defined. Unlike prior works, our method performs smoothing selectively in a learned sensitive subspace, suppressing sensitive variation while preserving task structure, without architectural changes or using sensitive labels at inference.

B Obfuscating Sensitive Information Improves Fairness

In this section, we explain the mathematical intuition to the design of the FairNVT framework. The goal of achieving prediction-level fairness, as measured by Demographic Parity (DP) and Equalized Odds (EO), is to ensure the model predictions are similar across different sensitive groups, i.e. $P(\hat{Y}|S=0) = P(\hat{Y}|S=1)$ for DP, and $P(\hat{Y}|S=0, Y=y) = P(\hat{Y}|S=1, Y=y)$ for EO. In typical debiasing pipelines, the predictions often depend on both the representation Z learnt from data X and the sensitive information, i.e. $\hat{Y} = f(Z = e(X), S)$. This motivates the use of fair representations that suppress sensitive information in the learned embedding, resulting in predictions of the form $\hat{Y} = f(Z = e(X)), Z \perp\!\!\!\perp S$. By limiting the model's access to sensitive attributes, such representation removes unconditional dependence on S, and encourage alignment of conditional prediction behavior across groups, thereby reducing disparities in both DP and EO.

We formalize the intuition that noising sensitive information in the task classifier embedding improves both prediction- and representation-level fairness. Let (X, Y, S) be the data pair that represents features, task and sensitive attributes respectively. Let $Z = e(X)$ be the encoded embedding of X from an encoder model e (e.g. the frozen backbone models). Given a realized embedding z, let c be a classifier model that predicts task attribute $\hat{Y}$ with $\mathbb{1}(c(z) > \tau)$ where given a threshold τ. In the case where both S, Y are binary attributes, the following result follows directly from the definition of total variation distance between two probability measures.

Lemma B.1. *If $Z \perp\!\!\!\perp S$, then $DP = 0$, $EO = 0$, $EOpp = 0$.*

Proof. Let A be the event that the classifier c predicts $\hat{Y} = 1$, i.e. $A = \{z : \mathbb{1}(c(z) > \tau) = 1\}$, and let P, Q be the conditional distribution of $Z|S = 0$ and $Z|S = 1$ where $P(A) = \Pr(\hat{Y} = 1|S = 0)$, $Q(A) = \Pr(\hat{Y} = 1|S = 1)$, then by the definition of demographic parity difference (DP) and total variation distance (δ_{TV}),

$$DP \coloneqq |P(\hat{Y} = 1|S = 0) - P(\hat{Y} = 1|S = 1)| = |P(A) - Q(A)| \leq \sup_A |P(A) - Q(A)| \coloneqq \delta_{TV}(P, Q).$$

If $Z \perp\!\!\!\perp S$, then $P = Q$ and $\delta_{TV}(P, Q) = 0$ hence $DP = 0$. For EO, we consider the conditional distributions of $Z|S = 0, Y = y$ and $Z|S = 1, Y = y$. Since $Z \perp\!\!\!\perp S$, it follows that $P(Z|S = s, Y = y) = P(Z|Y = y), s \in \{0, 1\}$. Therefore the distributions are identical across groups for each y, thus implying $EO = 0$. The same argument applies to *EOpp*. □

Although achieving independence between Z and S ($Z \perp\!\!\!\perp S$) is challenging in practice, noising the embedding subspace encoding relevant information of S provides a feasible way towards the target. Assuming that the data embedding Z can be decomposed into two the task (Z^t) and sensitive (Z^s) embeddings, $Z = (Z^t, Z^s)$, where $Z^t \perp\!\!\!\perp S$. If Z^s is obfuscated by a large amount of noise such that it is a pure random embedding N with $N \perp\!\!\!\perp S$, then it would imply $Z \perp\!\!\!\perp S$. As δ_{TV} upper-bounds the membership inference accuracy (Theorem 3.1 Aubinais et al. (2023)), we also expect lower prediction accuracy on the attribute S from Z. Our design of the FairNVT framework follows from such intuitions.

C Experiment Setup Details

Model Architectures We use the ViT-Base model [6] as the frozen backbone for the CelebA and UTKFace datasets. Both task and sensitive adapters are bottleneck adapters inserted into each Transformer block of the frozen backbone. Each adapter consists of a down-projection that maps hidden states to a lower-dimensional space and an up-projection that restores them to the original hidden dimension. The reduction factor is a tunable hyperparameter. The task and sensitive classification heads are Multi-Layer Perceptrons (MLPs) whose hidden layer size matches the respective embedding dimension; the number of hidden layers is also treated as a tunable hyperparameter. For evaluating representation-level fairness via attacker accuracies, we use an attacker network with the same architecture as the task classification head. Table 5 summarizes an example FairNVT architecture used in the experiment for the task attribute *expression (smiling)* and the sensitive attribute *gender (male)*. FairNVT for vision task trains only 5.4M parameters ($\sim 6\%$ of ViT-Base) by freezing the backbone and introducing lightweight adapters and classification heads, significantly reducing computational cost compared to full fine-tuning.

Implementation Details. We implement all models in PyTorch Paszke et al. (2019) and train them on a workstation equipped with an AMD EPYC 7H12 CPU (64 cores) with a NVIDIA A100 GPU. For both our method and the baselines, we train the models using AdamW (Loshchilov & Hutter, 2019; Kingma & Ba, 2015) with batch size 256 and default hyper parameters of $\beta_1 = 0.9$, $\beta_2 = 0.999$, a weight decay of 0.01, and a batch size of 256. The adapter architecture uses a reduction factor of 8 for the task branch and 16 for the sensitive branch. Training the debiasing framework for one run takes approximately 3 hours on a single A100 GPU. Inference requires 1.1 seconds for a batch of 256 samples.

[6] `google/vit-base-patch16-224`

Table 5: FairNVT architectural specifications for the experiment with task attribute *expression (smiling)* and sensitive attribute *gender (male)*. Layer dimensions are denoted as $N_{\text{weight\_in}} \times N_{\text{weight\_out}} + N_{\text{bias}}$. Task and sensitive adapter layers are attached after the final dense layer of the frozen ViT encoder (output dimension = 768) in all 11 encoder layers. Noise injection and embedding concatenation introduce no trainable parameters.

Architecture	**Layer**	**Specification**	**Output Size**
Task Adapter	down_projection	$(768 \times 96 + 96) \times 11$	96
	up_projection	$(96 \times 768 + 768) \times 11$	768
Sensitive Adapter	down_projection	$(768 \times 48 + 48) \times 11$	48
	up_projection	$(48 \times 768 + 768) \times 11$	768
Noise Injection	\	\	768
Embedding Concatenation	\	\	768×2
Task Clf Head	linear_0	$(768 \times 2) \times (768 \times 2) + (768 \times 2)$	768×2
	tanh_activation	\	768×2
	linear_1	$(768 \times 2) \times 2 + 2$	2
Sensitive Clf Head	linear_0	$768 \times 768 + 768$	768
	tanh_activation	\	768
	linear_1	$768 \times 2 + 2$	2

During training, We run a grid search over other sensitive hyperparameters including learning rates and loss weights, and report the best validation-selected results. Specifically, we perform grid-search hyperparameter tuning over the following ranges: adapter reduction factor $\{4, 8, 16\}$; number of hidden layers $\{0, 1, 2\}$; learning rates (searched by half orders of magnitude, e.g., 1e−1, 5e−2, 1e−2, etc., until the best run is not at a boundary value); gradient-clipping thresholds $\{1, 10, 100\}$; noise levels $\{1, 5, 10\}$; and loss-weight coefficients $\beta \in \{0, 0.1, 0.3, 0.5, 1.0, 3.0\}$.

For evaluation, accuracy and balanced accuracy are computed from the predicted and true task labels. Fairness metrics (DP, EO, and EOpp) are computed using the predicted and true task labels together with the true sensitive attributes. The attacker setup follows Kumar et al. (2023): in an independent run, the attacker receives the task-classifier embeddings as input X and the corresponding sensitive attributes Y from the training and test sets. The attacker is trained to predict Y from X until the training accuracy no longer improves significantly, and its test accuracy is reported as the attacker's ability to recover the sensitive attribute from the learned representation.

D More Experiment Results

D.1 Image-based Classification Results

More experiments on CelebA. Table 12 shows the results on more task, sensitive attribute pairs in the CelebA dataset. In most cases, we observe FairNVT showing a good balance between the prediction and fairness objectives, achieves better or comparable performances to the best baseline across different metrics.

Additional qualitative results on CelebA. We provide additional samples for the *expression (smiling)* task with *gender (male)* as the sensitive attribute. Heatmaps are computed with *SmoothGrad* on the predicted class logit by averaging input gradients over 25 Gaussian noised, normalized inputs ($\sigma = 0.10$), aggregating $|\nabla_x|$ across channels, bilinearly resizing, and applying a light 3×3 blur, with maps normalized independently per image so intensities reflect within panel variation. Warmer regions indicate stronger contributions to the output logit. FairNVT concentrates on expression-relevant areas (e.g., mouth, cheeks), suggesting reduced reliance on gender-correlated cues and improved fairness via task-specific evidence.

D.2 Text-based Classification Experiments and Results

Text-based task. We additionally evaluate on **text-based classification tasks** with Bert-Base as the frozen backbone to compare with baselines Kumar et al. (2023); Ravfogel et al. (2020). The dataset BIOS De-

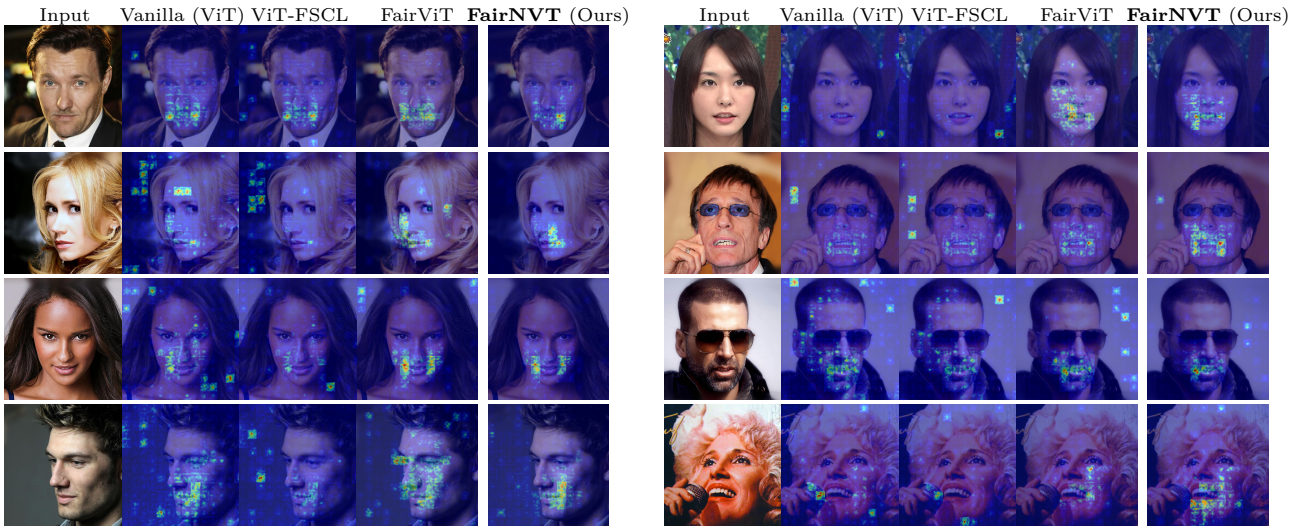


Figure 3: Additional examples: Gradient-based saliency map for the *expression (smiling)* as main task and *gender (male)* as sensitive attribute. Warmer regions indicate stronger contribution to the output logit. FairNVT primarily attends to expression-relevant areas (mouth/cheeks), demonstrating reduced reliance on gender-correlated cues.

Arteaga et al. (2019) consists of professional biographies with occupation labels, and the task is to predict occupation while evaluating fairness with respect to perceived gender.

Multi-class fairness loss. Since the task (*Profession*) in BIOS is a multi-class attribute, we extend the fairness loss in Eq. 3 to aggregate disparity with respect to the sensitive attribute across all task classes: let n_0, n_1 be the number of samples in a batch belonging to sensitive group $0, 1$ respectively, $p_k = p_\theta(\hat{y} = k|x)$ be the probability of predicting class k of label y, and $\mathbf{1}[\cdot]$ be the indicator function then,

$$L_{\text{dp}}^{\text{multi}}(\theta) = \sum_{k=1}^{c} \Big| \frac{1}{n_0} \sum_{i=1}^{n_0} \mathbf{1}[s_i = 0] p_{i,k} - \frac{1}{n_1} \sum_{j=1}^{n_1} \mathbf{1}[s_j = 1] p_{j,k} \Big|.$$

The overall loss remains the same form except substituting in the multi-class fairness loss, $L = L_{\text{ce}}^{\text{t}} + \beta_1 L_{\text{ce}}^{\text{s}} + \beta_2 L_{\text{orth}} + \beta_3 L_{\text{dp}}^{\text{multi}}$, where βs are hyperparameters representing weights on each loss.

Text-based fair classification baselines. We include the Vanilla setup and five baselines:

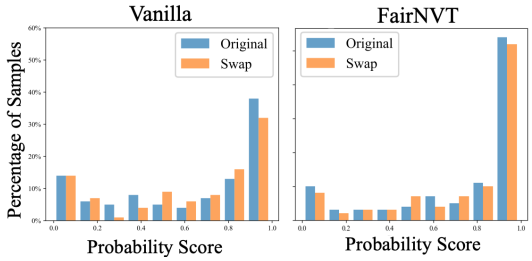
- **Vanilla-BERT** Devlin et al. (2019): Standard fine-tuning without fairness intervention.
- **FT-Debias** Kumar et al. (2023): Fine-tuning with adversarial debiasing objectives.
- **INLP** Ravfogel et al. (2020): Iteratively trains linear probes on the sensitive attribute and projects embeddings to remove the corresponding subspaces.
- **SUP** Shi et al. (2024): Projection-based concept removal that preserves task-relevant features while suppressing sensitive directions.
- **ConGater** Masoudian et al. (2024): Group-aware contrastive training to disentangle task and sensitive representations.
- **DAM** Kumar et al. (2023): Parameter-efficient debiasing using adapter fusion to reduce demographic leakage.

Comparison on BIOS. Table 6 reports results on BIOS De-Arteaga et al. (2019), where the task is multi-class *Profession*[7] and the sensitive attribute is *Gender*. We report the mean results and its standard deviation over 3 runs. Overall, FairNVT offers a favorable fairness–utility trade-off, pairing competitive accuracy with state-of-the-art DP and sensitive-attribute leakage close to the best baseline. Figure 4 compares logits for original and gender-swapped sentences, where pronouns are replaced with those of the opposite gender. FairNVT produces more similar distributions between the two, indicating reduced sensitivity to gender. Additional details, illustrative examples, and predicted scores are provided in the supplementary materials. These results highlight that FairNVT can effectively extend from the vision domain to textual embeddings.

Table 6: **Text-Based Classification task:** Comparing our method with baselines on Bios De-Arteaga et al. (2019) dataset, Task: Profession (Multi-Class), Sensitive Attribute: Gender. All reported values are scaled by $\times 10^2$.

Method	Acc(↑)	DP(↓)	Att.Acc(↓)
Vanilla-BERT	72.8±0.2	2.0±0.2	99.6±0.0
+FT-Debias	76.8±2.4	2.1±0.2	58.4±0.3
+INLP	76.4±0.1	1.7±0.1	**51.9±0.2**
+SUP	77.2±0.2	2.1±0.5	74.3±0.6
+DAM	80.3±0.4	2.2±0.5	60.6±0.2
+CONGATER	**82.4±0.5**	1.9±0.3	59.0±0.2
+FairNVT(Ours)	80.6±0.4	**1.6±0.1**	52.8±0.3

Figure 4: **Robustness to gender-indicator swapping on BIOS.** We plot the distribution of the model's confidence in predicting profession for the original text and its gender-swapped counterpart for 100 random samples. FairNVT (right) exhibits more overlapping distributions than Vanilla (left) in more confident predictions.



Qualitative results on BIOS. We evaluate fairness by comparing predictions on pairs of sentences that are identical except for words that indicate gender. Table 13 summarizes how the predicted profession probabilities change under these minimal substitutions. The vanilla model shows substantial shifts, whereas FairNVT produces more stable predictions across sentences that differ only in gender-indicative terms.

E More Ablation Results

Comparing different task classifier inputs. Table 7 shows the effect on accuracy and fairness metrics when the task classifier input changes. These results are trained with the same loss L as in Section 3, except nullifying the orthogonality loss when the sensitive embedding (e^s) is not present. When using the backbone frozen embedding h directly (row 1) or concatenating task (e_t) and sensitive embedding (e_s) without noise (row 2), it is more difficult to obtain fair outcomes when the sensitive information is not obfuscated, indicated by higher DP, EOpp, EO and Att.Acc values. Naively adding noise to h (row 3) could achieve good fairness outcomes but hurting the task performance. Although concatenating pure noise z with the task embedding e_t (row 4) slightly improves fairness metrics, it does not achieve the same level of fairness as FairNVT. Since z is not injected through the sensitive branch, it does not target and suppress sensitive information directly. We show in Table 4(App. E) that increasing attacker model complexity by adding more layers does not improve the attacker accuracy in FairNVT. We additionally test on alternative ways of fusing e_s, e_t. Aligning noise z with e_s before concatenating with e_t (row 5) improves task accuracy and achieves competitive DP, EOpp and EO, but leaks sensitive information as attacker accuracy increases. Fusing task and sensitive embedding with self-attention (row 6) preserves more sensitive information thus slightly hurting the fairness outcomes. Overall, simple concatenation of noisy e_s with e_t achieves the balance between accurate task prediction, fair outcomes and reducing sensitive information leakage.

[7]EO and EOpp condition on a binary label and are not directly applicable to multi-class tasks; DP remains applicable.

Table 7: **Effect of noise and projection choices on fairness and utility.** We assess variants of task classification head inputs constructed from frozen backbone output without having any adapter h, sensitive embedding e_s, task embedding e_t, and injected noise z. The comparison highlights how noise injection, projection, and attention choices influence task performance and fairness. All reported values are scaled by $\times 10^2$. We report the mean and standard deviation over 3 runs. Task: Expression (Smiling); Sensitive attribute: Gender (Male).

$Task^{clf}$ Inputs	Acc(↑)	BAcc(↑)	DP(↓)	EOpp(↓)	EO(↓)	Att.Acc(↓)
$[h]$	90.2±0.0	90.1±0.1	10.5±1.3	1.5±0.7	2.3±0.5	98.8±0.0
$[e_s, e_t]$	92.9±0.3	92.9±0.2	10.1±0.4	2.4±0.4	3.0±0.4	98.5±0.1
$[z, h]$	86.4±0.2	86.5±0.2	7.0±0.8	1.4±0.4	4.1±0.9	89.5±0.1
$[z, e_t]$	93.0±0.1	93.0±0.1	10.0±0.5	**0.5±0.3**	2.8±0.5	63.7±0.6
$[\frac{\langle z^i, e^i_s \rangle}{\|e^i_s\|} e^i_s, e_t]$	92.2±1.0	**93.2±0.9**	**9.8±0.8**	0.8±0.5	2.1±0.5	98.8±0.0
Attn $(e_s + z, e_t)$	91.9±0.2	91.7±0.2	10.2±1.1	1.3±0.5	2.8±1.0	54.5±0.2
FairNVT	**93.1±0.2**	93.0±0.3	9.9±0.3	0.8±0.3	**1.5±0.5**	**51.6±0.4**

Effect of ablating model components. Table 8 presents ablation results for different model components. The results are consistent with the main findings in Section 4: the DP loss primarily drives fairness improvements, noise injection reduces sensitive-attribute leakage, and the orthogonality loss further enhances fairness with minimal impact on task performance. Model components also exhibit interacting effects; in particular, combining DP loss with noise injection further decreases DP, EOpp, and EO scores, indicating that enhancing representation-level fairness can align with improvements in prediction-level fairness.

Table 8: **Ablation of FairNVT components on CelebA.** We toggle Demographic Parity loss (DP), Orthogonality loss (Orth), and Noise injection (Noise). ✓and ✗means the component is present and absent respectively. All reported values are scaled by $\times 10^2$ and show performance from a single run with the same seed.

Task/Sens.	DP Loss	Orth Loss	Noise	Acc(↑)	BAcc (↑)	DP (↓)	EOpp(↓)	EO(↓)	Att Acc(↓)
Task: Expression (Smiling)	✓	✗	✗	92.7	92.7	8.6	0.8	4.4	98.9
	✗	✓	✗	93.4	93.1	14.3	4.0	4.0	99.0
Sens.: Gender (Male)	✗	✗	✓	93.0	92.7	14.8	4.8	4.8	54.2
	✓	✓	✓	93.0	93.0	9.8	0.3	2.8	52.6
	✗	✗	✗	83.2	69.7	23.7	20.7	20.7	87.8
	✓	✗	✗	82.4	67.9	13.1	4.6	4.6	88.0
Task: Big Nose	✗	✓	✗	83.5	70.5	23.8	21.0	21.0	88.0
	✗	✗	✓	83.0	69.8	23.4	18.9	18.9	70.6
Sens. Age (Young)	✗	✓	✓	83.2	71.0	24.9	19.8	19.8	69.8
	✓	✗	✓	82.4	69.1	13.3	3.3	3.3	68.6
	✓	✓	✗	82.6	68.4	12.7	3.8	3.8	88.0
	✓	✓	✓	82.2	68.3	12.6	2.3	2.6	68.5

Sensitivity of loss weight coefficients. We analyze the effect of loss weight coefficients in Table 9, using the task and sensitive attributes *expression (smiling)* and *gender (male)*, respectively. A moderate orthogonality loss weight consistently achieves the best balance between task accuracy and fairness metrics, indicating that this setting effectively disentangles task and sensitive embeddings without degrading representation quality. Increasing the DP loss weight improves prediction-level fairness, particularly for demographic parity difference, which it directly optimizes, though with a gradual trade-off in task performance. Because EO and EOpp condition on specific label groups, they are naturally more sensitive to small prediction variations, yet we observe stable improvements at moderate DP weights. Overall, these trends highlight that the loss weights control the fairness–utility balance in a predictable manner, and tuning them allows FairNVT to adapt robustly across datasets and attribute combinations.

Sensitivity of embedding clipping. As discussed in Section 4, we clip the embeddings to an upper bound C before adding noise, which helps control the obfuscation of sensitive information. We analyze the sensitivity of the model to different values of the clipping threshold C. Changing C under a fixed noise multiplier (σ) has a combined effect: it alters the embedding magnitude while also changing the effective

Table 9: **Sensitivity of loss weight coefficients.** We evaluate the performance of FairNVT when the loss weight coefficients changes. All reported values are scaled by $\times 10^2$ and show performance from a single run with the same seed. Task: Expression (Smiling); Sensitive attribute: Gender (Male).

	Level	Acc(↑)	BAcc(↑)	DP(↓)	EOpp(↓)	EO(↓)	Att.Acc(↓)
Orth Loss	0	92.8	92.7	9.9	0.2	3.1	53.0
	0.01	92.8	92.8	10.2	0.4	2.5	53.2
	0.1	93.0	93.0	9.8	0.3	2.8	52.6
	1.0	92.8	92.7	10.4	0.7	2.4	52.1
DP Loss	0	93.2	92.8	14.5	4.8	4.8	52.9
	0.01	93.0	92.7	14.3	4.2	4.2	53.1
	0.3	93.0	93.0	9.8	0.3	2.8	52.6
	1.0	92.1	92.3	5.7	3.5	6.1	53.4

noise level, since the noise variance $\sigma^2 C^2$ scales with C (Table 10, rows 1-3). In this setting, smaller C values degrade representation-level fairness, as reflected by higher attacker accuracies. When controlling for noise variance (Table 10, rows 4-6), we observe that varying C produces no significant change in either task accuracy or fairness metrics, suggesting that the clipping operation itself has limited influence once the noise scale is fixed.

Table 10: **Sensitivity of embedding clipping threshold.** We evaluate the performance of FairNVT when the clipping threshold changes. All reported values are scaled by $\times 10^2$ and show performance from a single run with the same seed. Task: Expression (Smiling); Sensitive attribute: Gender (Male).

	Level	Acc(↑)	BAcc(↑)	DP(↓)	EOpp(↓)	EO(↓)	Att.Acc(↓)
Clip Threshold (with same noise multiplier)	1	93.1	93.1	9.7	0.4	2.9	89.0
	10	93.0	93.0	9.8	0.3	2.8	52.6
	100	92.3	92.2	10.6	0.9	1.9	53.1
Clip Threshold (with same noise amount)	1	92.8	92.8	9.8	0.2	2.9	52.6
	10	93.0	93.0	9.8	0.3	2.8	52.6
	100	92.8	92.8	9.7	0.3	2.7	52.2

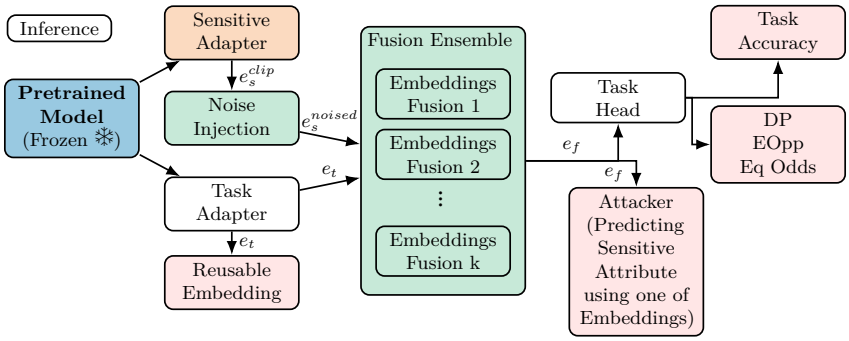


Figure 5: **An alternative inference time pipeline.** During inference, multiple noise samples produce a fused embedding e_f whose task predictions are aggregated by majority vote.

Sensitivity of number of noise draw at inference. Figure 5 shows an alternative inference time pipeline that is commonly used when random noise is drawn during training. Specifically, in the inference stage, multiple random noise can be drawn from the same distribution as optimized during training, where each noised sensitive embedding $e_{s,1}^{\text{noised}}, \ldots, e_{s,k}^{\text{noised}}$ is concatenated with the task embedding e_t to form k fused embeddings. The task prediction accuracy can be enhanced by majority voting from k fused embeddings. We examined such althernative pipeline in this section.

Table 11 reports results obtained when varying the number of noise draws during inference. The task prediction accuracies from a single noise draw are nearly identical to those from multiple draws, indicating that sensitive information is effectively disentangled from task-relevant features and that perturbing the sensitive subspace does not substantially alter task predictions. While majority voting over multiple noisy embeddings slightly improves task accuracy, it also marginally increases DP, EOpp, EO, and attacker accuracies, suggesting that aggregating multiple de-biased embeddings reintroduces a small amount of sensitive information. Overall, a single noise draw is sufficient to achieve strong task performance and fairness outcomes.

Table 11: **Sensitivity of number of noise draw at inference.** We evaluate the performance of FairNVT when the key hyperparameter value changes. All reported values are scaled by $\times 10^2$ and show performance from a single run with the same seed. Task: Expression (Smiling); Sensitive attribute: Gender (Male).

	Level	**Acc**(↑)	**BAcc**(↑)	**DP**(↓)	**EOpp**(↓)	**EO**(↓)	**Att.Acc**(↓)
Num. Noise Draw (Inference Time)	1	93.0	93.0	9.8	0.3	2.8	52.6
	10	93.2	93.2	9.9	0.4	3.3	53.3
	50	93.5	93.4	10.3	0.5	2.8	53.1

Table 12: **Image-Based Classification task:** Comparing our method with baselines on CelebA (Liu et al., 2015) dataset. All reported values are scaled by $\times 10^2$. We report the mean and standard deviation over 3 runs.

Method	Acc(↑)	BAcc(↑)	DP(↓)	EOpp(↓)	EO(↓)	Att.Acc(↓)
Vanilla	89.9 ±0.1	89.4±0.3	10.0±0.6	2.1±0.3	6.1±1.2	87.8±0.4
ViT-FSCL	88.7 ±0.1	88.0±0.1	7.4±0.8	**0.7±0.2**	2.5±0.1	87.4±0.1
FairViT	92.5±0.2	91.9±0.2	5.6±0.3	1.8±0.9	2.3±0.2	86.2±0.2
FairVPT	91.9±0.2	91.4±0.2	**1.7±1.0**	1.7±1.0	**2.0±0.3**	87.4±0.2
FairNVT(Ours)	**92.8±0.1**	**92.1±0.1**	5.8±0.3	1.7±1.0	2.2±0.2	**66.5±0.1**

(a) Task: Expression; Sensitive Attribute: Age (Young)

Method	Acc(↑)	BAcc(↑)	DP(↓)	EOpp(↓)	EO(↓)	Att.Acc(↓)
Vanilla	81.6 ±0.2	63.2±0.2	33.1±2.7	40.3±3.0	36.8±5.2	98.8±0.1
ViT-FSCL	80.4±1.1	64.8±0.0	24.7±0.2	35.2±0.1	24.7±0.2	97.8 ±0.1
FairViT	81.9±0.3	66.9±0.4	20.4±0.5	30.6±1.2	19.8±0.9	92.0±0.4
FairVPT	**83.0±0.5**	61.1±0.5	17.0±0.4	25.2±0.9	15.7±1.0	98.6±0.1
FairNVT(Ours)	81.2±0.1	**67.4±0.5**	**8.1±0.6**	**8.2±1.8**	**8.3±1.8**	**55.8±0.9**

(b) Task: Big Nose; Sensitive Attribute: Gender (Male)

Method	Acc(↑)	BAcc(↑)	DP(↓)	EOpp(↓)	EO(↓)	Att.Acc(↓)
Vanilla	84.4±0.6	81.2±0.5	10.3±1.0	8.5±1.0	7.7 ±2.0	87.7±0.4
ViT-FSCL	83.3 ±0.6	77.9±2.0	**5.3±0.8**	**2.0±0.4**	**1.3±0.3**	87.4±0.1
FairViT	86.6±0.4	83.7±0.3	9.0±0.5	3.5±1.6	2.8±0.6	86.4±0.3
FairVPT	84.2±0.6	82.2±0.4	7.9±0.6	2.8±1.0	2.9±0.5	88.1±0.2
FairNVT(Ours)	**87.0±0.5**	**84.0±0.5**	7.1±0.7	3.4±0.9	2.4±0.5	**66.8±0.1**

(c) Task: Wavy hair; Sensitive Attribute: Age(Young)

Method	Acc(↑)	BAcc(↑)	DP(↓)	EOpp(↓)	EO(↓)	Att.Acc(↓)
Vanilla	99.1 ±0.1	94.1±0.3	10.8±0.1	5.3±1.5	4.3 ±1.3	98.7±0.1
ViT-FSCL	99.1±0.1	95.0±0.2	10.8±0.4	3.9±0.4	2.4±0.2	97.9±0.2
FairViT	99.0±0.2	98.0±0.2	10.0±0.3	0.8±0.3	**0.6±0.3**	97.6±0.2
FairVPT	99.4±0.1	97.0±0.2	**9.6±0.3**	2.1±0.3	1.2±0.4	98.5±0.1
FairNVT(Ours)	**99.6±0.0**	**98.7±0.0**	11.2±0.1	**0.7±0.4**	0.7±0.4	**53.8±0.3**

(d) Task: Wearing glasses; Sensitive Attribute: Gender (Male)

Method	Acc(↑)	BAcc(↑)	DP(↓)	EOpp(↓)	EO(↓)	Att.Acc(↓)
Vanilla	99.1 ±0.1	95.4±0.2	13.7±0.2	7.0±0.6	6.3±1.6	88.1±0.2
ViT-FSCL	99.0±0.1	95.1±1.1	13.1±0.5	5.9±0.1	3.3±0.2	87.5±0.1
FairViT	99.1±0.3	**97.4±0.4**	13.0±0.7	2.7±0.6	2.9±0.5	89.4±0.6
FairVPT	**99.4±0.1**	96.8±0.3	12.9±0.7	**1.2±1.0**	**2.4±0.7**	87.6±0.3
FairNVT(Ours)	99.3±0.2	96.9±0.5	**12.4±1.0**	3.0±1.1	2.8±1.3	**67.3±0.1**

(e) Task: Wearing Glasses; Sensitive Attribute: Age (Young)

Method	Acc(↑)	BAcc(↑)	DP(↓)	EOpp(↓)	EO(↓)	Att.Acc(↓)
Vanilla	85.5±0.3	85.2±0.4	9.6±0.4	4.7±0.6	4.6±0.9	98.7±0.0
ViT-FSCL	82.6±0.5	81.8±0.6	7.5±2.0	1.3±0.7	2.5±1.7	97.6 ±0.1
FairViT	93.4±0.1	93.3±0.2	9.0±0.4	1.6±0.3	1.5±0.4	96.1±0.4
FairVPT	92.7±0.1	92.7±0.1	9.3±0.3	**0.3±0.8**	**0.6±0.7**	98.5±0.0
FairNVT(Ours)	**93.7±0.1**	**93.7±0.1**	**6.3±0.8**	0.9±0.1	1.5±0.6	**52.4±0.6**

(f) Task: Mouth Slightly Open; Sensitive Attribute: Gender(Male)

Method	Acc(↑)	BAcc(↑)	DP(↓)	EOpp(↓)	EO(↓)	Att.Acc(↓)
Vanilla	84.7±1.8	83.9±1.7	7.4±1.5	1.8±1.3	6.1±1.8	85.2±3.9
ViT-FSCL	83.8±0.1	82.9±0.1	5.7±1.2	1.5±1.3	2.3±1.1	87.4±0.1
FairViT	93.4±0.3	93.1±0.2	7.0±0.3	0.8±0.2	0.9±0.3	82.3±0.2
FairVPT	92.0±0.1	91.8±0.1	**4.4±0.5**	1.8±1.0	1.0±0.4	87.5±0.2
FairNVT(Ours)	**94.0±0.0**	**93.7±0.2**	4.6±0.2	**0.3±0.1**	**0.6±0.1**	**65.8±0.1**

(g) Task: Mouth Slightly Open; Sensitive Attribute: Age(Young)

Table 13: **Qualitative BIOS examples.** We show pairs of biography snippets that differ only in gender indicators. For each snippet, we display the model's predicted occupation and the prediction score for the ground-truth label. Vanilla model predictions vary substantially across genders, suggesting reliance on gender cues, whereas FairNVT yields more stable scores and consistent predictions, indicating improved robustness to gender indicators.

ID	BIO Snippet	Vanilla	FairNVT
1	He specializes in development economics, household economics, and personnel economics. In 2003 he received his Ph.D. in Economics from the London School of Economics...	professor (0.903)	professor (0.992)
1	She specializes in development economics, household economics, and personnel economics. In 2003 she received her Ph.D. in Economics from the London School of Economics...	professor (0.882)	professor (0.993)
2	Prosper was born and raised in Miami Beach, FL. He received his Bachelor's degree from Emory University and graduated with honors from the University of Miami School of Law...	attorney (0.971)	attorney (0.971)
2	Prosper was born and raised in Miami Beach, FL. She received her Bachelor's degree from Emory University and graduated with honors from the University of Miami School of Law...	attorney (0.939)	attorney (0.970)
3	She has been travelling the world, and worked, amongst others, on a documentary photography project in India with an orphanage...	photographer (0.643)	photographer (0.908)
3	He has been travelling the world, and worked, amongst others, on a documentary photography project in India with an orphanage...	photographer (0.729)	photographer (0.864)
4	She studied at EFET Paris and NYU New-York respectively. While working in a post-production, she develops her own photographic concept...	photographer (0.664)	photographer (0.966)
4	He studied at EFET Paris and NYU New-York respectively. While working in a post-production, he develops his own photographic concept...	photographer (0.804)	photographer (0.941)
5	He attended the University of California, San Francisco (UCSF), School of Medicine and subsequently trained at Children's Hospital Los Angeles for residency...	physician (0.717)	physician (0.818)
5	She attended the University of California, San Francisco (UCSF), School of Medicine and subsequently trained at Children's Hospital Los Angeles for residency...	physician (0.826)	physician (0.755)

ID	BIO Snippet	Vanilla	FairNVT
6	Dr. Cottrell attended medical school at the University of Missouri-Columbia School of Medicine. He is in-network for Anthem, Blue Cross/Blue Shield, Blue Shield, and more.	physician (0.773)	physician (0.923)
6	Dr. Cottrell attended medical school at the University of Missouri-Columbia School of Medicine. She is in-network for Anthem, Blue Cross/Blue Shield, Blue Shield, and more.	physician (0.791)	physician (0.920)
7	After spending two years at the University of Iowa, Kyle transferred to Chapman University, where he directed a superhero noir titled The League, about the 1960's superhero labor union of Chicago...	filmmaker (0.944)	filmmaker (0.808)
7	After spending two years at the University of Iowa, Kylie transferred to Chapman University, where she directed a superhero noir titled The League, about the 1960's superhero labor union of Chicago...	filmmaker (0.925)	filmmaker (0.866)
8	She has been a successful Dentist for the last 16 years. She is a BDS. She is currently associated with SMII Dental Art Studio in Koregaon Park, Pune...	dentist (0.991)	dentist (0.997)
8	He has been a successful Dentist for the last 16 years. He is a BDS. He is currently associated with SMII Dental Art Studio in Koregaon Park, Pune...	dentist (0.966)	dentist (0.943)
9	Downs was a fellow at Northwestern University's Academy for Alternative Journalism in 2004, and he earned a degree in English literature from University of California at Santa Barbara in 2002...	journalist (0.437)	journalist (0.917)
9	Downs was a fellow at Northwestern University's Academy for Alternative Journalism in 2004, and she earned a degree in English literature from University of California at Santa Barbara in 2002...	journalist (0.480)	journalist (0.885)
10	She graduated with honors in 2012. Having more than 5 years of diverse experiences, especially in NURSE PRACTITIONER, Melissa R Kludt affiliates with many hospitals including...	nurse (0.917)	nurse (0.890)
10	He graduated with honors in 2012. Having more than 5 years of diverse experiences, especially in NURSE PRACTITIONER, Miles R Kludt affiliates with many hospitals including...	nurse (0.624)	nurse (0.926)