

Protecting the Underrepresented: Data Debiasing via Asymmetric Adversarial Encoding for Fair Machine Learning

Deepanshu Aggarwal, Puspita Majumdar

Mastercard, India

deepanshu.aggarwal@mastercard.com, puspita.majumdar@mastercard.com

Abstract

Machine learning systems deployed in high-stakes domains often disadvantage underrepresented demographic groups because the training data itself encodes historical bias. Existing preprocessing methods typically transform all groups symmetrically or rebalance sample counts, without accounting for a basic asymmetry: the unprivileged group is often represented by fewer samples and may therefore be more vulnerable to additional distortion. We propose AdvEncoder, a training-time data debiasing framework based on asymmetric alignment: the latent reference space is learned from the unprivileged group, and the privileged group is adversarially aligned toward that reference. The goal is to reduce group-discriminative structure in the training data while minimizing additional distortion to the data-scarce group. AdvEncoder has two variants distinguished only by the discriminator’s conditioning. AdvEncoder-SP uses an unconditional discriminator and targets marginal alignment related to statistical parity, while AdvEncoder-EO uses a label-conditional discriminator and targets class-conditional alignment related to equal opportunity and equalized odds. We provide formal implications under exact and approximate alignment, while emphasizing that these are alignment-level guarantees rather than optimization or deployment guarantees. AdvEncoder is used only at training time: it outputs a debiased training set in the original feature space, after which any downstream classifier can be trained normally and evaluated on ordinary, untransformed test or deployment inputs. Across four tabular benchmarks (UCI Adult, LSAC, ACSIncome, and German Credit) and five downstream classifier families, AdvEncoder provides a strong fairness-utility trade-off. It does not uniformly dominate every baseline on every individual metric; rather, it consistently achieves near-best fairness while retaining competitive accuracy and avoiding the severe utility loss or single-metric failure modes observed for several baselines. Ablations confirm that both adversarial alignment and the direction of asymmetric anchoring toward the unprivileged group are important to the method’s effectiveness.

1 Introduction

Machine learning systems now mediate consequential decisions across nearly every domain of public life—credit eligibility, hiring, bail and sentencing, and university admission. The integration of these systems at scale makes their

fairness not merely a technical concern but a matter of social justice. An important and often underappreciated source of unfairness is the training data itself: when historical data reflects structural inequalities, models trained on it learn to reproduce and sometimes amplify those inequalities (Mehrabi et al. 2021; Barocas, Hardt, and Narayanan 2023).

A natural response is to intervene at the data level—a class of techniques known as preprocessing methods—spanning relabeling, perturbation, sampling, and reweighting (Kamiran and Calders 2009; Wang, Ustun, and Calmon 2019; Chawla et al. 2002; Kamiran and Calders 2012; Chai and Wang 2022). While these methods can improve aggregate fairness metrics, they share a fundamental limitation: they treat the data transformation as a *symmetric* problem, modifying all groups uniformly or optimizing a global objective that does not distinguish between groups. Representation learning approaches and optimal transport-based methods similarly process all groups through the same encoder or move both groups toward a shared barycenter (Oh et al. 2022; Islam, Chen, and Cai 2024; Feldman et al. 2015; Gordaliza et al. 2019). None of these methods account for a basic asymmetry: the unprivileged group is often represented by fewer samples and is therefore more vulnerable to additional distortion.

A concrete example. Consider the UCI Adult income-prediction task, where female applicants (the unprivileged group) are outnumbered roughly 2:1 by male applicants. A symmetric debiasing method would blur both groups toward a shared midpoint, distorting the already data-scarce female distribution and risking the loss of the limited statistical signal it carries. Our approach instead freezes a reference representation learned entirely from female data and pulls male records toward that reference—stripping the male-correlated features (e.g., occupation and hours-worked patterns reflecting historical labor-market segregation) that drive the disparity, while leaving the female distribution largely intact. The result is a debiased training set in which group membership is harder to predict from features, yet the information content of the smaller group is preserved.

Asymmetric alignment. Our motivating intuition is conservative rather than normative. The unprivileged group’s empirical distribution is supported by fewer samples and is therefore most at risk of information loss under aggressive transformation. The privileged group’s distribution, by contrast, often carries stronger signals of group membership and

historical advantage. We therefore propose to use the unprivileged group’s distribution as a *statistical anchor*—not as a normative target—and align only the privileged group toward it.

Our approach. We propose **AdvEncoder**, a training-time preprocessing framework that implements asymmetric alignment through a two-phase architecture. Phase 1 trains an autoencoder and main task classifier on unprivileged-group data, establishing a latent reference space. Phase 2 trains an adversarial encoder that maps privileged-group instances into this same space via a discriminator. **AdvEncoder-SP** uses an unconditional discriminator (marginal alignment); **AdvEncoder-EO** uses a label-conditional discriminator (class-conditional alignment). The label is required only during training-time construction of the debiased dataset.

AdvEncoder outputs a debiased *training* dataset in the original feature space. Downstream models are trained on this data and evaluated on ordinary, untransformed test examples—making AdvEncoder a training-set debiasing method rather than an input-sanitization method.

Contributions.

- We introduce the *asymmetric alignment principle*: learn a reference representation from the unprivileged group and align the privileged group toward it, placing less transformation burden on the data-scarce group.
- We instantiate this principle as **AdvEncoder**, a two-phase framework that outputs a debiased training dataset in the original feature space, requiring no architectural modification or test-time transformation.
- We propose two variants (SP and EO) targeting marginal and class-conditional alignment respectively, with formal implications under exact and approximate alignment.
- Across four tabular benchmarks and five classifier families, AdvEncoder achieves a strong fairness–utility trade-off, attaining near-best fairness with competitive accuracy. Ablations show that adversarial alignment and the direction of anchoring are both important.

2 Related Work

Bias mitigation strategies fall into *preprocessing*, *inprocessing*, and *postprocessing* categories (Mehrabi et al. 2021). Our approach is preprocessing.

Preprocessing approaches. Instance-modification methods include massaging (Kamiran and Calders 2009), k -NN label reassignment (Luong, Ruggieri, and Turini 2011), and feature perturbation (Wang, Ustun, and Calmon 2019). Distribution-modification methods include SMOTE (Chawla et al. 2002), CTGAN (Xu et al. 2019), and FAW (Chai and Wang 2022). While these can close accuracy gaps, they do not eliminate the distributional signal enabling discriminatory predictions (Wang et al. 2019), and treat group transformations symmetrically. A more powerful class maps data into a fair latent space: LFR (Zemel et al. 2013), VFAE (Louizos et al. 2016), adversarial variants (Edwards and Storkey 2016), FairGAN (Xu et al. 2018), and FUR (Kairouz et al. 2022). Most produce latent representations rather than feature-space data, and apply the same transformation to all groups.

Method	Type	Asym.	Min. U.	Output
LFR	Pre	✗	✗	Repr.
FairGAN	Pre	✗	✗	Synth. data
FUR	Pre	✗	✗	Repr.
FAW	Pre	✗	✗	Reweighted
FarconVAE	Pre/In	✗	✗	Repr.
LAFTTR	Pre	✗	✗	Repr.
AdapFair	Pre	✗	✗	Repr.
SLSD	In	✗	✗	Model
DAP [†]	In	✗	✗	Model
Ours	Pre	✓	✓	Data

Table 1: Theoretical comparison of AdvEncoder with related methods. **Type**: Pre = Preprocessing, In = Inprocessing. **Asym.**: applies asymmetric treatment to privileged vs. unprivileged groups. **Min. U.**: explicitly minimizes transformation of the unprivileged group. [†] DAP is an inprocessing reference point.

FarconVAE and SLSD. FarconVAE (Oh et al. 2022) disentangles sensitive and non-sensitive latents using paired observations, but is symmetric by design and unstable under high group imbalance (e.g., LSAC’s 8:1 ratio). Because it produces latents rather than feature-space data, like-for-like preprocessing comparison would require either training on its latents or adding a separate decoder. We therefore test the symmetric-vs-asymmetric distinction directly via our design-choice ablation (Section 6). SLSD (Islam, Chen, and Cai 2024) addresses domain alignment without sensitive labels and is complementary rather than competing.

Inprocessing methods. DAP (Sahyouni et al. 2025) introduces a differentiable adjusted-parity loss but is an *inprocessing* method: it cannot produce a debiased dataset for arbitrary downstream models. We include DAP as a reference point ([†] in tables) illustrating the inprocessing performance envelope.

Table 1 summarizes how AdvEncoder differs from prior work along the key dimensions of asymmetric treatment, minimization of unprivileged-group distortion, and feature-space output.

3 Proposed Method

We propose **AdvEncoder**, a preprocessing framework for fair machine learning that transforms biased training data by aligning the privileged group’s distribution with that of the unprivileged group in a shared latent space. The framework admits two variants that share an identical Phase 1 and the asymmetric alignment principle, and differ only in how the Phase 2 discriminator is conditioned: **AdvEncoder-SP** targets statistical parity via an unconditional discriminator, and **AdvEncoder-EO** targets equal opportunity and equalized odds via a label-conditional discriminator.

3.1 Problem Formulation

Let $\mathcal{D}_{\text{train}} = \{(x_i, s_i, y_i)\}_{i=1}^n$ be a training dataset with feature vector $x_i \in \mathbb{R}^m$, a *single binary* sensitive attribute $s_i \in \{0, 1\}$, and binary target $y_i \in \{0, 1\}$. By convention, $s = 0$ denotes the *unprivileged* group $\mathcal{D}_{s=0}$ (size n_0) and $s =$

1 the *privileged* group $\mathcal{D}_{s=1}$ (size n_1); in all settings we study $n_1 > n_0$, so the unprivileged group is the data-scarce one. We assume a single binary sensitive attribute throughout and defer multi-class and intersectional extensions to Section 8.

The goal of AdvEncoder is to construct a debiased *training* dataset

$$\hat{\mathcal{D}}_{\text{train}} = \{(\hat{x}_i, s_i, y_i)\}_{i=1}^n,$$

which can be used to train any downstream classifier. Importantly, AdvEncoder is not applied to test or deployment inputs. After a downstream model is trained on $\hat{\mathcal{D}}_{\text{train}}$, it is evaluated on ordinary untransformed examples from the test or deployment distribution. This protocol matches the practical use case in which a data owner releases a debiased training set to modeling teams, who then train standard models without modifying their inference pipelines.

AdvEncoder has two variants. AdvEncoder-SP encourages marginal alignment in the debiased training data,

$$p(\hat{x} \mid s = 0) \approx p(\hat{x} \mid s = 1), \quad (1)$$

which is related to statistical parity. AdvEncoder-EO encourages class-conditional alignment,

$$p(\hat{x} \mid y = c, s = 0) \approx p(\hat{x} \mid y = c, s = 1), \quad c \in \{0, 1\}, \quad (2)$$

which is related to equal opportunity and equalized odds.

The label y used by AdvEncoder-EO is required only when constructing the debiased training dataset. No label is required at test time because the learned encoder, discriminator, and decoder are not used during downstream inference.

3.2 Asymmetric Alignment Principle

Most distribution-alignment approaches for fairness move all groups toward a shared midpoint—either via optimal transport to a Wasserstein barycenter (Feldman et al. 2015; Gordaliza et al. 2019) or by applying a symmetric encoder to all data (Oh et al. 2022; Louizos et al. 2016). We argue this symmetry is inappropriate when groups differ in representational size.

The unprivileged group ($n_0 \ll n_1$) carries limited information about the true population distribution $p(x \mid s = 0)$. Any transformation applied to $\mathcal{D}_{s=0}$ can only preserve or reduce this information; it cannot recover signal absent due to underrepresentation. The privileged group’s empirical distribution $\hat{p}(x \mid s = 1)$, by contrast, over-encodes historically biased correlations between features and group membership—precisely the signal that enables discriminatory predictions. This motivates a principled asymmetry: *anchor* the unprivileged group’s distribution as the reference and *move only the privileged group* toward it. Formally, we seek a transformation G such that

$$p(G(\mathcal{D}_{s=1})) \approx p(\text{Enc}(\mathcal{D}_{s=0})), \quad (3)$$

where Enc is an encoder trained exclusively on $\mathcal{D}_{s=0}$.

Asymmetry of degree, not kind. The unprivileged group’s data is passed through the trained encoder-decoder pair and therefore undergoes a small reconstruction distortion; the privileged group undergoes the same encoder-decoder *plus* the adversarial alignment shift. The asymmetry is therefore

Dataset	Unprivileged MSE	Privileged MSE
UCI Adult	0.031±0.002	0.187±0.008
LSAC	0.018±0.001	0.142±0.006
ACSIncome	0.027±0.002	0.163±0.009
German Credit	0.044±0.004	0.231±0.014

Table 2: Per-group reconstruction MSE on the test set (mean±std over 5 seeds), in standardized feature space. The privileged group consistently incurs 4–6× higher MSE, confirming the asymmetric magnitude of transformation.

one of *magnitude*, not a binary between modified and unmodified data. We empirically verify this property in Table 2: the privileged group’s reconstruction MSE is 4–6× that of the unprivileged group across all four benchmarks, confirming that the alignment burden falls on the privileged group as designed.

Applicability. The principle is most clearly motivated when $n_0 < n_1$, the regime of all four benchmarks in our evaluation. When $n_0 \approx n_1$, the anchor choice should be guided by domain knowledge about historical disadvantage. We also clarify that we do not claim the unprivileged empirical distribution is normatively ideal or free from the effects of historical discrimination—we use it as a *conservative statistical reference*, since it is the distribution we have the least data about and therefore the one most at risk of additional distortion.

3.3 Architecture

AdvEncoder comprises five components trained in two sequential phases (Figures 1 and 2):

- **Encoder** $\text{Enc} : \mathbb{R}^m \rightarrow \mathbb{R}^h$, mapping unprivileged data to the latent reference space.
- **Decoder** $\text{Dec} : \mathbb{R}^h \rightarrow \mathbb{R}^m$, reconstructing data from latent representations.
- **Main task classifier** $C : \mathbb{R}^h \rightarrow [0, 1]$, a regularizer that ensures the latent space retains task-relevant information; not used at inference.
- **Adversarial encoder** $G : \mathbb{R}^m \rightarrow \mathbb{R}^h$, mapping privileged data into the same latent space.
- **Discriminator** Dis , predicting group membership from latent representations. Its input conditioning determines the fairness criterion (Section 3.5).

3.4 Phase 1: Anchored Autoencoder

In Phase 1, the autoencoder $f_\theta = (\text{Enc}, \text{Dec})$ and classifier C are trained *exclusively* on $\mathcal{D}_{s=0}$, establishing the latent reference distribution. Let $H_{s_0} = \text{Enc}(\mathcal{D}_{s=0})$. The decoder is trained with mean squared error reconstruction loss,

$$\mathcal{L}_{\text{rec}} = \frac{1}{n_0} \sum_{d_i \in \mathcal{D}_{s=0}} \|d_i - \text{Dec}(H_{s_0, i})\|^2, \quad (4)$$

and the classifier minimizes binary cross-entropy on H_{s_0} ,

$$\mathcal{L}_{\text{task}} = -\frac{1}{n_0} \sum_{i=1}^{n_0} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)], \quad (5)$$

Algorithm 1 AdvEncoder training-data debiasing

Require: Data $\mathcal{D}$; variant $v \in \{\text{SP}, \text{EO}\}$
1: Split $\mathcal{D}$ into $\mathcal{D}_0$ ($s=0$), $\mathcal{D}_1$ ($s=1$)
2: Train Enc, Dec, C on $\mathcal{D}_0$; freeze Enc, Dec
3: **if** $v = \text{SP}$ **then**
4: Train G, Dis_{SP} via Eq. (6)
5: **else**
6: Train G, Dis_{EO} via Eq. (7) (label-stratified batches)
7: **end if**
8: **for each** (x_i, s_i, y_i) **do**
9: $\hat{x}_i \leftarrow Dec(Enc(x_i))$ if $s_i=0$, else $Dec(G(x_i))$
10: Project categorical blocks of $\hat{x}_i$ to valid one-hot
11: **end for**
12: **return** $\hat{\mathcal{D}} = \{(\hat{x}_i, s_i, y_i)\}_{i=1}^n$

ensuring the latent space encodes task-relevant information. The total Phase 1 objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{task}}$. Once Phase 1 converges, all autoencoder parameters are frozen.

3.5 Phase 2: Adversarial Alignment

The adversarial encoder G and discriminator Dis are trained in a GAN-style min-max game (Goodfellow et al. 2014). The frozen Phase 1 encoder maps unprivileged training data to $H_{s_0} = Enc(\mathcal{D}_{s=0})$, while G maps privileged training data to $H_{s_1} = G(\mathcal{D}_{s=1})$. The discriminator receives either H alone or (H, y) , depending on the fairness criterion targeted by the variant. The label-conditioned variant uses labels only during this training-time debiasing phase.

AdvEncoder-SP (Unconditional Discriminator) Here $Dis_{\text{SP}} : \mathbb{R}^h \rightarrow \{0, 1\}$ predicts s from H alone:

$$\min_G \max_{Dis_{\text{SP}}} V_{\text{SP}}(G, Dis_{\text{SP}}) = \mathbb{E}_{x \sim p(\mathcal{D}_{s=0})} [\log Dis_{\text{SP}}(Enc(x))] + \mathbb{E}_{z \sim p(\mathcal{D}_{s=1})} [\log(1 - Dis_{\text{SP}}(G(z)))]. \quad (6)$$

At the equilibrium of this game, the marginal latent distributions of the two groups coincide, directly targeting statistical parity (Proposition 1).

AdvEncoder-EO (Conditional Discriminator) Here $Dis_{\text{EO}} : \mathbb{R}^h \times \{0, 1\} \rightarrow \{0, 1\}$ predicts s from (H, y) :

$$\begin{aligned} \min_G \max_{Dis_{\text{EO}}} V_{\text{EO}}(G, Dis_{\text{EO}}) &= \mathbb{E}_{(x,y) \sim p(\mathcal{D}_{s=0})} [\log Dis_{\text{EO}}(Enc(x), y)] \\ &+ \mathbb{E}_{(z,y) \sim p(\mathcal{D}_{s=1})} [\log(1 - Dis_{\text{EO}}(G(z), y))]. \end{aligned} \quad (7)$$

Conditioning on y forces G to match the *class-conditional* latent distributions separately, which is the condition required for equal opportunity and equalized odds (Proposition 2). AdvEncoder-EO thus enforces a strictly stronger condition than AdvEncoder-SP: no group identity signal *within each outcome class*, rather than only on average.

Discriminator design. Both variants use a *single* discriminator over the binary split. This keeps the min-max game simple but, like any GAN objective, carries a risk of mode collapse in G or residual group signal that the discriminator fails to penalize; we return to these risks, and to multi-discriminator extensions for intersectional settings, in Section 8.

3.6 Fairness Implications Under Exact and Approximate Alignment

We now state the fairness implications of the alignment objectives. These results should be interpreted carefully. They are not convergence guarantees for the adversarial optimization procedure, nor are they assumption-free deployment guarantees. Rather, they characterize what follows if the debiased training distributions produced by AdvEncoder are exactly or approximately aligned.

Let $T_0(x) = Dec(Enc(x))$ denote the transformation applied to unprivileged training instances, and let $T_1(x) = Dec(G(x))$ denote the transformation applied to privileged training instances. Let $\hat{X} = T_S(X)$ denote the transformed training feature vector. The downstream classifier is trained on samples from this debiased training distribution. In experiments, the trained classifier is then evaluated on untransformed test data, matching the training-time debiasing protocol described in Section 3.1.

Proposition 1 (Exact marginal alignment implies statistical parity on the aligned distribution). *If AdvEncoder-SP produces transformed training features satisfying $p(\hat{X} | s = 0) = p(\hat{X} | s = 1)$, then for any measurable classifier f evaluated on the aligned distribution,*

$$|p(f(\hat{X}) = 1 | s = 0) - p(f(\hat{X}) = 1 | s = 1)| = 0.$$

Proof sketch. If the group-conditional distributions of $\hat{X}$ are identical, then any measurable function of $\hat{X}$ has the same distribution across groups. Thus the positive prediction rate of $f(\hat{X})$ is identical for $s = 0$ and $s = 1$. $\square$

Proposition 2 (Exact class-conditional alignment implies equal opportunity and equalized odds on the aligned distribution). *If AdvEncoder-EO produces transformed training features satisfying $p(\hat{X} | y = c, s = 0) = p(\hat{X} | y = c, s = 1)$ for each $c \in \{0, 1\}$, then for any measurable classifier f evaluated on the aligned distribution,*

$$p(f(\hat{X}) = 1 | y = c, s = 0) = p(f(\hat{X}) = 1 | y = c, s = 1) \text{ for } c \in \{0, 1\}. \text{ Therefore equal opportunity difference and equalized odds difference are both zero on the aligned distribution.}$$

Proof sketch. For each label class c , the conditional feature distributions are identical across groups. Any measurable classifier therefore induces identical conditional prediction distributions across groups within that class. The case $c = 1$ gives equal true positive rates, and the case $c = 0$ gives equal false positive rates. $\square$

We also state the approximate version, which is more relevant in practice. Let $d_{\text{TV}}(P, Q)$ denote total variation distance.

Proposition 3 (Approximate alignment bounds fairness gaps). *If*

$$d_{\text{TV}}(p(\hat{X} | s = 0), p(\hat{X} | s = 1)) \leq \epsilon,$$

then for any measurable classifier f ,

$$\text{SP}(f) \leq \epsilon$$

on the aligned distribution. Similarly, if for each $c \in \{0, 1\}$,

$$d_{TV}(p(\hat{X} | y = c, s = 0), p(\hat{X} | y = c, s = 1)) \leq \epsilon_c,$$

then

$$\text{E.Opp}(f) \leq \epsilon_1, \quad \text{E.Odds}(f) \leq \epsilon_1 + \epsilon_0$$

on the aligned distribution.

Proof sketch. For any event A , total variation bounds $|P(A) - Q(A)|$. Taking $A = \{\hat{x} : f(\hat{x}) = 1\}$ gives the SP bound. Applying the same argument within each label stratum gives the equal opportunity and equalized odds bounds. $\square$

Interpretation. These propositions justify the choice of discriminator conditioning: unconditional discrimination targets marginal alignment, while label-conditioned discrimination targets class-conditional alignment. However, they do not prove that the GAN objective reaches the aligned distribution, and they do not by themselves guarantee fairness under arbitrary deployment-time distribution shift. Our empirical evaluation therefore directly measures fairness on held-out, untransformed test data after training downstream models on the debiased training set.

Remark on the SP/EO trade-off. AdvEncoder-EO’s class-conditional alignment does not generally imply statistical parity when $p(y | s = 0) \neq p(y | s = 1)$, because the marginal distributions $p(\hat{X} | s) = \sum_c p(\hat{X} | y = c, s)p(y = c | s)$ may still differ. This reflects the known incompatibility between statistical parity and equalized-odds-style criteria under unequal base rates (Chouldechova 2017). Practitioners should therefore choose the variant matching the deployment goal.

3.7 Debiased Data Reconstruction

Once Phase 2 converges, both groups’ latent representations are decoded by the shared decoder to produce the debiased dataset:

$$\hat{\mathcal{D}} = \text{Dec}(H_{s_0}, H_{s_1}). \quad (8)$$

The decoder, although trained only on unprivileged data, is applied to both groups because G ’s outputs are trained to occupy the same region of latent space as Enc ’s outputs. Categorical feature blocks are projected back to valid one-hot vectors after decoding. The sensitive attribute s is retained in $\hat{\mathcal{D}}$ to allow post-hoc auditing.

3.8 Implementation Details

Architectures. Enc is a single fully-connected layer $\mathbb{R}^m \rightarrow \mathbb{R}^{16}$ with ReLU; Dec is its mirror. C is a single linear layer with sigmoid output. G is a two-layer network with hidden dimensions [32, 16] and ReLU. Dis_{SP} is a two-layer network with hidden dimension 8 and sigmoid output, taking $H \in \mathbb{R}^{16}$. Dis_{EO} is identical except its input dimension is 17, formed by concatenating H with the label y .

Training. Phase 1 trains for 25 epochs with batch size 64; Phase 2 for 50 epochs with batch size 16. All parameters use Adam ($\text{lr} = 10^{-3}$). Categorical features are one-hot encoded; numerical features are standardized to zero mean and unit variance, fit on the training set only. Implementation is in TensorFlow.

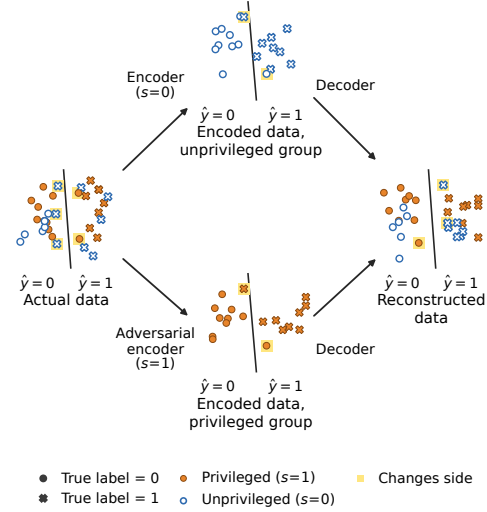


Figure 1: Visual illustration of AdvEncoder. Models trained on original data produce biased outcomes (*left*). Phase 1 trains an autoencoder on the unprivileged group, establishing a reference latent distribution (*top*). Phase 2 aligns the privileged group’s latent representation via adversarial training (*bottom*); AdvEncoder-SP uses an unconditional discriminator, AdvEncoder-EO uses a label-conditional one. The decoder reconstructs a debiased dataset (*right*).

Dataset	Sens.	Train	Test	Pos.	Dim	Priv/Unpriv.
Adult	Gender	30,888	7,722	24.1%	49	Male/Female
LSAC	Race	17,926	4,481	94.5%	8	White/Non-white
ACSIncome	Sex	156,532	39,133	37.2%	9	Male/Female
German	Age	800	200	70.0%	47	Older/Younger

Table 3: Dataset summary. The unprivileged group is the smaller group in all four experimental settings.

4 Experimental Setup

4.1 Datasets

We evaluate on four publicly available tabular benchmarks (Table 3). In all settings, the unprivileged group is the smaller group in training data. **UCI Adult** (Becker and Kohavi 1996): 48,842 instances, 14 raw attributes, predicting income $> \$50K$; sensitive attribute is sex (female unprivileged). **LSAC** (Wightman 1998): 22,407 law school records, predicting bar passage; race (non-white unprivileged). **ACSIncome** (Ding et al. 2021): California 2018 ACS via folkttables, $\$50K$ threshold; sex (female unprivileged). **German Credit** (Hofmann 1994): 1,000 loan applications with binary credit-risk labels; age (younger unprivileged). All datasets use 80/20 train-test splits with five random seeds. All preprocessing methods are fit only on the training split; metrics are computed on the held-out test split using untransformed test features.

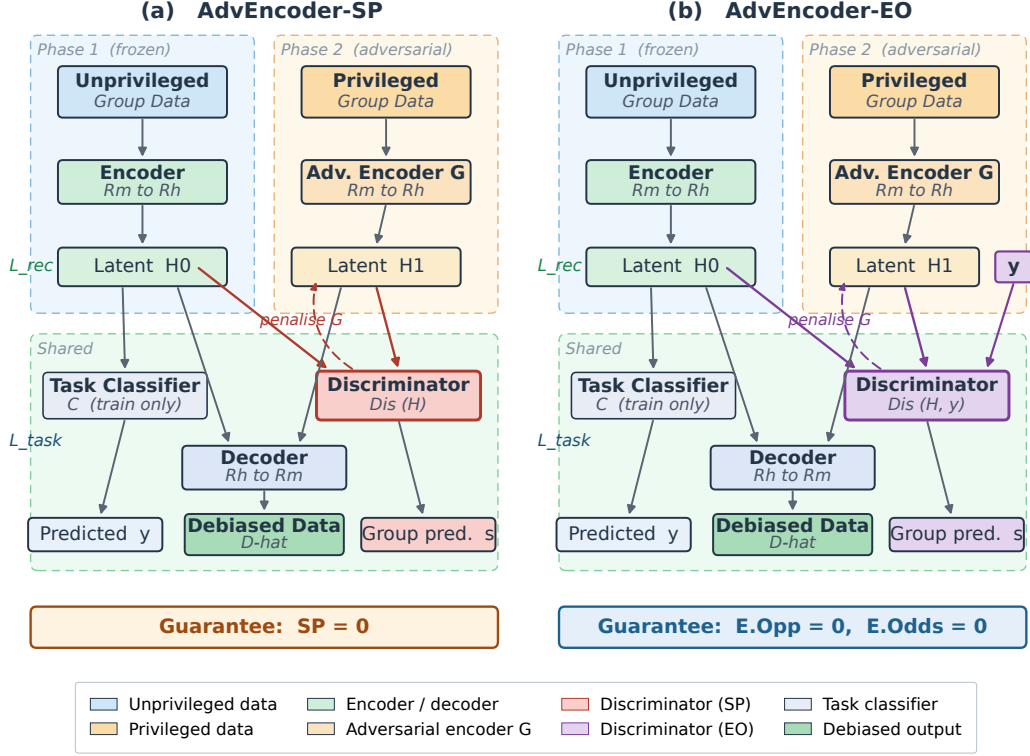


Figure 2: Block diagram of the AdvEncoder framework. (a) AdvEncoder-SP uses an unconditional discriminator $Dis(H)$, targeting marginal latent alignment (statistical parity). (b) AdvEncoder-EO uses a label-conditional discriminator $Dis(H, y)$, targeting class-conditional alignment (equal opportunity and equalized odds). All components are used only at training time.

4.2 Evaluation Metrics

We use accuracy as the utility metric and three standard group fairness metrics. Let $\hat{y}$ denote the prediction, y the true label, s the sensitive attribute. **Statistical Parity Difference** (Dwork et al. 2012): $SP = |p(\hat{y}=1 | s=1) - p(\hat{y}=1 | s=0)|$. **Equal Opportunity Difference** (Hardt, Price, and Srebro 2016): $E.Opp = |p(\hat{y}=1 | y=1, s=1) - p(\hat{y}=1 | y=1, s=0)|$. **Equalized Odds** (Hardt, Price, and Srebro 2016) is the sum of absolute TPR and FPR differences across groups. Lower values indicate fairer predictions.

4.3 Baselines

We compare against eleven methods spanning every major preprocessing family, selecting a widely-used representative of each so that comparisons reflect design paradigms rather than implementation quirks. *Sampling-based* preprocessing: **Undersampling**, **SMOTE** (Chawla et al. 2002), **CTGAN** (Xu et al. 2019). *Fairness-aware* preprocessing: **LFR** (Zemel et al. 2013), **FUR** (Kairouz et al. 2022), **FAW** (Chai and Wang 2022), **LAFTR** (Madras et al. 2018), **AdapFair** (Huang et al. 2024). We additionally include **Zhang** (Zhang, Lemoine, and Mitchell 2018) (adversarial fairness baseline) and **DAP[†]** (Sahyouni et al. 2025) (inprocessing reference, [†]). **Baseline** denotes the unmodified-data results. The fair-representation family (LFR, FUR, LAFTR, AdapFair) is well represented, covering both non-adversarial

and adversarial encoders; FarconVAE-style symmetric encoders output latents rather than feature-space data and are thus not like-for-like preprocessing competitors, so we instead isolate the symmetric-vs-asymmetric design axis directly via the Symmetric AE ablation (Section 6), which holds our architecture and adversarial alignment fixed and varies only the anchoring. All baselines are run under a unified protocol with five seeds and logistic regression as the downstream classifier (unless noted in Section 5.2).

4.4 Protocol

All preprocessing is fit on training only. For AdvEncoder, the encoder, decoder, G , and discriminator are used only to construct $\hat{D}_{train}$. Downstream classifiers are trained on the debaised training data and evaluated on the original, untransformed test split. No AdvEncoder component is used at test time. AdvEncoder-EO uses labels during construction of the debaised dataset, but labels are not required at test or deployment time.

All preprocessing baselines feed a logistic regression classifier with `class_weight="balanced"` (logistic regression hyperparameter to give equal weightage to both classes). To validate model-agnosticity, we additionally evaluate AdvEncoder with random forest, XGBoost, SVM, and MLP downstream classifiers (Section 5.2). Reported values are averaged over five random seeds with standard deviations.

5 Results and Analysis

We evaluate both AdvEncoder variants against eleven baselines on four datasets. All logistic regression classifiers use `class_weight="balanced"`; results are averaged over five seeds. Downstream models are trained on the debiased training data and evaluated on untransformed held-out test data. Our goal is not to claim numerical dominance on every metric: some baselines achieve lower values on particular metrics in particular datasets, but these often come with substantial accuracy loss, instability, or degradation of other fairness criteria. We evaluate methods by their overall fairness-utility profile and whether observed behavior matches the intended alignment objective. DAP is marked $\dagger$ as an inprocessing reference, not a like-for-like preprocessing competitor.

5.1 Main Results

Tables 4 and 5 report main results across all four datasets. Several patterns are consistent.

Sampling-based methods are unreliable for fairness. Undersampling, SMOTE, and CTGAN do not consistently reduce group disparities. SMOTE raises SP on LSAC from 0.03 to 0.22 and substantially increases EO; CTGAN worsens SP on German Credit from 0.10 to 0.24. Balancing sample counts does not necessarily remove group-discriminative structure.

Single-metric improvements can harm other criteria. Several baselines achieve strong values on one fairness metric while worsening another. FUR obtains very low SP on Adult and ACSIncome, but its E.Opp and EO are worse than the original baseline on both. This motivates reporting all three fairness metrics rather than the most favorable one.

Representation-learning baselines can be strong but dataset-sensitive. LFR achieves perfect or near-perfect fairness on LSAC and German Credit, but incurs large accuracy drops on Adult and ACSIncome and performs poorly on Adult SP. LAFTR is generally more stable but provides smaller fairness improvements.

AdvEncoder provides a strong fairness-utility trade-off. AdvEncoder does not dominate every baseline on every metric, but consistently provides competitive accuracy together with strong fairness across all three metrics. On Adult, AdvEncoder-EO achieves the best EO among preprocessing methods while keeping accuracy within ~ 1.6 points of the original classifier. On ACSIncome, AdvEncoder-EO achieves the lowest EO among preprocessing methods and retains substantially higher accuracy than LFR. On German Credit, AdvEncoder substantially reduces the large baseline EO gap while avoiding the instability of sampling-based methods. On LSAC, where several methods already achieve near-zero disparities, AdvEncoder remains competitive.

Variant behavior matches the alignment objective. AdvEncoder-SP usually obtains lower SP than AdvEncoder-EO, while AdvEncoder-EO usually obtains lower E.Opp and EO. This pattern is consistent with Section 3.6: unconditional alignment targets marginal group differences, label-conditional alignment targets within-label differences.

Utility cost. AdvEncoder incurs a modest accuracy cost: ~ 1 –2 points on Adult and ACSIncome, negligible on LSAC, and

~ 0.5 points on German Credit under logistic regression. This is substantially smaller than LFR’s loss while still producing large fairness reductions.

5.2 Model-Agnostic Evaluation

A central design claim of AdvEncoder is that its output is a debiased training dataset rather than a constrained model. We test whether downstream fairness improvements persist across heterogeneous classifier families. For each dataset and seed, AdvEncoder is used once to construct a debiased training set; we then train five classifier families (LR, RF, XGBoost, SVM, MLP) on that data and evaluate on the original test set.

Baseline disparity is classifier-dependent; AdvEncoder fairness is not. On German Credit, baseline EO ranges from 10.14 (MLP) to 36.53 (LR)—a $3.6\times$ swing. After AdvEncoder-EO debiasing, EO ranges only from 3.12 to 4.00, a tight sub-percentage-point band. Adult shows the same pattern: AdvEncoder-EO yields EO between 3.60 and 4.45 across all five classifiers. The fairness improvement is therefore a property of the debiased data, not an interaction with a specific downstream model.

The two-variant ordering holds for every classifier. Propositions 1 and 3 predict that AdvEncoder-SP achieves lower SP, and AdvEncoder-EO lower EO. This ordering holds for all 10 dataset-classifier combinations in Table 6. The theoretical implications thus translate into empirically observable behavior independent of the downstream model.

Practical implication. A data-engineering team can produce a debiased training set once and provide it to multiple modeling teams using different classifier families. This differs from inprocessing methods, which require integration into each model class, and from representation methods that require test-time encoding.

6 Ablation Study

We conduct three complementary ablations: a component ablation, a design-choice ablation, and a flipped-anchor ablation that tests whether the *direction* of asymmetric alignment matters. All ablations use AdvEncoder-EO as the reference configuration, since equal opportunity is the primary fairness criterion emphasized throughout the paper.

6.1 Component Ablation

Table 7 compares three settings: (i) *w/o G & Dis*, where the adversarial encoder and discriminator are removed (autoencoder-only); (ii) *w/o Dis*, where G is trained via MSE alignment instead of adversarial feedback; and (iii) the full proposed model.

The autoencoder-only variant provides limited and inconsistent fairness improvement, and on German Credit causes severe utility collapse (accuracy 56.50%) while worsening fairness relative to baseline. MSE alignment achieves low SP on Adult (1.00) but E.Opp rises to 31.00 and EO to 17.00, demonstrating that low SP can be achieved while substantially distorting opportunity-based fairness. Only the full adversarial model is stable across all four datasets, and on LSAC nearly eliminates equal-opportunity disparity entirely (E.Opp = 0.00, EO = 0.14).

Method	UCI Adult				LSAC			
	Acc	SP ↓	E.Opp ↓	EO ↓	Acc	SP ↓	E.Opp ↓	EO ↓
Baseline	84.79±0.00	0.18±0.00	0.08±0.00	0.15±0.00	94.65±0.00	0.03±0.00	0.02±0.00	0.09±0.00
Undersampling	84.71±0.12	0.18±0.01	0.08±0.01	0.16±0.01	94.68±0.09	0.04±0.01	0.02±0.01	0.13±0.02
SMOTE	82.75±0.38	0.09±0.03	0.14±0.03	0.15±0.03	92.21±0.47	0.22±0.04	0.16±0.03	0.57±0.05
CTGAN	84.07±0.52	0.17±0.03	0.07±0.02	0.15±0.03	94.60±0.19	0.03±0.02	0.03±0.02	0.10±0.06
FUR	80.47±0.22	0.01±0.01	0.20±0.02	0.22±0.02	94.79±0.02	0.01±0.01	0.00 ±0.00	0.02±0.01
Zhang	84.50±0.18	0.12±0.02	0.07±0.01	0.10±0.02	94.82±0.02	0.00 ±0.00	0.00 ±0.00	0.01±0.01
LAFTTR	84.54±0.21	0.11±0.02	0.08±0.02	0.11±0.03	94.70±0.08	0.02±0.01	0.01±0.01	0.05±0.02
FAW	84.67±0.05	0.01 ±0.00	0.06±0.01	0.13±0.01	94.20±0.04	0.11±0.01	0.07±0.01	0.36±0.02
LFR	76.77±0.36	0.79±0.35	0.06±0.01	0.08±0.01	94.78±0.03	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00
AdapFair	83.12±0.31	0.05±0.01	0.18±0.02	0.21±0.02	94.77±0.04	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00
DAP [†]	83.73±0.15	0.05±0.01	0.20±0.02	0.21±0.02	94.78±0.04	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00
AdvEnc-SP	83.50±0.10	0.04±0.01	0.04±0.01	0.09±0.01	94.55±0.05	0.00 ±0.00	0.00 ±0.00	0.01±0.01
AdvEnc-EO	83.20±0.10	0.10±0.01	0.03 ±0.01	0.04 ±0.01	94.60±0.05	0.01±0.01	0.00 ±0.00	0.00 ±0.00

Table 4: Results on UCI Adult and LSAC. Accuracy in %. **Bold** indicates best per fairness metric among preprocessing methods. [†] DAP is an inprocessing reference point only.

Method	ACSIIncome				German Credit			
	Acc	SP ↓	E.Opp ↓	EO ↓	Acc	SP ↓	E.Opp ↓	EO ↓
Baseline	76.72±0.08	0.05±0.00	0.03±0.00	0.06±0.00	71.00±0.00	0.10±0.00	0.23±0.00	0.36±0.00
Undersampling	76.72±0.14	0.05±0.01	0.03±0.01	0.06±0.01	70.50±1.12	0.10±0.03	0.10±0.05	0.18±0.08
SMOTE	75.32±0.33	0.05±0.02	0.04±0.02	0.08±0.02	70.50±0.63	0.15±0.04	0.27±0.05	0.43±0.09
CTGAN	75.54±0.41	0.05±0.02	0.04±0.02	0.08±0.02	69.83±1.52	0.24±0.04	0.19±0.06	0.44±0.06
FUR	75.04±0.15	0.01 ±0.01	0.07±0.02	0.14±0.03	69.50±0.71	0.02±0.02	0.03±0.05	0.03±0.05
Zhang	75.13±0.45	0.05±0.02	0.02±0.01	0.04±0.02	70.00±1.00	0.13±0.05	0.11±0.06	0.28±0.10
LAFTTR	75.40±0.20	0.04±0.01	0.03±0.01	0.05±0.02	70.50±1.67	0.07±0.04	0.07±0.06	0.14±0.13
FAW	73.50±0.10	0.03±0.01	0.03±0.01	0.06±0.01	72.00±0.04	0.05±0.01	0.13±0.01	0.35±0.01
LFR	68.80±0.42	0.04±0.01	0.03±0.01	0.05±0.01	70.00±0.20	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00
AdapFair	76.76±0.14	0.04±0.00	0.04±0.01	0.08±0.01	73.50±0.88	0.03±0.01	0.11±0.04	0.22±0.04
DAP [†]	76.34±0.22	0.03±0.01	0.06±0.02	0.06±0.02	74.17±0.85	0.03±0.01	0.04±0.04	0.13±0.06
AdvEnc-SP	75.95±0.10	0.02±0.01	0.03±0.01	0.05±0.01	70.50±0.32	0.01±0.01	0.03±0.01	0.05±0.02
AdvEnc-EO	75.80±0.10	0.03±0.01	0.02 ±0.01	0.03 ±0.01	70.50±0.32	0.02±0.01	0.02±0.01	0.03±0.02

Table 5: Results on ACSIncome (sex) and German Credit (age). Conventions as in Table 4.

6.2 Design-Choice Ablation

Table 8 evaluates two design choices. *CLF loss only* removes the reconstruction objective, retaining only the task classifier. *Symmetric AE* trains the autoencoder on both groups, removing the asymmetric anchoring while preserving the adversarial alignment.

Removing reconstruction substantially weakens the method—most dramatically on LSAC, where accuracy collapses to 78.80%. The symmetric variant is consistently weaker on most datasets: on German Credit, symmetric training worsens SP from 2.30 to 19.00 and E.Opp from 2.10 to 14.00. On Adult, symmetric training worsens SP from 10.00 to 15.00 and EO from 4.45 to 7.00, while on LSAC it raises SP from 1.41 to 7.00 and EO from 0.14 to 12.00. These results provide direct empirical support for both the reconstruction objective and the asymmetric anchoring principle: removing either substantially degrades the method’s effectiveness.

6.3 Flipped Anchor Ablation: Does the Direction Matter?

The asymmetric alignment principle (Section 3.2) prescribes a specific direction: train the autoencoder on the unprivileged group and align the privileged group toward it. A natural question is whether the *direction* matters, or whether any one-sided alignment would yield comparable fairness gains. **Setup.** We construct a flipped variant, **AdvEnc-EO (flipped)**, that reverses the anchor: the autoencoder is trained exclusively on the *privileged* group, and G maps *unprivileged* data into this privileged-anchored latent space: $p(G(\mathcal{D}_{s=0})) \approx p(Enc(\mathcal{D}_{s=1}))$. All other components—architecture, training schedule, discriminator conditioning, latent dimension, and downstream classifier—are identical to AdvEncoder-EO. Conceptually, the proposed direction is a “bring-down” strategy that strips privilege-correlated features out of the privileged group’s data, while the flipped direction is a “bring-up” strategy that synthetically endows the unprivileged group with privileged-like features while leaving the privileged data

Clf.	Setting	Adult			German		
		ACC	SP↓	EO↓	ACC	SP↓	EO↓
LR	Baseline	84.79	17.74	15.35	71.00	9.87	36.53
	AdvEnc-SP	83.50	4.00	9.00	70.50	1.15	5.20
	AdvEnc-EO	83.20	10.00	4.45	70.50	2.30	3.12
RF	Baseline	84.72	18.12	13.30	73.50	16.71	21.51
	AdvEnc-SP	82.80	4.30	8.80	69.00	1.20	6.00
	AdvEnc-EO	81.90	9.10	4.20	68.50	2.40	3.80
XGB	Baseline	86.62	17.48	12.06	71.50	5.46	23.31
	AdvEnc-SP	83.60	3.70	8.00	68.50	1.40	6.50
	AdvEnc-EO	82.40	8.20	3.60	68.00	2.50	4.00
SVM	Baseline	85.48	16.66	14.47	75.50	10.65	19.12
	AdvEnc-SP	83.80	4.50	8.50	71.00	1.60	5.80
	AdvEnc-EO	83.50	9.50	4.00	70.50	2.80	3.50
MLP	Baseline	85.21	16.64	12.46	71.50	0.43	10.14
	AdvEnc-SP	83.70	4.10	7.90	69.00	0.80	5.20
	AdvEnc-EO	83.30	8.80	3.80	68.50	1.80	3.20

Table 6: Compatibility across classifier families on Adult and German Credit. Debiasing transformation held fixed; only downstream classifier varies. ACC in %; SP and EO in percentage points. **Bold** marks lowest EO per classifier block.

untransformed.

Results. Table 9 reports the comparison on Adult and German Credit. The proposed direction substantially outperforms the flipped variant on both datasets and both fairness metrics. On Adult, AdvEncoder-EO achieves EO = 4.45 vs. 10.00 for the flipped variant (more than $2\times$ larger fairness gap) and SPD = 10.00 vs. 15.00. On German Credit, the gap is even more pronounced: SPD = 2.30 vs. 6.00 ($2.6\times$) and EO = 3.20 vs. 11.00 ($3.4\times$).

Interpretation. Three observations follow. *First*, the direction of asymmetric alignment is not interchangeable: if any one-sided alignment sufficed, the flipped variant would have produced comparable fairness, but it leaves substantial residual disparity on both datasets and both metrics. This rules out the alternative hypothesis that AdvEncoder’s gains derive from one-sided distribution-matching in general rather than from anchoring on the unprivileged group specifically. *Second*, the flipped variant’s slightly higher accuracy on both datasets (Adult: 84.30 vs. 83.20; German: 70.65 vs. 70.50) is consistent with the conceptual analysis: because it preserves the privileged group’s original features unaltered, the downstream classifier retains access to the full discriminative signal—including signal correlated with group membership. The proposed direction sacrifices a small amount of accuracy precisely because it removes this signal. *Third*, the flipped variant’s residual disparity reflects a heterogeneous training distribution: original privileged features alongside synthetically uplifted unprivileged features. This heterogeneity reintroduces the group-discriminative signal that the alignment was meant to remove.

The flipped-anchor ablation thus provides direct empirical validation of the asymmetric alignment principle’s specific directional claim: alignment must move the privileged group

Dataset	Condition	ACC	SP	E.Opp	EO
Adult	Baseline	84.79	17.74	8.00	15.35
	w/o G & Dis	82.27	15.00	3.00	8.00
	w/o Dis	80.98	1.00	31.00	17.00
	Proposed	83.20	10.00	3.20	4.45
German	Baseline	71.00	9.87	23.00	36.53
	w/o G & Dis	56.50	21.00	9.00	27.00
	w/o Dis	69.00	21.00	15.00	26.00
	Proposed	70.50	2.30	2.10	3.20
ACIncome	Baseline	76.72	4.55	3.20	5.91
	w/o G & Dis	76.16	5.00	9.00	6.00
	w/o Dis	75.65	2.00	7.00	5.00
	Proposed	75.80	3.23	2.34	3.49
LSAC	Baseline	94.65	2.79	2.00	8.97
	w/o G & Dis	94.60	5.00	3.00	8.00
	w/o Dis	94.69	3.00	1.00	6.00
	Proposed	94.60	1.41	0.00	0.14

Table 7: Component ablation. Metrics in percentage points.

Dataset	Condition	ACC	SP	E.Opp	EO
Adult	Proposed	83.20	10.00	3.20	4.45
	CLF loss only	78.76	4.00	35.00	19.00
	Symmetric AE	82.49	15.00	3.00	7.00
German	Proposed	70.50	2.30	2.10	3.20
	CLF loss only	70.00	9.00	16.00	11.00
	Symmetric AE	67.50	19.00	14.00	16.00
ACIncome	Proposed	75.80	3.23	2.34	3.49
	CLF loss only	75.67	3.00	7.00	5.00
	Symmetric AE	73.19	5.00	2.00	3.00
LSAC	Proposed	94.60	1.41	0.00	0.14
	CLF loss only	78.80	3.00	4.00	10.00
	Symmetric AE	94.69	7.00	4.00	12.00

Table 8: Design-choice ablation. Metrics in percentage points.

toward the unprivileged anchor, not the reverse.

7 Broader Societal Impact

The datasets used in this work reflect high-stakes domains where algorithmic bias has documented real-world consequences. We discuss three implications of deploying AdvEncoder.

Beneficiaries and deployment. The primary beneficiaries are underrepresented groups whose data currently receives less reliable statistical support. Practitioners benefit from the model-agnostic output: a debiased dataset can be distributed to multiple downstream teams without requiring them to modify training pipelines. Regulators and auditors may benefit from the separability of the debiasing step. AdvEncoder is a training-time method that does not transform deployment inputs and therefore does not require labels, sensitive attributes, or learned encoders at inference. Practitioners should verify the sensitive-attribute split using domain knowledge, evaluate models on untransformed data drawn from the expected deployment distribution, and conduct ongoing monitoring for

Dataset	Setting	ACC	SP ↓	EO ↓
Adult	Baseline	84.79±0.00	17.74±0.00	15.35±0.00
	AdvEnc-EO	83.20±0.10	10.00 ±0.95	4.45 ±0.81
	AdvEnc-EO (flip)	84.30±0.17	15.00±2.10	10.00±1.50
German	Baseline	71.00±0.00	9.87±0.00	36.53±0.00
	AdvEnc-EO	70.50±0.32	2.30 ±0.84	3.20 ±2.42
	AdvEnc-EO (flip)	70.65±0.54	6.00±1.80	11.00±2.50

Table 9: Flipped anchor ablation. Metrics in percentage points; **bold** marks lowest value per metric per dataset.

distribution shift. AdvEncoder currently operates on a single binary sensitive attribute and does not address intersectional identities, label measurement bias, or feedback loops.

Choosing between AdvEncoder-SP and AdvEncoder-EO. The two-variant design surfaces a normative choice. AdvEncoder-SP targets statistical parity and is appropriate under disparate-impact regulations or when equality of outcome is the policy goal. AdvEncoder-EO targets equal opportunity and equalized odds and is appropriate when qualified members of each group should have equal access (e.g., loans among genuinely creditworthy applicants). Propositions 1 and 3 make explicit that these goals are not jointly achievable under unequal base rates (Chouldechova 2017); the choice is therefore a substantive ethical decision rather than a technical one.

Normative status of the reference and risk of fairness washing. We emphasize that AdvEncoder uses the unprivileged distribution as a *conservative statistical anchor*, not a normative target for outcomes. In Adult, for example, the female income distribution reflects documented historical wage discrimination (Barocas, Hardt, and Narayanan 2023); anchoring on this distribution does not endorse the observed outcomes as the desired target. The debiasing goal is to reduce *separability* of group membership in the feature space, not to validate observed disparities. Relatedly, a practitioner could apply AdvEncoder, observe improved benchmark metrics, and use this to deflect regulatory scrutiny while the underlying system remains harmful in deployment. Improved benchmark fairness is necessary but not sufficient for fair deployment; the debiased dataset should be one component of a broader fairness audit, complemented by third-party review, stakeholder consultation, and ongoing deployment monitoring.

8 Limitations

Training-time scope and alignment-level guarantees. AdvEncoder debiases the training dataset, not deployment inputs. The formal implications in Section 3.6 apply to the aligned training distribution and characterize what follows under exact or approximate alignment; they do not prove that the adversarial optimization converges to such alignment, nor do they extend automatically to every deployment distribution. Empirical evaluation on held-out, untransformed test data is therefore essential, and deployment should monitor for distribution shift.

Adversarial training instability. Phase 2 relies on a GAN-

style min-max objective that can be sensitive to hyperparameters and seeds. We average over multiple seeds, but practical deployment should monitor discriminator loss, alignment quality, and downstream fairness. A stable non-adversarial variant retaining asymmetric alignment is a promising direction.

Single binary attribute and discriminator design.

AdvEncoder-EO conditions the discriminator on training labels, assuming they are meaningful proxies for the construct of interest; if labels encode historical or measurement bias, equalizing conditional distributions may not remove underlying injustice. AdvEncoder also assumes a single binary sensitive attribute. Extending it to multi-class or intersectional settings—via a multi-class discriminator over group cells or per-attribute discriminators—is natural in principle but raises two coupled difficulties. First, anchor selection becomes harder: with many small intersectional cells, it is unclear which cell should serve as the reference, and the smallest cells that most need protection are the least reliable anchors. Second, our single discriminator over the binary split, while simple, shares the failure modes noted above: it can induce mode collapse in G or leave residual group signal in the latent if it fails to fully penalize separability. A stable multi-attribute alignment objective with a principled anchor-selection rule is the most important direction for future work.

Fairness–utility trade-off and reconstruction validity. AdvEncoder reduces fairness gaps at a modest but nonzero accuracy cost, and some baselines achieve lower values on individual metrics in particular datasets. Its main advantage is the consistent fairness–utility trade-off rather than strict per-metric dominance. Decoded categorical blocks are projected to valid one-hot encodings, ensuring syntactic but not full semantic validity of cross-feature relationships.

9 Conclusion

We presented AdvEncoder, a training-time data debiasing framework based on *asymmetric distribution alignment*: rather than moving all groups symmetrically toward a shared midpoint, AdvEncoder learns a latent reference space from the unprivileged group and adversarially aligns the privileged group toward it. The motivation is conservative—the group with fewer samples should bear less transformation burden. AdvEncoder-SP and AdvEncoder-EO target marginal and class-conditional alignment respectively, with formal implications under exact and approximate alignment. Across four benchmarks and five classifier families, AdvEncoder achieves a strong fairness–utility trade-off: not strict dominance on every metric, but consistent near-best fairness with competitive accuracy. Ablations confirm that both the adversarial component and the direction of anchoring are important. Future work should extend asymmetric alignment to intersectional and multi-class sensitive attributes, incorporate stronger semantic constraints for reconstructed tabular data, and develop more stable non-adversarial alignment objectives.

References

- Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.
- Becker, B.; and Kohavi, R. 1996. Adult. <https://doi.org/10.24432/C5XW20>.
- Chai, J.; and Wang, X. 2022. Fairness with Adaptive Weights. In *International Conference on Machine Learning (ICML)*, 2853–2866.
- Chawla, N. V.; Bower, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16: 321–357.
- Chouldechova, A. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2): 153–163.
- Ding, F.; Hardt, M.; Miller, J.; and Schmidt, L. 2021. Retiring Adult: New Datasets for Fair Machine Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 6478–6490.
- Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through Awareness. In *Innovations in Theoretical Computer Science Conference*, 214–226.
- Edwards, H.; and Storkey, A. 2016. Censoring Representations with an Adversary. In *International Conference on Learning Representations (ICLR)*.
- Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and Removing Disparate Impact. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268.
- Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27.
- Gordaliza, P.; Del Barrio, E.; Fabrice, G.; and Loubes, J.-M. 2019. Obtaining Fairness using Optimal Transport Theory. In *International Conference on Machine Learning (ICML)*, 2357–2365.
- Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29.
- Hofmann, H. 1994. Statlog (German Credit Data). <https://doi.org/10.24432/C5NC77>.
- Huang, Y.; et al. 2024. AdapFair: Ensuring Adaptive Fairness for Machine Learning Operations. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Islam, R.; Chen, H.; and Cai, Y. 2024. Fairness without Demographics through Shared Latent Space-Based Debiasing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38.
- Kairouz, P.; Liao, J.; Huang, C.; Vyas, M.; Welfert, M.; and Sankar, L. 2022. Generating Fair Universal Representations Using Adversarial Models. *IEEE Transactions on Information Forensics and Security*, 17: 1970–1985.
- Kamiran, F.; and Calders, T. 2009. Classifying without Discriminating. In *International Conference on Computer, Control and Communication*, 1–6.
- Kamiran, F.; and Calders, T. 2012. Data Preprocessing Techniques for Classification without Discrimination. *Knowledge and Information Systems*, 33(1): 1–33.
- Louizos, C.; Swersky, K.; Li, Y.; Welling, M.; and Zemel, R. 2016. The Variational Fair Autoencoder. In *International Conference on Learning Representations (ICLR)*.
- Luong, B. T.; Ruggieri, S.; and Turini, F. 2011. k -NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 502–510.
- Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning Adversarially Fair and Transferable Representations. In *International Conference on Machine Learning (ICML)*, 3384–3393.
- Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6): 1–35.
- Oh, C.; Won, H.; So, J.; Kim, T.; Kim, Y.; Choi, H.; and Song, K. 2022. Learning Fair Representation via Distributional Contrastive Disentanglement. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1295–1305.
- Sahyouni, B.; Vowels, M.; Chen, L.; and Hadfield, S. 2025. Differential Adjusted Parity for Learning Fair Representations. *arXiv preprint arXiv:2502.09765*.
- Wang, H.; Ustun, B.; and Calmon, F. 2019. Repairing without Retraining: Avoiding Disparate Impact with Counterfactual Distributions. In *International Conference on Machine Learning (ICML)*, 6618–6627.
- Wang, T.; Zhao, J.; Yatskar, M.; Chang, K.-W.; and Ordonez, V. 2019. Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. In *IEEE International Conference on Computer Vision (ICCV)*, 5310–5319.
- Wightman, L. F. 1998. LSAC National Longitudinal Bar Passage Study. Technical report, Law School Admission Council.
- Xu, D.; Yuan, S.; Zhang, L.; and Wu, X. 2018. FairGAN: Fairness-aware Generative Adversarial Networks. In *IEEE International Conference on Big Data*, 570–575.
- Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; and Veeramachaneni, K. 2019. Modeling Tabular Data Using Conditional GAN. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.
- Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning Fair Representations. In *International Conference on Machine Learning (ICML)*, 325–333.
- Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 335–340.