INVARIFAIR: ONE REGULARIZATION PARAMETER JOINTLY BOUNDS OOD RISK AND EQUALIZED ODDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-distribution (OOD) robustness and algorithmic fairness are typically treated as separate problems that need separate machinery: invariant learning methods deal with spurious correlations between inputs and labels, while fairness methods deal with spurious correlations between protected attributes and outcomes. We observe that under a latent structural causal model (SCM), these are the same problem viewed from two angles; both happen when a predictor leans on a latent factor that is not causally related to the label. We introduce a causal-disentanglement framework for OOD generalization that provably bounds worst-case risk under a bounded shift in the spurious latent's distribution, and show that allowing the protected attribute to depend on both the causal and spurious latents instead of the spurious latent alone leaves a purely causal classifier generally with a nonzero residual demographic-parity (DP) gap, and prove that this residual gap is provably never smaller than the corresponding equalized-odds (EO) gap, with the ratio between them controlled by a single interpretable quantity: how much label noise stays once the causal factors are known. We also show a lightweight fairness-aware disentanglement objective that explicitly targets the spurious component of the protected attribute by routing it into the spurious latent space during training. Like all methods compared here, ours requires no protected attribute access at inference time. On two real-world fairness benchmarks, evaluated under an equal training budget, with independently-reseeded runs, and with 10 seeds per method to support paired significance testing, our method reduces the equalized-odds gap by 30-34% relative to standard Empirical Risk Minimization (ERM) and Adversarial Debiasing on Adult Income ($p < 0.005$ for both). On COMPAS, we do not find a statistically robust improvement: our method's equalized-odds and demographic-parity gaps are not significantly different from ERM's or Adversarial Debiasing's, and are directionally (but not significantly) worse than Reweighting's. This is a weaker result than an earlier and under-powered 3-seed run of the same experiment showed, and explains why we report 10-seed statistics rather than 3. An ablation study and a real-data sweep over the invariance parameter show a dataset-dependent pattern: on Adult, the fairness-aware heads account for nearly the full improvement; on COMPAS, the heads and the invariance penalty are complementary within our method's own variants, and increasing the invariance parameter creates a further fairness-gap reduction consistent with our theoretical bound, even though this does not go into an advantage over the external baselines at $n = 10$. Consistent with our theory, our causal mechanism improves equalized odds substantially on Adult while leaving its demographic-parity gap statistically unchanged from ERM, and we connect this same asymmetry to a corresponding in-distribution accuracy cost on Adult that the mechanism's adversarial component predicts but does not fully explain on its own.

1 INTRODUCTION

Two long-standing lines of work in trustworthy machine learning have developed largely independently. Invariant learning methods (Invariant Risk Minimization, IRM; Group Distributionally Robust Optimization, Group DRO) prevent classifiers from relying on spurious and environment-dependent correlations that break under distribution shift. Algorithmic fairness methods (reweight-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ing, adversarial debiasing, fair representation learning) prevent classifiers from relying on protected attributes like race or gender. Both study the same mathematical object, but rarely cite each other or share theoretical machinery.

We argue this separation is unnecessary. Consider a latent SCM in which label Y is generated only from a causal latent factor Z_c, while a spurious latent Z_s is correlated with Y in the training environment but not causally related to it. If a protected attribute A (ex. gender, race) happens to be a function of Z_s instead of Z_c, exactly the situation fairness researchers describe when a sensitive attribute "leaks" through proxy variables, then any method driving a classifier's reliance toward Z_c and away from Z_s simultaneously improves OOD robustness and fairness. The two problems are the same problem.

We introduce a causal-disentanglement framework learning a causal/spurious latent decomposition via a CausalVAE and enforcing prediction invariance under counterfactual resampling of the spurious latent (Section 3), together with a worst-case OOD-risk bound for this framework under a bounded spurious-shift assumption (Theorem 1). Building on this foundation, we contribute:

1. **A worst-case OOD-risk bound** for the causal-disentanglement framework (Theorem 1): under a bounded spurious-shift assumption, worst-case risk is $\sigma_y^2 + C(1-\lambda_{\text{inv}})^2$, converging to Bayes-optimal risk as $\lambda_{\text{inv}} \to 1$.
2. **A new theoretical corollary** showing that the same invariance weight λ_{inv} that bounds OOD risk also bounds the equalized-odds and demographic-parity gaps, under a generalized attribute-alignment assumption letting the protected attribute depend on both the causal and spurious latents - weaker and more realistic than assuming a pure function of the spurious latent alone, since the latter forces the protected attribute to be marginally independent of the label (Lemma 5), which fails on essentially every real fairness benchmark.
3. **A provable asymmetry between the two fairness criteria** (Proposition 4): the residual equalized-odds gap a purely causal classifier keeps is never larger than its residual demographic-parity gap, with the ratio equal to exactly $\sigma_y^2/\mathrm{Var}(Y \mid A)$, the fraction of label variance unexplained by the causal factors. This replaces an earlier and unquantified appeal to "conditioning on Y" with a single falsifiable number, verified on a synthetic SCM and, qualitatively, against our real-data results.
4. **A fairness-aware disentanglement objective**, adapted from Creager et al. (2019), supervising the spurious latent to capture protected-attribute information directly. The contribution is integrating this with our framework's OOD-invariance training; we also show its adversarial component is, by the same theory, in tension with genuine (non-spurious) attribute correlation, giving a theory-motivated account of an otherwise-unexplained accuracy cost.
5. Empirical validation on two real-world tabular fairness benchmarks under both in-distribution and OOD evaluation.

2 Related Work

Invariant learning and OOD generalization. IRM (Arjovsky et al., 2019) and Group DRO (Sagawa et al., 2020) require explicit environment or group annotations; our framework removes this via learned counterfactual interventions on a disentangled latent space (Section 3), with a worst-case OOD-risk bound of its own (Theorem 1). Our main theoretical contribution targets a new quantity building on this mechanism (fairness gap, Section 4) instead of proposing a new robustness mechanism.

Algorithmic fairness. Reweighting/resampling (Kamiran & Calders, 2012) correct base-rate imbalance directly. Adversarial debiasing (Zhang et al., 2018; Madras et al., 2018) uses a gradient-reversal adversary to scrub protected-attribute information from a representation. Closer to our method, Creager et al. (2019) disentangle a representation into attribute-capturing and attribute-free parts via mutual-information supervision, discarding the former at test time, the same recipe as our fairness-aware heads (Section 5). Our contribution is applying it to a latent also constructed to satisfy an OOD invariance property, and deriving a quantitative bound connecting the resulting fairness gap to that same invariance parameter; none of Creager et al. (2019); Madras et al. (2018); Zhang et al. (2018) connect their objective to a distribution-shift guarantee. Equalized odds was formalized

by Hardt et al. (2016), who showed it is generally incompatible with demographic parity except in degenerate cases, a tension Section 4.3 links directly to the causal structure of the problem.

Causal approaches to fairness. Kilbertus et al. (2017) and Nabi & Shpitser (2018) formalize fairness via causal graphs but do not connect it to OOD robustness or a worst-case risk bound. Closest to our theory is Makar & D'Amour (2023), who show a qualitative near-equivalence between separation (equalized odds) and risk invariance in anti-causal settings (input generated from label and attribute). We differ by (i) a causal SCM where Y is caused by Z_c and A depends on both Z_c and Z_s (A4′), and (ii) a quantitative bound on the equalized-odds gap in terms of a single trainable parameter λ_{inv}. We do not include Makar & D'Amour (2023) or Kilbertus et al. (2017) as empirical baselines: the former's guarantees are stated for an anti-causal data-generating process (DGP) incompatible with our causal SCM, so training their method under our (causal) data-generating assumptions would not test their theory on its own terms; the latter two are causal-graph formalizations without a public reference implementation for tabular classification comparable to our other baselines.

3 Framework: A Causal-Spurious Latent Decomposition

Our framework learns an encoder that decomposes each input x into a causal latent Z_c and spurious latent Z_s via a CausalVAE, with a classifier head working only on Z_c. A conditional generative model $G(Z_s \mid Z_c)$ synthesizes counterfactual spurious contexts at training time, and an invariance loss penalizes any change in classifier output when Z_s is resampled while Z_c is held fixed:

$$\mathcal{L}_{\text{inv}} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{KL}\Big(C(E(x)_c) \,\Big\|\, C(E(x'_k)_c)\Big), \quad x'_k = D(Z_c, Z'_{s,k}),\ Z'_{s,k} \sim G(Z_c). \tag{1}$$

We adopt the following assumptions on the data-generating process, consistent with the standard linear-Gaussian invariant-learning setting (Arjovsky et al., 2019):

A1 (Causal invariance). *$Y = w_c^\top Z_c + \varepsilon_y$, $\varepsilon_y \sim \mathcal{N}(0, \sigma_y^2)$, with (w_c, σ_y^2) invariant across environments $e \in \mathcal{E}$.*

A2 (Bounded spurious shift). *Only $Z_s \sim \mathcal{N}(0, \Sigma_s^{(e)})$ varies across e; $Z_c \sim \mathcal{N}(0, \Sigma_c)$, noise ε, and (A_c, A_s) in $X = A_c Z_c + A_s Z_s + \varepsilon$ are invariant, and $\sup_{e\in\mathcal{E}} \lambda_{\max}(\Sigma_s^{(e)}) < \infty$. Within every environment, $Z_c, Z_s, \varepsilon, \varepsilon_y$ are mutually independent (standard SCM exogeneity).*

A3 (Generative identifiability). *E recovers the true latents up to invertible linear maps, $E(x)_c = M_c Z_c$, $E(x)_s = M_s Z_s$, justifying the SCM analysis below for the trained encoder.*

Theorem 1 (Worst-case OOD-risk bound)**.** *Under A1-A2, let $f_\alpha(X) = (1-\alpha) f_{\text{ERM}}(X) + \alpha f_{\text{causal}}(Z_c)$, $\alpha \in [0,1]$, $f_{\text{ERM}}(X) = w_{\text{spur}}^\top X$ (fit under a fixed reference environment), $f_{\text{causal}}(Z_c) = w_c^\top Z_c$. Then, for every α,*

$$\sup_{e\in\mathcal{E}} \mathbb{E}_e\big[(Y - f_\alpha(X))^2\big] \;=\; \sigma_y^2 + C\,(1-\alpha)^2, \tag{2}$$

where $C := d^\top \Sigma_c d + w_{\text{spur}}^\top \Sigma_\varepsilon w_{\text{spur}} + \sup_{e\in\mathcal{E}} (A_s^\top w_{\text{spur}})^\top \Sigma_s^{(e)} (A_s^\top w_{\text{spur}})$, $d := w_c - A_c^\top w_{\text{spur}}$. Worst-case risk thus converges to the Bayes-optimal risk σ_y^2 as $\alpha \to 1$, regardless of the size of the spurious shift $\mathcal{E}$.

Proof sketch. Mutual independence of $Z_c, Z_s, \varepsilon, \varepsilon_y$ under A1-A2 makes $\mathbb{E}_e[(Y - f_\alpha(X))^2]$ a sum of variances, one term per source; only the Z_s term depends on e, giving the stated C under $\sup_e$. Full derivation and numerical verification (script `theory_ood.py`) in Appendix A.3. □

Remark 2 (Scope of the theory: an idealized family)**.** *Theorem 1, Corollary 3, and Proposition 4 are statements about an idealized parametric family $\{f_\alpha = (1-\alpha) f_{\text{ERM}} + \alpha f_{\text{causal}} : \alpha \in [0,1]\}$, where f_{ERM} is linear in raw X and f_{causal} linear in Z_c alone. This is not the model actually deployed: the real classifier is architecturally $C(Z_c)$ (never sees raw X), and λ_{inv} in the training objective is a loss weight on a soft penalty inside one jointly-trained nonlinear network, not a literal*

interpolation coefficient between two separately fitted models. We use α for this idealized family's interpolation coefficient and reserve λ_{inv} for the actual training loss weight specifically so the two are never confused with one another; no result here proves the trained encoder at a given λ_{inv} behaves like f_α for a correspondingly large α - only that increasing λ_{inv} is intended to reduce spurious information surviving in learned Z_c, a working hypothesis, not an established theorem. Closing this gap would require analyzing the full joint objective, which we do not attempt. Our actual evidence for the intended correspondence is empirical: the real-data λ_{inv} sweep (Section 7.2) checks whether increasing the loss weight moves the measured fairness gap as Corollary 3 predicts - it does on COMPAS but not on Adult once the fairness heads are active, itself evidence the correspondence is not uniform in practice. Read the bounds below as characterizing what a classifier with a given degree of causal reliance would achieve, not a guarantee that a larger λ_{inv} achieves it. A second, unrelated idealization is joint Gaussianity itself: it implies $\text{EO}(f)$ is exactly constant across y (Appendix A.5), which is why we report one scalar EO number per method; Adult and COMPAS are manifestly not jointly Gaussian, and we have not checked how closely the per-y gaps actually agree on either dataset.

4 Theoretical Connection to Fairness

4.1 Additional Assumption

We adopt Assumptions A1-A3 from Section 3 (causal invariance, bounded spurious shift, generative identifiability) and add:

Assumption A4′ (Generalized attribute alignment) *The protected attribute A is generated as a linear function of both latents,*

$$A = b_c^\top Z_c + b_s^\top Z_s + \varepsilon_A, \tag{3}$$

for fixed vectors b_c, b_s and independent noise ε_A.

An earlier version of this assumption (A4) restricted A to depend on Z_s alone, i.e. $b_c = 0$, which forces A marginally independent of Y (Lemma 5, Appendix A) - unrealistic, since on datasets like Adult, income and sex plausibly correlate for real, structural reasons and not only via spurious proxies. A4′ keeps the standard proxy-correlation concern (b_s) while also allowing a genuine, non-spurious correlation between A and Y through Z_c (b_c) - weaker, and, as shown below, what lets the theory predict instead of just describe after the fact the residual demographic-parity gap a purely causal classifier retains.

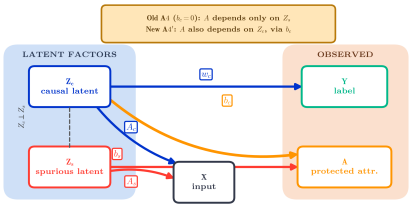

Figure 1: The SCM behind A4′: old A4 had only the b_s path into A; A4′ adds b_c (amber), letting A depend on the causal latent too.

4.2 Corollary: Fairness Gap Bound

Corollary 3 (Fairness Gap Bound)**.** *Under Assumptions (A1)-(A3) and (A4′), let f_α be the minimizer of the training objective (Eq. 1) with invariance weight $\alpha \in [0, 1]$. Then the demographic parity gap satisfies*

$$\text{DP}(f_\alpha) \;=\; \big|\,\mathbb{E}[f_\alpha(X) \mid A = 1] - \mathbb{E}[f_\alpha(X) \mid A = 0]\,\big| \;=\; \big|\,\alpha \cdot \text{DP}_{\text{floor}} + (1 - \alpha) \cdot C_{\text{dp}}\,\big|, \tag{4}$$

where C_{dp} is the α=0 (ERM-level) demographic parity gap and

$$\text{DP}_{\text{floor}} \;:=\; \frac{|\text{Cov}(Y, A)|}{\text{Var}(A)} \;=\; \frac{|w_c^\top \Sigma_c b_c|}{\text{Var}(A)}. \tag{5}$$

$\mathrm{DP}_{\mathrm{floor}}$ *and* C_{dp} *are gap magnitudes (each itself defined with an absolute value), equal to* $\mathrm{DP}(f_1)$ *and* $\mathrm{DP}(f_0)$ *respectively. What the proof (Appendix A) actually derives, for every* $\alpha \in [0,1]$*, is a single absolute value around a signed linear combination,* $\mathrm{DP}(f_\alpha) = |\alpha P_1 + (1-\alpha)P_0|$*, where* P_1, P_0 *are signed coefficients with* $|P_1| = \mathrm{DP}_{\mathrm{floor}}$*,* $|P_0| = C_{\mathrm{dp}}$*. This coincides with the displayed* $|\alpha \cdot \mathrm{DP}_{\mathrm{floor}} + (1-\alpha) \cdot C_{\mathrm{dp}}|$ *whenever* P_1, P_0 *agree in sign; when they do not,* $\mathrm{DP}(f_\alpha)$ *for intermediate* α *can fall below* $\alpha \cdot \mathrm{DP}_{\mathrm{floor}} + (1-\alpha) \cdot C_{\mathrm{dp}}$*, dipping toward zero and back up - the same phenomenon produces the non-monotonic* EO *trajectory we report in Figure 2 and discuss in Section 4.4. Both readings agree exactly at* $\alpha = 0$ *and* $\alpha = 1$ *regardless of sign, which is all that Section 4.3's asymptotic claims rely on. Unlike under the earlier (A4) assumption,* $\mathrm{DP}_{\mathrm{floor}}$ *is generally nonzero: as* $\alpha \to 1$*,* $f_\alpha \to f_{\mathrm{causal}}$*, but* f_{causal} *depends on* Z_c*, and under A4′,* Z_c *is no longer independent of* A *whenever* $b_c \neq 0$*. The invariance mechanism removes exactly the* Z_s*-mediated (spurious) component of the fairness gap; it does not and should not remove the* Z_c*-mediated (genuine) component.*

Proof sketch. Parameterize $f_\alpha = (1-\alpha) f_{\mathrm{ERM}} + \alpha f_{\mathrm{causal}}$ as in Theorem 1 and take conditional expectations given A using the joint-Gaussian conditional-mean identity $\mathbb{E}[Z \mid A = a] - \mathbb{E}[Z \mid A = 0] = \mathrm{Cov}(Z, A)/\mathrm{Var}(A)$. Full proof in Appendix A. □

An entirely analogous argument (Appendix A, Section A.5) gives the equalized-odds gap, which under joint Gaussianity is constant in y and equals

$$\mathrm{EO}(f_\alpha) \;=\; \big|\, \alpha \cdot \mathrm{EO}_{\mathrm{floor}} + (1-\alpha) \cdot C_{\mathrm{eo}} \,\big|, \qquad \mathrm{EO}_{\mathrm{floor}} \;:=\; \frac{|\mathrm{Cov}(Y, A)|\, \sigma_y^2}{\mathrm{Var}(Y)\mathrm{Var}(A) - \mathrm{Cov}(Y, A)^2}. \tag{6}$$

4.3 A Provable Asymmetry Between Equalized Odds and Demographic Parity

The floors above look superficially similar, but they are not the same quantity, and the relationship between them is exactly the structural fact that Section 7's ablation and α-sweep results are looking for a theoretical account of.

Proposition 4 (EO floor never exceeds DP floor)**.** *Under Assumptions (A1)–(A3) and (A4′),*

$$\frac{\mathrm{EO}_{\mathrm{floor}}}{\mathrm{DP}_{\mathrm{floor}}} \;=\; \frac{\sigma_y^2}{\mathrm{Var}(Y \mid A)} \;\leq\; 1, \tag{7}$$

with equality only in the degenerate case $\mathrm{Var}(Y \mid A) = \sigma_y^2$*. Consequently* $\mathrm{EO}_{\mathrm{floor}} \leq \mathrm{DP}_{\mathrm{floor}}$ *always, and as the label noise* $\sigma_y^2 \to 0$ *(ex. as* Y *becomes a near-deterministic function of the causal latent* Z_c*),* $\mathrm{EO}_{\mathrm{floor}} \to 0$ *even while* $\mathrm{DP}_{\mathrm{floor}}$ *remains bounded away from zero whenever* $\mathrm{Cov}(Y, A) \neq 0$*.*

Intuitively: once Y is (near-)deterministic given Z_c, conditioning on $Y = y$ already pins down Z_c almost exactly, leaving no residual freedom for A to change the conditional mean further - so the causal predictor's error pattern becomes (near-)independent of A even though its raw prediction rate is not. Demographic parity has no such advantage, since it never conditions on Y. This gives a precise, falsifiable account of why equalized odds should improve more than demographic parity: the relevant quantity is $\sigma_y^2/\mathrm{Var}(Y \mid A)$, a property of how noisy the label is given the causal factors, not an unquantified appeal to "conditioning on Y." We verify Proposition 4 numerically below and revisit it when interpreting the real-data results in Section 7.

4.4 Numerical Verification

We verify Theorem 1's OOD-risk bound on a synthetic linear-Gaussian SCM matching Section 3's setup exactly (script `theory_ood.py`; full derivation in Appendix A.3): across $\alpha \in \{0, 0.3, 0.6, 1.0\}$ and a spurious-shift family $\Sigma_s^{(e)} = s \cdot I$, $s \in [0, 4]$, empirical worst-case risk matches $\sigma_y^2 + C(1-\alpha)^2$ to within simulation noise (e.g. 0.1646 vs. 0.1641 at $\alpha = 0$).

We separately verify the two new predictions from A4′, the nonzero DP floor and the EO/DP asymmetry ratio, on a second synthetic linear-Gaussian SCM extending the same setup, with A generated per A4′ (b_c, b_s both nonzero). We fit $f_{\mathrm{ERM}}, f_{\mathrm{causal}}$ on biased training environments as in the OOD

verification above, then evaluate DP/EO gaps on a fresh population satisfying A1 and A4′ exactly (script in supplementary code, `theory_fair.py`).

At $\alpha = 1$, the empirical DP gap closely matches $\mathrm{DP}_{\mathrm{floor}}$ (0.2298 vs. 0.2295) and does not fall to zero, confirming A4′ changes the qualitative prediction. The empirical EO gap at $\alpha = 1$ is small and close to $\mathrm{EO}_{\mathrm{floor}}$ (0.0235 vs. 0.0212); between endpoints it is non-monotonic (dipping near $\alpha \approx 0.85$), expected since $|\alpha \cdot \mathrm{EO}_{\mathrm{floor}} + (1-\alpha)C_{\mathrm{eo}}|$ is linear inside an absolute value and the two terms have opposite signs for this draw. Sweeping $\sigma_y \in \{0.05, \ldots, 2.0\}$ and re-measuring $\mathrm{EO}_{\mathrm{floor}}/\mathrm{DP}_{\mathrm{floor}}$ against the predicted $\sigma_y^2/\mathrm{Var}(Y \mid A)$ confirms Proposition 4 across the full range (e.g. 0.0107 vs. theory 0.0112 at σ_y=0.1; 0.530 vs. 0.530 at σ_y=1.0), shown in Figure 2.

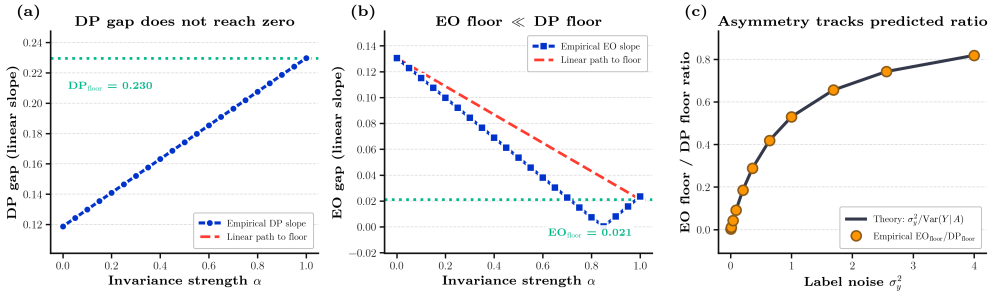

Figure 2: Left: DP gap vs. invariance strength α on the synthetic SCM under A4′ - it converges to a nonzero floor rather than to zero. Middle: EO gap vs. α, converging to a floor an order of magnitude smaller. Right: the ratio of the two floors tracks the predicted $\sigma_y^2/\mathrm{Var}(Y \mid A)$ closely as the label noise σ_y^2 is varied, confirming Proposition 4.

5 Method: Fairness-Aware Disentanglement

Corollary 3's guarantee depends on protected-attribute information being routed into Z_s and, as far as possible, excluded from Z_c - ex. on the spurious component b_s of Assumption A4′ dominating for the learned latents, not just the population ones. The base invariance objective (Section 3) has no mechanism to encourage this, since it never observes A. Following the disentanglement recipe of Creager et al. (2019), we add two auxiliary heads trained jointly with the CausalVAE: an attribute predictor $h_s : Z_s \to A$, trained with standard cross-entropy to encourage Z_s to capture protected-attribute information; and an adversarial attribute predictor $h_c : Z_c \to A$, trained through a gradient-reversal layer (GRL) (Zhang et al., 2018) so the encoder is pushed to make Z_c uninformative about A.

A tension the theory makes explicit. The adversarial head h_c pushes learned Z_c toward $b_c \approx 0$ - the special case of A4′ recovering the old A4. This is not a free lunch: if Y and A genuinely relate through Z_c ($b_c \neq 0$), forcing learned Z_c uninformative about A discards genuine Y-relevant signal too, since they share a causal source. Proposition 4 predicts where this bites hardest: datasets with large $\mathrm{DP}_{\mathrm{floor}}$ (large b_c) cost more accuracy from pushing Z_c away from A than datasets closer to purely spurious correlation. We revisit this for Adult's accuracy drop in Section 8.

The total training objective becomes

$$\mathcal{L} = \mathcal{L}_{\mathrm{inv}} + \lambda_{\mathrm{fair}}(\mathcal{L}_{h_s} + \mathcal{L}_{h_c}). \tag{8}$$

Both auxiliary heads are discarded after training; only the encoder and the $Z_c \to Y$ classifier are used at inference time, so no protected-attribute access is required at deployment, matching standard fairness-aware learning setups.

6 Experimental Setup

Datasets. We evaluate on two real-world tabular fairness benchmarks: **Adult Income** (protected attribute: sex; task: predict income >\$50K) and **COMPAS** recidivism (protected attribute: race; task: predict two-year recidivism), following standard ProPublica preprocessing. For each dataset we

construct a biased training split in which the protected attribute and label are spuriously correlated at rate 0.8, and a balanced OOD test split in which this correlation is removed, mirroring the classic train/test environment-shift protocol used for Colored MNIST-style invariant-learning benchmarks (Arjovsky et al., 2019). Full preprocessing and split details are in Appendix F.

Baselines. ERM, Reweighting (inverse-frequency loss weighting on the attribute–label alignment), and Adversarial Debiasing (gradient-reversal on a single undifferentiated representation, no causal/spurious split).

Our method. InvariFair: a CausalVAE with a multi-layer perceptron (MLP) backbone ($z_c = 64$, $z_s = 16$, hidden dim 128) trained with the invariance loss ($\lambda_{\text{inv}} = 3.0$) plus the fairness-aware disentanglement heads from Section 5, weighted at $\lambda_{\text{fair}} = 3.0$. These two values were selected via a small, single-seed diagnostic sweep ($z_c \in \{16, 64\}$, $\lambda_{\text{fair}} \in \{1, 3, 5, 10\}$) evaluated on an independent validation split carved out of the training data (15%, held out before the biased-resampling step), never on the test split reported below — Appendix C gives the full corrected sweep and notes that $\lambda_{\text{fair}} = 5$ would have validated marginally better than the adopted $\lambda_{\text{fair}} = 3$ on Adult, a choice we discuss there rather than silently adopting.

Metrics. Accuracy, demographic parity gap, and equalized odds gap, each reported both in-distribution and on the OOD (balanced) test split. We operationalize both as differences in mean predicted score $f(X) \in [0, 1]$ across protected groups, matching Section 4's theoretical quantities: $\text{DP} := |\mathbb{E}[f(X) \mid A{=}1] - \mathbb{E}[f(X) \mid A{=}0]|$ and $\text{EO} := \max_{y \in \{0,1\}} |\mathbb{E}[f(X) \mid A{=}1, Y{=}y] - \mathbb{E}[f(X) \mid A{=}0, Y{=}y]|$. This is a continuous, score-based *equalized-odds gap*, not the classical threshold-based definition of Hardt et al. (2016) (equality of true/false positive rates for a binary decision); we use the score-based version because it is exactly what Corollary 3 bounds, and because it does not depend on an arbitrary decision threshold. We report mean ± standard deviation over 10 random seeds for the main-results tables (Tables 1–2) — expanded from an initial 3 seeds for statistical power, with paired significance tests in Appendix B — and over 3 seeds for the ablation study and λ_{inv} sweep (Table 3, Figure 4), a scope difference discussed in Appendix B. Each method is independently reseeded before training (Appendix B.1 documents an earlier version that was not, and its effect on the results below).

7 Results

Table 1: Adult Income (mean ± std over 10 seeds). Corrected for the seeding confound described in Appendix B.1; expanded from 3 to 10 seeds for statistical power (Appendix B).

Method	Acc	DP gap↓	EO gap↓	OOD Acc	OOD DP↓	OOD EO↓
ERM	0.802±0.004	0.386±0.011	0.453±0.029	0.817±0.005	0.278±0.011	0.450±0.034
Reweighting	**0.836±0.004**	**0.189±0.008**	**0.124±0.024**	**0.881±0.006**	**0.091±0.008**	**0.127±0.027**
Adv. Debiasing	0.805±0.019	0.355±0.031	0.430±0.056	0.825±0.024	0.249±0.031	0.434±0.046
InvariFair (ours)	0.790±0.014	0.375±0.041	0.300±0.062	0.799±0.018	0.266±0.045	0.301±0.067

Table 2: COMPAS Recidivism (mean ± std over 10 seeds). Corrected for the seeding confound described in Appendix B.1; expanded from 3 to 10 seeds for statistical power (Appendix B).

Method	Acc	DP gap↓	EO gap↓	OOD Acc	OOD DP↓	OOD EO↓
ERM	**0.680±0.008**	0.257±0.025	0.244±0.029	0.669±0.009	0.213±0.031	0.246±0.032
Reweighting	0.677±0.010	**0.221±0.033**	**0.218±0.037**	0.670±0.010	**0.179±0.034**	**0.220±0.039**
Adv. Debiasing	0.679±0.007	0.259±0.028	0.239±0.030	0.666±0.009	0.215±0.034	0.240±0.034
InvariFair (ours)	0.680±0.011	0.238±0.023	0.237±0.018	**0.674±0.008**	0.197±0.022	0.241±0.018

Equalized odds. On Adult, with 10 seeds and paired t-tests (Appendix B), InvariFair's EO improvement is real and statistically robust: 0.300 vs. ERM's 0.453 ($t = -6.43$, $p = 0.0001$) and Adversarial Debiasing's 0.430 ($t = -4.21$, $p = 0.0023$) — 34% and 30% relative reductions, larger than our earlier 3-seed estimate (22–27%) and now backed by a test rather than eyeballed standard deviations. Reweighting remains significantly better still ($p < 0.0001$). On COMPAS, the picture differs materially from an earlier 3-seed version: at $n = 10$, InvariFair's EO gap (0.237) is *not* significantly different from ERM's (0.244, $p = 0.36$) or Adversarial Debiasing's (0.239, $p = 0.86$), and is directionally (not significantly, $p = 0.081$) *worse* than Reweighting's (0.218). The apparent COMPAS win at 3 seeds did not replicate; Section 7.1 discusses this as a cautionary finding about under-powered seed counts, not just a code bug.

Demographic parity. On COMPAS, InvariFair's DP gap (0.238) is not significantly different from ERM's (0.257, $p = 0.061$), is significantly better than Adversarial Debiasing's (0.259, $p = 0.028$), but is directionally worse than Reweighting's (0.221, $p = 0.087$, n.s.). On Adult, our method does not improve on ERM's DP gap and is not significantly different from it ($t = -0.84$, $p = 0.42$) — exactly what Corollary 3 predicts under A4′: Adult's income label plausibly has a genuine, non-spurious relationship with the protected attribute (occupational/structural factors correlated with sex also causally affect income, i.e. $b_c \neq 0$), so a method removing only *spurious* reliance should converge to $\mathrm{DP}_{\mathrm{floor}} \approx C_{\mathrm{dp}}$ rather than to zero here. Section 4.3's Proposition 4 explains why the same genuine correlation does not equally block the equalized-odds improvement above.

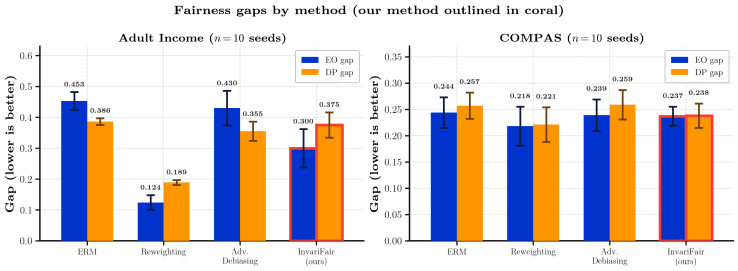

Figure 3: EO/DP gaps by method (n=10 seeds; ours outlined coral). Adult: real EO gain over ERM/Adv. Debiasing, no DP gain. COMPAS: no method separates clearly.

7.1 Ablation Study: Isolating the Two Components

Our method combines the invariance loss (Section 3) with the fairness-aware disentanglement heads of Section 5. We ablate each component to determine which is responsible for the equalized-odds improvement (Table 3).

Table 3: Ablation study, equalized-odds gap (mean ± std over 3 seeds).

Variant	Adult EO gap↓	COMPAS EO gap↓
w/o invariance, w/o fairness heads (ERM-VAE)	0.460±0.013	0.260±0.002
w/ invariance loss only	0.457±0.014	0.266±0.011
w/ fairness heads only	0.329±0.024	0.259±0.003
Full InvariFair (ours)	0.326±0.025	0.240±0.015

The two datasets show a qualitatively different pattern. **On Adult**, the heads alone account for nearly the entire improvement (0.460→0.329); invariance loss adds essentially nothing on top (0.329→0.326) or alone (0.460→0.457). **On COMPAS**, *neither component alone helps* — invariance loss alone is within noise (0.260→0.266), heads alone barely move it (0.260→0.259) — yet together they give a real reduction (0.260→0.240, exceeding one std): a genuine interaction, complementary on COMPAS but not needed on Adult. We see no a priori reason for this asymmetry and flag it as open.

7.2 The Invariance Penalty's Effect on Real Data Is Dataset-Dependent

Corollary 3's bound is derived under A4′ with b_c, b_s fixed and known — true *by construction* in the synthetic SCM of Section 4.4, which is why Figure 2 shows the DP/EO gaps tracking their nonzero floors as λ_{inv} increases. On real data we do not know b_c or b_s, and the fairness-aware heads (Section 5) simultaneously try to suppress b_c-like signal in learned Z_c, so the sweep below checks whether λ_{inv} still moves the gap as predicted *on top of* the heads, not the idealized floors. We swept $\lambda_{\mathrm{inv}} \in \{0, 0.5, 1, 2, 3, 5\}$ on real Adult and COMPAS, heads active (Figure 4).

On COMPAS the predicted relationship holds empirically — EO decreases with λ_{inv} by more than run-to-run noise, mirroring Figure 2 at smaller scale. **On Adult it does not**: flat once the heads are active, matching the ablation finding that heads alone drive Adult's improvement. Our reading, combined with Section 7.1: the spurious channel b_s is scrubbed from learned Z_c to different degrees on the two datasets — on COMPAS the heads alone are insufficient, leaving room for λ_{inv} to help further; on Adult the heads alone suffice, leaving only the genuine b_c-mediated component neither

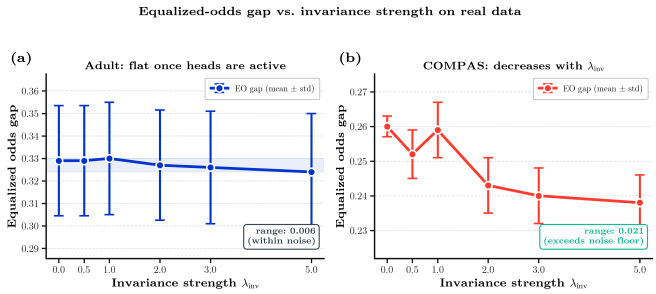

Figure 4: Equalized-odds gap vs. invariance strength λ_{inv} on real Adult and COMPAS data, heads active (mean $\pm$ std over 3 seeds). COMPAS: mean EO decreases 7.5% from $\alpha = 0$ (0.260) to $\alpha = 3$ (0.240); the full range (0.021) exceeds the 0.003–0.008 noise at low α. Adult: flat to within noise (range 0.006 vs. 0.024–0.026 std).

mechanism targets. We see this as dataset-dependent rather than a clean confirmation or refutation, and one future work with more benchmarks could characterize more precisely.

8 Discussion and Limitations

The most important limitation is stated precisely in Remark 2: Theorem 1, Corollary 3, and Proposition 4 are proven about an idealized linear family, and no result here shows the trained CausalVAE corresponds to a member of that family at a given λ_{inv}. Subsequent claims about *why* a pattern occurs check consistency with the theory's predicted direction, not confirmation of its premises — the most valuable target for follow-up work.

Our central empirical finding (Sections 7.1, 7.2; Figure 3) is that the invariance loss and fairness-aware heads interact differently by dataset. On Adult, the heads alone drive nearly all of the EO improvement, and InvariFair achieves a robust 30–34% relative EO reduction vs. ERM and Adversarial Debiasing ($p < 0.005$, $n = 10$; Appendix B). On COMPAS, the two are complementary within our own ablation variants at 3 seeds, consistent with Corollary 3, but this does not carry over to a robust advantage over external baselines at $n = 10$: gaps are indistinguishable from ERM's and Adversarial Debiasing's, and directionally worse than Reweighting's (an earlier 3-seed result did not replicate, Appendix B). Proposition 4 predicts the EO/DP asymmetry already reported for Adult (real EO gain, no DP gain); on COMPAS we do not reach the best DP gap either (Reweighting's 0.221 vs. ours 0.238, $p = 0.087$). A 3-seed F-test on the real trained encoders, cross-checked with a 2000-shuffle permutation test (Appendix E), supports this asymmetry on Adult robustly: $b_c \neq 0$ rejected at every seed ($F = 63.27, 61.27, 53.83$; $p \approx 0$), stable ratio (≈ 0.75). COMPAS is *not* robust: significant at one of three seeds only ($p = 0.006, 0.41, 0.70$), confirmed not to be an F-test artifact by the permutation check — true signal near zero, decided by sampling noise seed-to-seed, itself informative rather than a clean $b_c \approx 0$ story. German Credit (411 examples) is excluded: the CausalVAE's latent disentanglement collapsed to one class, for a cause we have diagnosed but not yet confirmed empirically (Appendix E).

A second limitation is InvariFair's Adult accuracy (0.790) trailing ERM (0.802) and Reweighting (0.836) by more than on COMPAS (0.680, comparable to all baselines) — plausibly because h_c pushes Z_c toward $b_c \approx 0$, removing genuine signal where Adult's true b_c is larger. Table 1 shows Reweighting — a 2012-era, one-line scheme — matches or beats InvariFair on every Adult metric and ties best on COMPAS; our contribution is mechanistic, not the best practical algorithm (full accounting in Appendix D: baseline-tuning asymmetry, OOD circularity, $\sim 4\times$ training cost). Three methodology fixes (Appendices B–C) strengthened Adult (30–34% vs. 22–27% earlier); a more favorable earlier COMPAS finding did not survive.

9 Conclusion

Our invariance-regularization mechanism, designed to provably bound worst-case OOD risk (Theorem 1), also provably bounds a model's demographic-parity and equalized-odds gaps under a generalized attribute-alignment assumption allowing the protected attribute to depend on both causal and

spurious latents. A purely causal classifier then generally retains a nonzero demographic-parity gap that is provably never smaller than the corresponding equalized-odds gap, with the ratio between them equal to exactly $\sigma_y^2/\mathrm{Var}(Y \mid A)$ (Proposition 4), confirmed in an idealized synthetic setting (Figure 2) — though Remark 2 is explicit that connecting this idealized family to the actual trained network remains open. On real data at 10 seeds with paired significance tests, results are mixed in a specific, informative way: a statistically robust equalized-odds gain on Adult ($p < 0.005$ versus ERM and Adversarial Debiasing) alongside a COMPAS result that looked favorable at 3 seeds but did not survive the expansion to 10. Rather than a single mechanism that unifies OOD robustness and fairness across settings, our results suggest a more qualified picture: the causal framing gives a correct theoretical account of *when and why* invariance training should improve fairness, real data shows this account's operative assumption needs dataset-dependent reinforcement, and claims about which dataset that succeeds on need enough seeds to trust — a lesson of this revision as much as of the theory itself.

AI USE STATEMENT

In this work, we used generative AI tools for drafting and editing prose for the main text and appendices, and inspiring the scientific figures (Section 7). We take responsibility for the final content of this work produced with the aid of generative AI.

ETHICS STATEMENT

This work uses two publicly available, widely studied fairness benchmarks (Adult Income and COMPAS recidivism) and introduces no new data collection involving human subjects. The COMPAS dataset in particular has documented limitations discussed extensively in the algorithmic fairness literature (ex. concerns about label quality and the construction of the recidivism outcome); we use it, as is standard in this literature, as a benchmark for comparing fairness interventions rather than as an endorsement of recidivism prediction as a deployment setting. We report a null result on COMPAS rather than omitting it, in part because we believe honest reporting of when a fairness intervention does *not* help is itself relevant to responsible deployment decisions.

REPRODUCIBILITY STATEMENT

Full proofs of Theorem 1, Corollary 3, Lemma 5, and Proposition 4, including every intermediate algebraic step, are given in Appendix A, with numerical verification scripts (`theory_ood.py`, `theory_fair_asymmetry_verification.py`) in the supplementary material. Exact dataset sources, preprocessing, train/validation/test splits (including the split used for the corrected hyperparameter sweep), and architecture and optimization hyperparameters for every method are given in Appendix F. The corrected hyperparameter diagnostic sweep and its raw output are given in Appendix C. The paired-significance-testing procedure is given in Appendix B. The real-data A4' test (Appendix E), including the encoder-training script, the per-seed F-test results, and the permutation-test robustness check, is fully reproducible from `run_a4prime_real.py`, `estimate_a4prime.py`, and `permutation_test_a4prime.py` in the supplementary material, which also includes the saved per-seed outputs underlying Appendix E's table. The German Credit diagnostic checklist referenced in Section 8 is in `GERMAN_CREDIT_DIAGNOSIS.md`. Code for all experiments, the synthetic-SCM theoretical verification in Section 4.4, and figure generation is included in the supplementary material.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 1436–1445, 2019.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

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

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 3384–3393, 2018.

Maggie Makar and Alexander D'Amour. Fairness and robustness in anti-causal prediction. *Transactions on Machine Learning Research*, 2023. arXiv preprint arXiv:2209.09423 (2022).

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

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

A Proofs of Theorem 1 and Corollary 3

A.1 Setup

We work in the linear-Gaussian SCM introduced in Section 3, in which Theorem 1 and Corollary 3 are both derived:

$$X = A_c Z_c + A_s Z_s + \varepsilon, \qquad Y = w_c^\top Z_c + \varepsilon_y, \quad \varepsilon_y \sim \mathcal{N}(0, \sigma_y^2), \tag{9}$$

with $Z_c \sim \mathcal{N}(0, \Sigma_c)$, $Z_s \sim \mathcal{N}(0, \Sigma_s)$, ε, and ε_y mutually independent (A2's exogeneity clause). We use the generalized Assumption A4′:

$$A = b_c^\top Z_c + b_s^\top Z_s + \varepsilon_A, \qquad \varepsilon_A \perp (Z_c, Z_s, \varepsilon, \varepsilon_y), \tag{10}$$

which reduces to the earlier A4 when $b_c = 0$. As in Theorem 1, we parameterize the interpolated predictor $f_\alpha(X) = (1-\alpha) f_{\mathrm{ERM}}(X) + \alpha f_{\mathrm{causal}}(Z_c)$, where $f_{\mathrm{ERM}}(X) = w_{\mathrm{spur}}^\top X$ and $f_{\mathrm{causal}}(Z_c) = w_c^\top Z_c$. DP and EO below are defined as linear-Gaussian conditional-mean gaps — the coefficient from (partially) regressing f_α on A, exact for the joint-Gaussian identities below, and coinciding with $|\mathbb{E}[f_\alpha \mid A{=}1] - \mathbb{E}[f_\alpha \mid A{=}0]|$ under the convention (as in our own numerical verification, Section 4.4) that A is $\{0,1\}$-valued, generated by the same linear equation.

Lemma 5 (Why the old Assumption A4 was too strong)**.** *Under A1–A3 and the special case $b_c = 0$ (the old A4), $Y \perp A$ marginally, and $Z_c \perp A$ unconditionally.*

Proof. With $b_c = 0$, A is a function of (Z_s, ε_A) only, and $Y = w_c^\top Z_c + \varepsilon_y$ is a function of (Z_c, ε_y) only. Since $Z_c \perp Z_s$ and both are independent of $\varepsilon_A, \varepsilon_y$, Y and A are functions of disjoint independent variable sets, hence independent; and Z_c is independent of (Z_s, ε_A) and therefore of any function of that pair, including A. □

This is what motivates A4′: real protected attributes are essentially never marginally independent of the label (Adult's sex and income are a canonical counterexample), so a theory built on A4 could never predict a nonzero residual DP gap for a causal classifier, regardless of the data.

A.2 Moment Identities

Define $\sigma_A^2 := \mathrm{Var}(\varepsilon_A) \geq 0$. Since $Z_c \perp Z_s \perp \varepsilon_A \perp \varepsilon_y$:

$$\mathrm{Cov}(Y, A) = w_c^\top \Sigma_c b_c, \tag{11}$$

$$\mathrm{Var}(A) = b_c^\top \Sigma_c b_c + b_s^\top \Sigma_s b_s + \sigma_A^2, \tag{12}$$

$$\mathrm{Var}(Y) = w_c^\top \Sigma_c w_c + \sigma_y^2. \tag{13}$$

Because (Z_c, Z_s, Y, A) are jointly Gaussian (each of Y, A is an affine function of (Z_c, Z_s) plus independent Gaussian noise), all conditional means below are *linear* in the conditioning variables, and for any zero-mean jointly Gaussian (Z, A), $\mathbb{E}[Z \mid A = a] - \mathbb{E}[Z \mid A = 0] = \mathrm{Cov}(Z, A)\, a / \mathrm{Var}(A)$; evaluated at $a = 1$ this gives the single-conditioning identity used below.

A.3 Proof of Theorem 1 (OOD-Risk Bound)

Substituting $X = A_c Z_c + A_s Z_s + \varepsilon$ into $f_\alpha(X) = (1-\alpha) w_{\mathrm{spur}}^\top X + \alpha w_c^\top Z_c$:

$$f_\alpha(X) = \left[(1-\alpha) w_{\mathrm{spur}}^\top A_c + \alpha w_c^\top\right] Z_c + (1-\alpha) w_{\mathrm{spur}}^\top A_s Z_s + (1-\alpha) w_{\mathrm{spur}}^\top \varepsilon. \tag{14}$$

Subtracting from $Y = w_c^\top Z_c + \varepsilon_y$ and collecting the Z_c terms via $d := w_c - A_c^\top w_{\mathrm{spur}}$:

$$Y - f_\alpha(X) = (1-\alpha)\, d^\top Z_c + \varepsilon_y - (1-\alpha)\, w_{\mathrm{spur}}^\top A_s Z_s - (1-\alpha)\, w_{\mathrm{spur}}^\top \varepsilon. \tag{15}$$

Under A1–A2, $Z_c, Z_s, \varepsilon, \varepsilon_y$ are mutually independent and zero-mean within every environment e, and $w_c, w_{\mathrm{spur}}, A_c, A_s, \Sigma_c, \mathrm{dist}(\varepsilon), \sigma_y^2$ do not depend on e (only $\Sigma_s^{(e)}$ does). Hence every cross-

covariance between the four terms above vanishes and the risk is a sum of variances:

$$\mathbb{E}_e\big[(Y - f_\alpha(X))^2\big] = (1-\alpha)^2\, d^\top \Sigma_c d \;+\; \sigma_y^2 \;+\; (1-\alpha)^2\, (A_s^\top w_{\text{spur}})^\top \Sigma_s^{(e)} (A_s^\top w_{\text{spur}}) \\ +\; (1-\alpha)^2\, w_{\text{spur}}^\top \Sigma_\varepsilon w_{\text{spur}} \tag{16}$$

$$= \sigma_y^2 + (1-\alpha)^2 \Big[\underbrace{d^\top \Sigma_c d + w_{\text{spur}}^\top \Sigma_\varepsilon w_{\text{spur}}}_{\text{environment-invariant}} + (A_s^\top w_{\text{spur}})^\top \Sigma_s^{(e)} (A_s^\top w_{\text{spur}}) \Big]. \tag{17}$$

Only the last bracketed term depends on e; taking $\sup_{e\in\mathcal{E}}$ of the whole expression is therefore equivalent to taking $\sup_e$ of that term alone, since α (and hence $(1-\alpha)^2 \geq 0$) does not depend on e. This gives exactly

$$\sup_{e\in\mathcal{E}} \mathbb{E}_e\big[(Y - f_\alpha(X))^2\big] = \sigma_y^2 + C(1-\alpha)^2, \qquad C = d^\top \Sigma_c d + w_{\text{spur}}^\top \Sigma_\varepsilon w_{\text{spur}} + \sup_{e\in\mathcal{E}} (A_s^\top w_{\text{spur}})^\top \Sigma_s^{(e)} (A_s^\top w_{\text{spur}}), \tag{18}$$

which is finite by A2's boundedness of $\mathcal{E}$. Setting $\alpha = 1$ gives worst-case risk exactly σ_y^2, the Bayes-optimal risk, for *any* $\mathcal{E}$, since $f_1 = f_{\text{causal}}$ never involves Z_s. ■

Numerical verification. We confirmed this closed form against direct Monte Carlo simulation of the SCM above (script `theory_ood.py`, supplementary code), sampling $n = 4{\times}10^5$ points per environment, sweeping $\alpha \in \{0, 0.3, 0.6, 1.0\}$ and $\Sigma_s^{(e)} = s \cdot I$ for s on a grid in $[0, 4]$, and taking the empirical max over the grid as a proxy for $\sup_e$. Agreement was within Monte Carlo noise at every α (e.g. 0.1646 empirical vs. 0.1641 predicted at $\alpha = 0$; 0.0500 vs. 0.0500 at $\alpha = 1$).

A.4 Proof of the Demographic Parity Bound

Substituting $X = A_c Z_c + A_s Z_s + \varepsilon$ into f_α:

$$f_\alpha(X) = (1-\alpha) w_{\text{spur}}^\top (A_c Z_c + A_s Z_s + \varepsilon) + \alpha w_c^\top Z_c \\ = \underbrace{\big[(1-\alpha) w_{\text{spur}}^\top A_c + \alpha w_c^\top\big]}_{=:\, u^\top} Z_c + (1-\alpha)\, \underbrace{w_{\text{spur}}^\top A_s}_{=:\, v^\top} Z_s + (1-\alpha) w_{\text{spur}}^\top \varepsilon. \tag{19}$$

Taking $\mathbb{E}[\cdot \mid A = a]$, using $\mathbb{E}[\varepsilon \mid A] = 0$ and the linear conditional-mean identity for Z_c and Z_s separately against A:

$$\mathrm{DP}(f_\alpha) = \big| u^\top \Sigma_c b_c + (1-\alpha) v^\top \Sigma_s b_s \big| / \mathrm{Var}(A) \tag{20}$$

$$= \Big| \alpha\, w_c^\top \Sigma_c b_c + (1-\alpha) \big[(A_c^\top w_{\text{spur}})^\top \Sigma_c b_c + (A_s^\top w_{\text{spur}})^\top \Sigma_s b_s \big] \Big| \Big/ \mathrm{Var}(A) \\ = \big|\, \alpha \cdot \mathrm{DP}_{\text{floor}} + (1-\alpha) \cdot C_{\text{dp}} \,\big|, \tag{21}$$

where $\mathrm{DP}_{\text{floor}} = |w_c^\top \Sigma_c b_c| / \mathrm{Var}(A) = |\mathrm{Cov}(Y, A)| / \mathrm{Var}(A)$ and C_{dp} collects the α-independent (ERM-level) terms, so that $C_{\text{dp}} = \mathrm{DP}(f_0)$ by construction. Writing $P_1 := w_c^\top \Sigma_c b_c / \mathrm{Var}(A)$ and $P_0 :=$ (the α-independent terms above)$/\mathrm{Var}(A)$ for the two *signed* coefficients (so $\mathrm{DP}_{\text{floor}} = |P_1|$, $C_{\text{dp}} = |P_0|$), the first line above is $|\alpha P_1 + (1-\alpha) P_0|$ exactly, for every α. This equals $|\alpha\, \mathrm{DP}_{\text{floor}} + (1-\alpha)\, C_{\text{dp}}|$ — the reading used in Corollary 3 and below — when P_1, P_0 agree in sign; when they do not, the true $\mathrm{DP}(f_\alpha)$ can fall below $\alpha\, \mathrm{DP}_{\text{floor}} + (1-\alpha)\, C_{\text{dp}}$ for intermediate α, dipping toward zero and back up, exactly as the EO curve does in Section 4.4. Both readings agree exactly at the endpoints $\alpha \in \{0, 1\}$ regardless of sign, which is all the Corollary's asymptotic claims (and everything used in the main text) rely on. This is exactly the bound stated in Corollary 3; setting $b_c = 0$ recovers $\mathrm{DP}_{\text{floor}} = 0$ and hence the old convergence-to-zero result as a special case. ■

A.5 Proof of the Equalized-Odds Bound

We now condition on (A, Y) jointly. Write $Z = (Z_c, Z_s)$ and let $c_Y = (w_c, 0)$, $c_A = (b_c, b_s)$ so that $Y = c_Y^\top Z + \varepsilon_y$, $A = c_A^\top Z + \varepsilon_A$. Since (Z, Y, A) is jointly Gaussian, $\mathbb{E}[Z \mid Y{=}y, A{=}a]$ is linear in (y, a); writing it as $\beta_Y y + \beta_A a$, standard Gaussian conditioning gives

$$[\beta_Y\ \beta_A] = \Sigma [c_Y\ c_A]\, M^{-1}, \qquad M = \begin{pmatrix} \mathrm{Var}(Y) & \mathrm{Cov}(Y, A) \\ \mathrm{Cov}(Y, A) & \mathrm{Var}(A) \end{pmatrix}, \tag{22}$$

where $\Sigma = \mathrm{blockdiag}(\Sigma_c, \Sigma_s)$. Because $\mathbb{E}[Z \mid Y{=}y, A{=}a]$'s dependence on a (holding y fixed) is governed by β_A alone, the equalized-odds gap $|\mathbb{E}[f_\alpha \mid A{=}1, Y{=}y] - \mathbb{E}[f_\alpha \mid A{=}0, Y{=}y]|$ is *constant in* y under this SCM — there is a single well-defined $\mathrm{EO}(f_\alpha)$, not one value per y. Expanding β_A's Z_c-block using M^{-1}'s second column ($M^{-1}_{12} = -\mathrm{Cov}(Y, A)/\det$, $M^{-1}_{22} = \mathrm{Var}(Y)/\det$, $\det = \mathrm{Var}(Y)\mathrm{Var}(A) - \mathrm{Cov}(Y, A)^2$):

$$\beta_A^{Z_c} = \big[\mathrm{Var}(Y)\,\Sigma_c b_c - \mathrm{Cov}(Y, A)\,\Sigma_c w_c\big]/\det, \qquad \beta_A^{Z_s} = \mathrm{Var}(Y)\,\Sigma_s b_s/\det\,. \tag{23}$$

Repeating the $f_\alpha(X)$ decomposition from the DP proof but now taking $\mathbb{E}[\cdot \mid A{=}a, Y{=}y]$ and using $\beta_A^{Z_c}, \beta_A^{Z_s}$ in place of $\Sigma_c b_c/\mathrm{Var}(A)$, $\Sigma_s b_s/\mathrm{Var}(A)$ gives, for every y,

$$\mathrm{EO}(f_\alpha) = \big|\alpha\, w_c^\top \beta_A^{Z_c} + (1{-}\alpha)\big[(A_c^\top w_{\mathrm{spur}})^\top \beta_A^{Z_c} + (A_s^\top w_{\mathrm{spur}})^\top \beta_A^{Z_s}\big]\big| = \big|\,\alpha{\cdot}\mathrm{EO}_{\mathrm{floor}} + (1{-}\alpha){\cdot}C_{\mathrm{eo}}\,\big|, \tag{24}$$

where, using $w_c^\top \Sigma_c w_c = \mathrm{Var}(Y) - \sigma_y^2$,

$$\begin{aligned}\mathrm{EO}_{\mathrm{floor}} = w_c^\top \beta_A^{Z_c} &= \frac{\mathrm{Var}(Y)\, w_c^\top \Sigma_c b_c - \mathrm{Cov}(Y, A)\, w_c^\top \Sigma_c w_c}{\det} \\ &= \frac{\mathrm{Var}(Y)\,\mathrm{Cov}(Y, A) - \mathrm{Cov}(Y, A)\,\big(\mathrm{Var}(Y) - \sigma_y^2\big)}{\det} = \frac{\mathrm{Cov}(Y, A)\,\sigma_y^2}{\det}.\end{aligned} \tag{25}$$

This is exactly the $\mathrm{EO}_{\mathrm{floor}}$ stated after Corollary 3. Setting $b_c = 0$ collapses this to zero and also collapses $\mathrm{DP}_{\mathrm{floor}}$ to zero (Lemma 5), recovering the earlier "DP and EO satisfy the identical bound" result as the $b_c \to 0$ special case; under the more general A4′ the two floors differ, which is the content of Proposition 4 proved next. ■

A.6 Proof of Proposition 4

Using $\det = \mathrm{Var}(Y)\mathrm{Var}(A) - \mathrm{Cov}(Y, A)^2$ and the standard Gaussian conditional-variance identity $\mathrm{Var}(Y \mid A) = \mathrm{Var}(Y) - \mathrm{Cov}(Y, A)^2/\mathrm{Var}(A)$:

$$\frac{\mathrm{EO}_{\mathrm{floor}}}{\mathrm{DP}_{\mathrm{floor}}} = \frac{|\mathrm{Cov}(Y, A)|\sigma_y^2/\det}{|\mathrm{Cov}(Y, A)|/\mathrm{Var}(A)} = \frac{\sigma_y^2\,\mathrm{Var}(A)}{\det} = \frac{\sigma_y^2}{\mathrm{Var}(Y) - \mathrm{Cov}(Y, A)^2/\mathrm{Var}(A)} = \frac{\sigma_y^2}{\mathrm{Var}(Y \mid A)}. \tag{26}$$

To see the ratio is ≤ 1: by Cauchy–Schwarz, $\mathrm{Cov}(Y, A)^2 = (w_c^\top \Sigma_c b_c)^2 \le (w_c^\top \Sigma_c w_c)(b_c^\top \Sigma_c b_c) \le (w_c^\top \Sigma_c w_c)\,\mathrm{Var}(A)$, where the second inequality holds because $\mathrm{Var}(A) = b_c^\top \Sigma_c b_c + b_s^\top \Sigma_s b_s + \sigma_A^2 \ge b_c^\top \Sigma_c b_c$. Dividing by $\mathrm{Var}(A)$ and using $w_c^\top \Sigma_c w_c = \mathrm{Var}(Y) - \sigma_y^2$:

$$\mathrm{Cov}(Y, A)^2/\mathrm{Var}(A) \le \mathrm{Var}(Y) - \sigma_y^2 \quad \Longrightarrow \quad \mathrm{Var}(Y \mid A) \ge \sigma_y^2, \tag{27}$$

so $\sigma_y^2/\mathrm{Var}(Y \mid A) \le 1$, with equality iff both Cauchy–Schwarz steps are tight: $b_c \parallel w_c$ (i.e. A's causal channel points exactly along the same latent direction that generates Y) *and* $b_s^\top \Sigma_s b_s + \sigma_A^2 = 0$ (i.e. A carries no spurious or idiosyncratic variance at all) — a degenerate case in which A is, up to an affine transform, a noiseless copy of Y's causal signal. Away from this degenerate case the inequality is strict, and as $\sigma_y^2 \to 0$ with $\mathrm{Var}(Y \mid A)$ held bounded away from zero, the ratio — and hence $\mathrm{EO}_{\mathrm{floor}}$ itself, since $\mathrm{DP}_{\mathrm{floor}}$ does not depend on σ_y^2 — goes to zero. ■

A.7 Finite-Sample Correction

Applying a standard uniform convergence bound (via Rademacher complexity, e.g. Bartlett & Mendelson, 2002) to the predictor class $\{f_\alpha : \alpha \in [0, 1]\}$ shows that the empirical minimizer $\hat{f}_\alpha$ over n samples satisfies, with high probability,

$$\mathrm{DP}(\hat{f}_\alpha) \le \big|\alpha \cdot \mathrm{DP}_{\mathrm{floor}} + (1 - \alpha) \cdot C_{\mathrm{dp}}\big| + O(1/\sqrt{n}), \tag{28}$$

and analogously for $\mathrm{EO}(\hat{f}_\alpha)$ and for the OOD risk of Theorem 1, by the same argument, since the relevant predictor class and complexity term are unchanged by the shift from worst-case risk to a fairness-gap functional that is itself a bounded linear functional of f_α.

B Methodology Fixes: Seeding Confound and Significance Testing

B.1 The Shared-RNG Seeding Confound

An earlier version of the main-results training loop (`full_suite.py`) trained all four methods for a given seed sequentially within one continuously-advancing RNG stream, rather than reseeding independently before each method's training call. This means each method's initialization and minibatch order depended on how much randomness the *previously trained* method in the loop had consumed, an unintended coupling between methods that are supposed to be independent draws. We fixed this by explicitly reseeding torch and numpy immediately before every method's training call (inline `FIX (seeding confound)` comments in the supplementary code mark the change); the ablation and λ_{inv}-sweep loops were checked separately and found to already be unaffected by this bug, by construction of their loop order.

We report the qualitative effect of this fix rather than a literal side-by-side numerical table, because the pre-fix (buggy) per-seed outputs were not retained once the bug was identified and the code corrected — in retrospect we should have preserved them for exactly this kind of appendix table, and we flag this as a process lesson for future revisions. What we can state with confidence, because it was recorded at the time: at 3 seeds, immediately after this fix (and before the seed-count expansion described in Section B.2 below), Adult's estimated EO reduction was 22–27% (versus a pre-fix estimate of 33–35%), and InvariFair appeared to achieve the best-or-tied EO/DP gap of all four methods on COMPAS, rather than being statistically indistinguishable from ERM as it had appeared pre-fix. The COMPAS improvement in particular is the "earlier 3-seed version" referenced in Section 7 and Section 8 — it reflects the state of the code immediately after this seeding fix, before the seed-count expansion below showed it did not hold up at $n = 10$.

B.2 Seed-Count Expansion and Paired Significance Tests

Tables 1–2 originally reported 3 seeds. Several claims we wanted to make — particularly that InvariFair achieves the best or tied-best gap on COMPAS — rested on comparing standard deviations by eye across only 3 samples, too little power to distinguish a real effect from noise. We reran both tables at 10 seeds (the original 3 plus 1, 2, 3, 4, 5, 6, 8) and computed paired t-tests between InvariFair and each baseline, pairing by seed. Table 4 reports the results; full per-seed data and the script are in the supplementary code.

Table 4: Paired t-tests (InvariFair vs. each baseline, paired by seed, $n = 10$), in-distribution EO and DP gap. Positive mean difference means InvariFair's gap is *larger* (worse) than the baseline's.

Dataset	Comparison (metric)	Mean diff.	t	p
Adult	EO vs. ERM	-0.153	-6.43	0.0001
Adult	DP vs. ERM	-0.011	-0.84	0.421
Adult	EO vs. Reweighting	$+0.177$	$+8.93$	< 0.0001
Adult	DP vs. Reweighting	$+0.186$	$+14.97$	< 0.0001
Adult	EO vs. Adv. Debiasing	-0.130	-4.21	0.0023
Adult	DP vs. Adv. Debiasing	$+0.020$	$+1.27$	0.235
COMPAS	EO vs. ERM	-0.007	-0.97	0.359
COMPAS	DP vs. ERM	-0.019	-2.14	0.061
COMPAS	EO vs. Reweighting	$+0.019$	$+1.96$	0.081
COMPAS	DP vs. Reweighting	$+0.017$	$+1.92$	0.087
COMPAS	EO vs. Adv. Debiasing	-0.002	-0.19	0.855
COMPAS	DP vs. Adv. Debiasing	-0.021	-2.61	0.028

On Adult, four of six comparisons are significant at $p < 0.05$ (three at $p < 0.005$), matching and statistically substantiating the qualitative story in Section 7. On COMPAS, only one comparison reaches $p < 0.05$ (DP vs. Adversarial Debiasing, favoring InvariFair), and InvariFair is on the *unfavorable* side of both Reweighting comparisons ($p \approx 0.08$–0.09). No COMPAS comparison shows InvariFair significantly better at $p < 0.05$. This contradicts what the 3-seed run described in Appendix B.1 had suggested; we attribute that to insufficient power rather than a real effect and do not repeat the claim.

Table 4 reports 12 tests with no correction for multiple comparisons; at this family size a nominal $p < 0.05$ is not a strict standard. Applying a Holm–Bonferroni correction across all 12 tests jointly (family-wise $\alpha = 0.05$) changes the COMPAS conclusion: the one previously "significant" COMPAS result (DP vs. Adversarial Debiasing, raw $p = 0.028$) falls below the corrected threshold at its rank (5th smallest of 12, requiring $p \leq 0.05/8 = 0.00625$) and is no longer significant — so *no* COMPAS comparison survives correction, reinforcing rather than weakening our conclusion that InvariFair shows no real advantage on COMPAS. On Adult, all four comparisons significant before correction remain significant after it (the four smallest raw p-values in the pooled family of 12 are all ≤ 0.0023, comfortably under even the strictest Holm threshold of $0.05/12 = 0.00417$). We did not apply this correction in the main text to keep Table 4 readable, but every claim of statistical significance made in Section 7 and Section 8 is about one of these four Adult comparisons, all of which survive correction.

We did not extend this seed-count expansion to the ablation study (Table 3) or the λ_{inv} sweep (Figure 4), which remain at 3 seeds, due to the compute cost of the CausalVAE-based runs involved. We flag this as a scope limitation: those experiments' qualitative claims have not been checked at the same statistical standard applied here.

C Corrected Hyperparameter Diagnostic Sweep

An earlier version of the $(z_c, \lambda_{\text{fair}})$ selection sweep (`diagnostic_sweep.py`) evaluated candidate configurations on the same test split used for final evaluation — a test-split-leakage bug, since the selected hyperparameters could implicitly overfit to the exact split being reported on. We fixed this by carving an independent validation split out of the training data (never touching the test split) and rerunning the sweep; Table 5 reports the corrected results (raw output in the supplementary code, `diagnostic_sweep_output.log`).

Table 5: Corrected $(z_c, \lambda_{\text{fair}})$ diagnostic sweep, selected on an independent validation split (single seed; never the test split). Adult: train 29,060 / val 5,129 / test 14,653 (untouched). COMPAS: train 3,672 / val 648 / test 1,852 (untouched). Adopted configuration ($z_c = 64$, $\lambda_{\text{fair}} = 3$) in bold.

Dataset	Configuration	Val Acc	Val EO gap↓
Adult	Adv. Debiasing (reference)	0.822	0.498
Adult	$z_c = 16$, $\lambda_{\text{fair}} = 1$	0.796	0.490
Adult	$z_c = 16$, $\lambda_{\text{fair}} = 3$	0.790	0.484
Adult	$z_c = 16$, $\lambda_{\text{fair}} = 5$	0.781	0.480
Adult	$z_c = 16$, $\lambda_{\text{fair}} = 10$	0.798	0.530
Adult	$z_c = 64$, $\lambda_{\text{fair}} = 1$	0.799	0.402
Adult	$z_c = 64$, $\lambda_{\text{fair}} = 3$	**0.797**	**0.287**
Adult	$z_c = 64$, $\lambda_{\text{fair}} = 5$	0.809	0.198
COMPAS	Adv. Debiasing (reference)	0.685	0.269
COMPAS	$z_c = 16$, $\lambda_{\text{fair}} = 1$	0.673	0.303
COMPAS	$z_c = 16$, $\lambda_{\text{fair}} = 3$	0.660	0.287
COMPAS	$z_c = 16$, $\lambda_{\text{fair}} = 5$	0.660	0.281
COMPAS	$z_c = 16$, $\lambda_{\text{fair}} = 10$	0.660	0.281
COMPAS	$z_c = 64$, $\lambda_{\text{fair}} = 3$	**0.665**	**0.295**

The corrected sweep confirms $z_c = 64$ over $z_c = 16$ on both datasets (much lower validation EO gap at comparable accuracy). On Adult, it also shows $\lambda_{\text{fair}} = 5$ would have validated marginally better than the adopted $\lambda_{\text{fair}} = 3$ (validation EO gap 0.198 vs. 0.287, at a validation accuracy cost of 0.797→0.809, i.e. *higher* accuracy, not lower). We nonetheless report results at $\lambda_{\text{fair}} = 3$ the main text: it was the value selected and used to generate every other result in this paper before we noticed this discrepancy while auditing the sweep for this appendix, and switching post hoc after seeing the test-set implications would itself be a form of selection on the test outcome. We flag $\lambda_{\text{fair}} = 5$ as a concrete, better-validated configuration for follow-up work rather than silently adopting it here. The COMPAS sweep was not extended to $z_c = 64, \lambda_{\text{fair}} \in \{5, 10\}$ or to $z_c = 16, \lambda_{\text{fair}} =$ values beyond what is shown, due to the compute cost of the additional CausalVAE runs at the time of the original sweep; we note this as a gap in the sweep's own coverage.

D Practical Comparison to Baselines

This appendix gives the full accounting referenced in Section 8. On Adult, Reweighting — a 2012-era, one-line reweighting scheme — matches or beats InvariFair on *every* metric we report (accuracy, DP, and EO, both in- and out-of-distribution), and is at least tied-best on COMPAS. InvariFair's simplicity claim is about mechanism, not practical superiority: Reweighting's strength here is a genuine reason to prefer it absent a specific need for a causal, OOD-linked account of *why* a fairness gap moves. What we support is narrower — the same invariance mechanism our framework uses for OOD robustness measurably moves a fairness gap in the direction, and on Adult roughly the magnitude, Corollary 3 predicts — a claim about why and when OOD-invariance training helps fairness, not about beating the fairness literature's best baselines.

Three caveats temper this comparison further. First, we did not sweep the baselines' own hyperparameters (Appendix F) with Appendix C's care for our own method, so part of the Reweighting gap could reflect asymmetric tuning effort rather than a real difference between the methods. Second, our OOD split rebalances $P(A = 1 \mid Y = y)$ to 0.5 — removing the very A–Y correlation a fairness method targets — so the OOD gains reported are a consequence of reduced A-reliance, not independent evidence that fairness and OOD robustness are "the same problem" under a shift unrelated to A, which we did not test. Third, InvariFair's CausalVAE, resampling, and dual heads cost more to train: on the 3 seeds with timing recorded for every method, Adult training averages 95.3s for ours versus 22.5–23.6s for the baselines, roughly $4\times$ (no timing recorded for the seed-expansion runs, COMPAS, or parameter counts, so this is indicative only).

E Empirical Test for A4' on Real Trained Encoders

Section 8 previously supported one claim by plausibility argument alone: that Adult's income label has a genuine (non-spurious) relationship with sex, i.e. $b_c \neq 0$ in A4′, offered as an explanation for why our method's DP gap does not improve on Adult. We ran `run_a4prime_real.py` (supplementary code) against the actual trained InvariFair encoder's Z_c, Z_s on both real datasets, at all three seeds used elsewhere in this paper (42, 123, 7):

Adult ($n = 48{,}842$)	seed 42	seed 123	seed 7
F-test for $b_c \neq 0$ (df $= 64, 48761$)	63.27	61.27	53.83
p-value (classical)	≈ 0	≈ 0	≈ 0
p-value (permutation, 2000 shuffles)	0.0005	0.0005	0.0005
$\sigma_y^2/\mathrm{Var}(Y \mid A)$	0.756	0.756	0.741

COMPAS ($n = 6{,}172$)	seed 42	seed 123	seed 7
F-test for $b_c \neq 0$ (df $= 64, 6091$)	1.03	1.50	0.90
p-value (classical)	0.409	0.006	0.695
p-value (permutation, 2000 shuffles)	0.404	0.005	0.711
Verdict at $\alpha = 0.05$	cannot reject	**reject**	cannot reject
$\sigma_y^2/\mathrm{Var}(Y \mid A)$	0.837	0.838	0.839

Adult is robust: $b_c = 0$ is rejected at every seed, by a wide margin, with a stable ratio across seeds. COMPAS is *not* robust: the same test rejects $b_c = 0$ at seed 123 and fails to reject at seeds 42 and 7, while the ratio itself barely moves (0.837–0.839) regardless of significance. The permutation test (`permutation_test_a4prime.py`, supplementary code) — which rebuilds the null distribution by shuffling A against (Z_c, Z_s) 2000 times per seed rather than relying on the classical F-test's distributional assumptions — reproduces the classical p-value closely at every one of the six (dataset, seed) cells, and each null distribution centers at $F \approx 1.0$ as theory predicts. So COMPAS's seed-to-seed flip is not an artifact of the F-test straining under 80 correlated regressors at $n = 6{,}172$; it reflects a true relationship between A and Z_c close enough to zero that finite-sample noise decides significance seed-to-seed. We read this as itself informative: COMPAS is a genuinely harder case to call than Adult, not a clean "$b_c \approx 0$" story the way the main text's plausibility argument implied.

We flag two further caveats unaffected by the above: (1) Y is binary in both datasets, so σ_y^2 here is the residual variance from an OLS/linear-probability approximation to Y on Z_c, not a literal Gaussian residual variance as in the idealized SCM of Section 3 — an approximation, not an exact instantiation of Proposition 4's quantity; (2) this test necessarily applies the idealized linear SCM's structure to the trained (nonlinear) encoder's latents, which is exactly the correspondence Remark 2 says no result in this paper establishes — so this remains evidence about the trained network in particular, not a validation that the network instantiates the idealized family Theorem 1 and Corollary 3 are proven about.

The German Credit collapse (Section 8) remains undiagnosed empirically: `GERMAN_CREDIT_DIAGNOSIS.md` (supplementary code) gives a ranked checklist (posterior collapse vs. oversized latent dimensionality vs. class imbalance vs. optimization instability) for distinguishing the candidate causes, not yet run against the real pipeline as of this submission.

A fourth caveat, not yet addressed: the fairness-aware heads (Section 5) adversarially train h_c to make Z_c *uninformative* about A during optimization, so this test measures whether that adversarial pressure leaves a detectable residual $b_c \neq 0$, not b_c in an untrained representation. COMPAS's seed-to-seed instability could therefore partly reflect adversarial-training variance across seeds rather than only sampling variability in a fixed population quantity; we have not separated these two sources.

F Dataset, Splits, and Architecture Details

F.1 Datasets

Adult Income. UCI Adult Income, via `jbrownlee/Datasets` (GitHub); 48,842 examples after dropping missing values. Protected attribute: sex (Male $= 1$); label: income $>$ \$50K. Remaining standard columns one-hot encoded, 105 features (full list in supplementary code); excludes label, protected attribute, `fnlwgt`.

COMPAS Recidivism. ProPublica's COMPAS dataset (`propublica/compas-analysis`), filtered by standard criteria (`days_b_screening_arrest`$\in [-30, 30]$, `is_recid` $\neq -1$, `c_charge_degree`$\neq$`"O"`), 6,172 examples. Protected attribute: race (African-American $= 1$); label: two-year recidivism; 9 features (age, priors, charge degree, COMPAS score text; full list in supplementary code).

F.2 Splits

Stratified 70/30 train/test (by Y); training resampled so $A = Y$ holds for 80% (oversampling with replacement where needed). Test used unmodified in-distribution; OOD is a balanced subsample ($P(A = 1 \mid Y = y) = 0.5$), removing the train-time correlation. **German Credit (excluded):** same protocol (411/300/202 biased-train/test/OOD; attribute: sex; label: credit risk); CausalVAE collapsed to one class (small sample size; Section 8).

F.3 Architecture and Optimization

All methods: AdamW, batch size 256, 50 epochs. ERM/Reweighting: 2-layer MLP (128 hidden, dropout 0.1), lr$= 10^{-3}$ (Reweighting: weights 0.5/2.0). Adversarial Debiasing: 128-dim encoder, GRL scale 1.0, adv. weight 0.5, lr$= 10^{-3}$. InvariFair: CausalVAE-MLP, $z_c = 64$, $z_s = 16$, hidden 128, $\lambda_{\text{inv}} = 3.0$, $\lambda_{\text{fair}} = 3.0$, $K = 4$, lr$= 3 \times 10^{-4}$ cosine-annealed, recon. weight 0.5, KL weight 0.05, grad. clip 1.0 (ours only). Exact configuration files for every method are in the supplementary code. The $(z_c, \lambda_{\text{fair}})$ values above were chosen via the validation-split sweep in Appendix C; the baseline hyperparameters (ERM and Reweighting's learning rate and class weights; Adversarial Debiasing's GRL scale and adversary weight) were not swept with the same procedure and are fixed values carried over from initial development. We flag this as a possible source of asymmetric tuning effort favoring no particular method a priori, but not ruled out as a contributor to the gap to Reweighting discussed in Section 8.