ENDPOINT EXPOSURE AND CONSTRAINT COVERAGE IN MIXUP UNDER SPURIOUS CORRELATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Selective Mixup changes both interpolation and endpoint sampling. We study these effects using two clean-endpoint baselines. The policy-matched baseline follows the original pair sampler and coefficient weights, whereas the marginal-matched baseline reproduces pooled group-position exposure with independently sampled endpoint groups. For binary labels and attributes, we define Disentangling Pair Exposure (DPE) as the probability that a sampled pair differs in exactly one variable. DPE and complementary same-polarity coverage jointly determine endpoint-polarity balance. In a linear squared-loss model with $\lambda \sim \mathrm{Beta}(\alpha, \alpha)$ and finite $\alpha > 0$, Mixup and endpoint training uniquely identify the noiseless core and spurious signal coefficients if and only if both endpoint polarities have positive mass. Positive DPE is sufficient but not necessary for noiseless signal identification, while pair topology can change objective curvature. Experiments on synthetic data, Colored MNIST, biased CIFAR-10, Waterbirds, and CelebA show that endpoint sampling alone can produce substantial changes in worst-group accuracy. With balanced synthetic endpoint groups at source agreement probability $\rho = 0.98$, mean Mixup worst-group accuracy ranges from 90.34% to 90.59% as DPE varies from zero to one. The remaining Mixup–endpoint differences depend on the pair relation and interpolation space. These results motivate evaluating selective Mixup together with clean-endpoint baselines derived from its sampling policy.

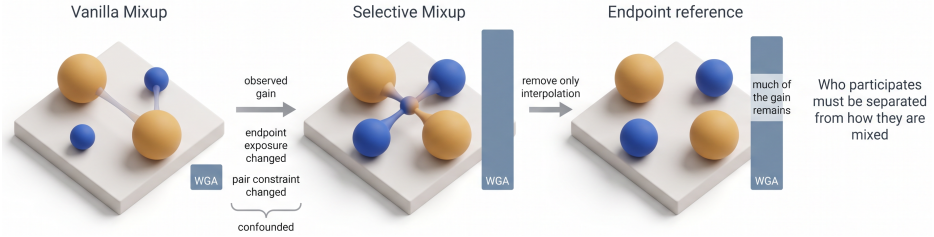

Figure 1: **Pair selection changes exposure and interpolation.** Sphere size denotes group participation across endpoint positions; bridges denote interpolation. Policy-matched endpoint references retain the pair sampler and coefficient weights but remove interpolation. Bars are illustrative.

1 INTRODUCTION

When selective Mixup improves worst-group accuracy (WGA), the change may arise from interpolation, from the distribution of sampled endpoints, or from both. This distinction is particularly relevant under spurious correlations, where predictors can perform well on common groups while failing on rare label–attribute combinations (Geirhos et al., 2020; Sagawa et al., 2020).

Mixup trains on interpolated examples and targets (Zhang et al., 2018). Prior work has studied its effects through local linearity, calibration, and regularization (Guo et al., 2019; Thulasidasan et al., 2019; Carratino et al., 2022). Related methods modify the representation layer, spatial mixing operator, or target weighting (Verma et al., 2019; Yun et al., 2019; Harris et al., 2020; Kim et al., 2020; Chou et al., 2020). LISA constructs pairs conditional on labels and environments, including same-label/different-environment and different-label/same-environment pairs (Yao et al., 2022). When

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.

environments correspond to spurious attributes, such policies also alter which groups participate in training. This sampling effect is relevant because balancing alone can improve group robustness without requiring interpolation between training examples (Idrissi et al., 2022).

We study selective Mixup using clean-endpoint objectives derived from the same sampling policies. The policy-matched endpoint reference applies the sampled coefficient weights to losses on the two original endpoints. The marginal-matched reference instead samples endpoint groups independently from the policy's pooled group-position exposure. These objectives provide two comparisons between interpolation and training on the endpoint distributions induced by selective pairing (Figure 1).

We also characterize pair topology for binary labels Y and attributes A, both encoded in $\{-1, +1\}$. The first letter of a topology denotes whether labels are the same (S) or different (D), and the second denotes the corresponding attribute relation, giving SS, SD, DS, and DD. We define Disentangling Pair Exposure (DPE) as the total probability of SD and DS pairs. Equivalently, DPE is the mass of pairs that differ in polarity $S = YA$. Together with complementary SS/DD coverage, it gives an exact decomposition of endpoint-polarity balance. A linear squared-loss analysis separates signal identification from objective curvature. Both Mixup and endpoint training uniquely identify the noiseless core and spurious signal coefficients when both endpoint polarities have positive mass. Positive DPE is sufficient but not necessary: SS/DD support at both polarities satisfies the same identification condition. Pair topology can nevertheless change the curvature of the objective.

Experiments on synthetic data, Colored MNIST, biased CIFAR-10, Waterbirds, and CelebA compare endpoint exposure, pair topology, and interpolation space. Under balanced synthetic endpoint exposure, mean WGA changes little as DPE varies from zero to one. On Waterbirds and CelebA, policy contrasts have the same direction under Mixup and clean-endpoint training. Their magnitudes, and the Mixup–endpoint differences, vary across datasets, pair relations, and endpoint implementations.

We provide **(i)** clean-endpoint baselines for analyzing the sampling distributions induced by selective Mixup; **(ii)** exact binary coverage identities and a characterization of the signal-identification condition shared by Mixup and endpoint training; and **(iii)** controlled experiments that vary endpoint exposure, pair topology, and interpolation space. Additional related work appears in Appendix A.

2 What a pairing policy changes

Let $(X, Y, A) \sim \mathcal{D}$, where $Y, A \in \{-1, +1\}$ are the label and spurious attribute, respectively, and $G = (Y, A) \in \mathcal{G} = \{-1, +1\}^2$ is the group. Let $\mathcal{D}_{\text{eval}}$ denote the evaluation distribution. For a prediction $\widehat{Y}$, define

$$a_g = \mathbb{P}_{\mathcal{D}_{\text{eval}}}(\widehat{Y} = Y \mid G = g), \qquad \pi_g^{\text{eval}} = \mathbb{P}_{\mathcal{D}_{\text{eval}}}(G = g).$$

Worst-group accuracy, group-balanced accuracy, and average accuracy are

$$\text{WGA} = \min_{g \in \mathcal{G}} a_g, \qquad \text{GBA} = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} a_g, \qquad \text{Avg} = \sum_{g \in \mathcal{G}} \pi_g^{\text{eval}} a_g.$$

Every evaluation group is represented. Controlled evaluation splits are group-balanced, so Avg equals GBA; real benchmarks use their standard evaluation splits.

2.1 Endpoint exposure

Indices i and j denote the first (anchor) and second (partner) endpoints, respectively, of an ordered sampled pair. A pairing policy induces a distribution P over such pairs. Its group-pair matrix, position-specific group marginals, and pooled group-position exposure are

$$M_{g,h} = \mathsf{P}(G_i = g, G_j = h), \tag{1}$$

$$E_g^{(i)} = \sum_h M_{g,h}, \qquad E_g^{(j)} = \sum_h M_{h,g}, \tag{2}$$

$$E_g^{\text{pos}} = \frac{1}{2}\left(E_g^{(i)} + E_g^{(j)}\right). \tag{3}$$

Thus E_g^{pos} is the probability that a uniformly chosen endpoint position belongs to group g. Write $E^{\text{pos}} = (E_g^{\text{pos}})_{g \in \mathcal{G}}$ for the resulting group distribution.

Let f_θ be a predictor with parameters θ, and let ℓ be an additive per-example loss. For an observed endpoint (x_k, y_k, a_k), the target is $y_k^{\mathrm{tar}} = y_k$ for squared loss and $y_k^{\mathrm{tar}} = z_k = (y_k + 1)/2 \in \{0, 1\}$ for binary cross-entropy. Independently of the pair, draw $\lambda \sim \mathrm{Beta}(\alpha, \alpha)$, with $\alpha > 0$. Input Mixup uses

$$\widetilde{x} = \lambda x_i + (1 - \lambda) x_j, \qquad \widetilde{y}^{\mathrm{tar}} = \lambda y_i^{\mathrm{tar}} + (1 - \lambda) y_j^{\mathrm{tar}}.$$

For feature Mixup, write $f_\theta = c_\theta \circ h_\theta$, where h_θ is the encoder and c_θ is the classifier head. The head receives

$$\widetilde{h} = \lambda h_\theta(x_i) + (1 - \lambda) h_\theta(x_j)$$

with the same mixed target.

The **policy-matched endpoint reference** retains the pair and coefficient draws but trains on the two clean endpoints:

$$\mathcal{L}_{\mathrm{end}} = \mathbb{E}_{(i,j),\lambda}\left[\lambda \ell(f_\theta(x_i), y_i^{\mathrm{tar}}) + (1 - \lambda)\ell(f_\theta(x_j), y_j^{\mathrm{tar}})\right]. \tag{4}$$

It mixes neither inputs, representations, nor targets. Assume that every source group has positive probability and that each endpoint, conditional on its group, follows the source within-group distribution, as under uniform within-group sampling. Define the source group risk

$$R_g(f_\theta) = \mathbb{E}_{\mathcal{D}}\left[\ell(f_\theta(X), Y^{\mathrm{tar}}) \mid G = g\right].$$

For a pointwise predictor, meaning that the prediction for one example does not depend on the other examples in its minibatch, independence of λ and $\mathbb{E}[\lambda] = 1/2$ give

$$\mathcal{L}_{\mathrm{end}} = \sum_{g \in \mathcal{G}} E_g^{\mathrm{pos}} R_g(f_\theta). \tag{5}$$

Under these population assumptions, pooled exposure determines the clean-endpoint objective.

The **marginal-matched endpoint reference** independently draws both endpoint groups from $\widehat{E}^{\mathrm{pos}}$, the originating policy's finite-run estimate of pooled exposure. It then samples examples uniformly within each group and equally weights the two clean losses. Under the same population assumptions, for a fixed target $\widehat{E}^{\mathrm{pos}}$, its objective is

$$\mathcal{L}_{\mathrm{marg}}(f_\theta \mid \widehat{E}^{\mathrm{pos}}) = \sum_{g \in \mathcal{G}} \widehat{E}_g^{\mathrm{pos}} R_g(f_\theta). \tag{6}$$

When $\widehat{E}^{\mathrm{pos}} = E^{\mathrm{pos}}$, this equals Equation (5). Policy matching retains position-specific marginals, pair coupling, and coefficient weights. Marginal matching uses independent groups from the pooled distribution and equal endpoint weights. We use both procedures to measure empirical sensitivity to these training choices at a fixed target pooled exposure. Empirical policy-matched runs preserve the ordered-pair draws, coefficient draws, and optimizer-update count. For B pairs, input Mixup evaluates B mixed inputs, whereas its reference evaluates $2B$ clean inputs. Feature-space comparisons use two backbone passes in both procedures. With separate endpoint passes, $E^{(i)}$ and $E^{(j)}$ describe the group composition seen by each pass. Minibatch composition, gradient covariance, and normalization-state updates therefore remain properties of the training procedure beyond the additive population identity. Appendix D gives the implementation details.

2.2 Direct and complementary coverage

Let $T \in \{\mathrm{SS}, \mathrm{SD}, \mathrm{DS}, \mathrm{DD}\}$ denote pair topology (Figure 2), with $p_\tau = \mathbb{P}(T = \tau)$. Define

$$\mathrm{DPE}_{\mathrm{inv}} = p_{\mathrm{SD}}, \qquad \mathrm{DPE}_{\mathrm{sens}} = p_{\mathrm{DS}}, \qquad \mathrm{DPE} = p_{\mathrm{SD}} + p_{\mathrm{DS}}. \tag{7}$$

The invariant component SD varies the attribute at a fixed label; the sensitivity-oriented component DS varies the label at a fixed attribute. DPE is therefore the probability that exactly one metadata variable differs. The polarity $S = YA$ indicates label–attribute agreement $(+1)$ or disagreement (-1). For $\tau \in \{\mathrm{SS}, \mathrm{DD}\}$ and $s \in \{-1, +1\}$, let $m_{\tau,s} = \mathbb{P}(T = \tau, S_i = S_j = s)$ and $n_s = m_{\mathrm{SS},s} + m_{\mathrm{DD},s}$. Pooled complementary coverage is

$$C_{\mathrm{coll}} = 2 \min(n_+, n_-), \tag{8}$$

the balanced SS/DD mass available across polarities. Let $e_- = \frac{1}{2}[\mathbb{P}(S_i = -1) + \mathbb{P}(S_j = -1)]$ be the negative-polarity endpoint-position mass.

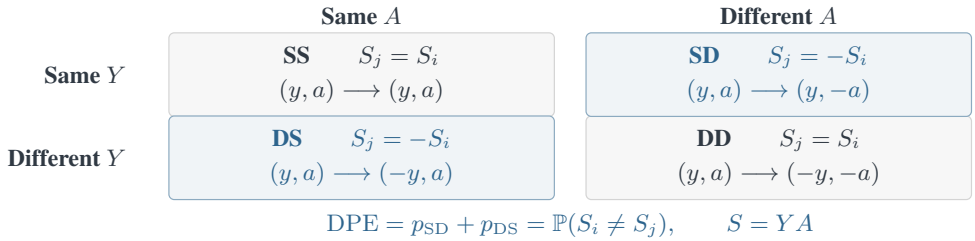

Figure 2: **Pair topology and polarity.** Rows and columns specify label and attribute relations. Blue off-diagonal cells flip polarity and contribute to DPE; diagonal cells preserve polarity. SS/DD provide complementary coverage only when both polarities are represented across pairs. Arrows map groups, not individual examples.

Proposition 1 (Coverage identity)**.** *For every ordered pair distribution over binary Y, A,*

$$\mathrm{DPE} = \mathbb{P}(S_i \neq S_j), \qquad e_- = \frac{\mathrm{DPE}}{2} + n_-, \tag{9}$$

and

$$\mathrm{DPE} + C_{\mathrm{coll}} = 2\min(e_-, 1 - e_-). \tag{10}$$

DPE measures cross-polarity pairs, and complementary coverage measures balanced same-polarity support; their sum is exactly endpoint-polarity balance. Equal DPE need not imply equal endpoint exposure, and equal polarity balance can arise from different allocations of direct and complementary support. Polarity balance is also coarser than four-group balance.

2.3 The fixed-marginal coverage budget

For position-specific masses $\eta_i = \mathbb{P}(S_i = -1)$ and $\eta_j = \mathbb{P}(S_j = -1)$, every binary coupling satisfies

$$|\eta_i - \eta_j| \leq \mathrm{DPE} \leq \min\{\eta_i + \eta_j,\, 2 - \eta_i - \eta_j\}. \tag{11}$$

If both positions retain the source polarity marginal $\rho = \mathbb{P}_{\mathcal{D}}(S = +1)$, then

$$\mathrm{DPE} \leq 2\min\{\rho, 1 - \rho\}. \tag{12}$$

For $\rho \geq 1/2$, the bound is $2(1 - \rho)$. Exceeding this budget requires changing at least one endpoint polarity marginal, not merely rearranging pairs. To illustrate, let $\pi_{y,a} = \mathbb{P}_{\mathcal{D}}(Y = y, A = a)$, $\pi_Y(y) = \sum_a \pi_{y,a}$, and $\pi_A(a) = \sum_y \pi_{y,a}$. With source-distributed anchors and all requested partner groups available, only-SD maps (y, a) to $(y, -a)$, while only-DS maps it to $(-y, a)$. Hence

$$E^{\mathrm{pos,SD}}_{y,a} = \frac{1}{2}(\pi_{y,a} + \pi_{y,-a}) = \frac{\pi_Y(y)}{2}, \tag{13}$$

$$E^{\mathrm{pos,DS}}_{y,a} = \frac{1}{2}(\pi_{y,a} + \pi_{-y,a}) = \frac{\pi_A(a)}{2}. \tag{14}$$

Only-SD balances attributes within labels while retaining the source label marginal; only-DS balances labels within attributes while retaining the source attribute marginal. Both have DPE 1, but generally different endpoint distributions.

2.4 A reporting decomposition

For policy P, let $A_{\mathrm{mix}}(P)$ and $A_{\mathrm{end}}(P)$ denote mean WGA under Mixup and its policy-matched endpoint reference, and define the residual $\Delta(P) = A_{\mathrm{mix}}(P) - A_{\mathrm{end}}(P)$. For policies P, Q,

$$A_{\mathrm{mix}}(P) - A_{\mathrm{mix}}(Q) = A_{\mathrm{end}}(P) - A_{\mathrm{end}}(Q) + \Delta(P) - \Delta(Q). \tag{15}$$

Equation (15) expresses the observed Mixup policy contrast as the sum of a policy-matched endpoint contrast and a differential Mixup–endpoint residual. The endpoint term includes the finite-run effects of the retained sampler in addition to its pooled exposure. We use this identity as a descriptive decomposition and compute all terms within seed before aggregation. Section 4 introduces matched-marginal topology controls and interpolation-space comparisons.

3 ZERO-DPE IDENTIFICATION OF SIGNAL COEFFICIENTS

A linear population model separates signal identification from the curvature induced by pair topology. Let $X \in \mathbb{R}^d$, $d \geq 2$, follow

$$X = \mu_y Y u + \mu_a A v + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_d), \tag{16}$$

where u, v are orthonormal core and spurious directions, $\mu_y, \mu_a > 0$, $\sigma \geq 0$, and I_d is the $d \times d$ identity matrix. Noise is independent of (Y, A); endpoint observations are independent conditional on their selected groups, and λ is independent of both endpoints.

For $f_w(x) = w^\top x$, define $r = \mu_y w^\top u - 1$ and $t = \mu_a w^\top v$. The target signal coefficients are $r = t = 0$, equivalently $w^\top u = 1/\mu_y$ and $w^\top v = 0$. For $\lambda \sim \mathrm{Beta}(\alpha, \alpha)$, write $\kappa_\alpha = \mathbb{E}[(2\lambda - 1)^2] = (2\alpha + 1)^{-1}$ and $\beta_\alpha = \mathbb{E}[\lambda^2 + (1 - \lambda)^2] = (\alpha + 1)/(2\alpha + 1)$.

3.1 RANDOM PAIRING AND THE DIRECT-COVERAGE BUDGET

Independent random endpoints with source positive-polarity mass $\rho \in [1/2, 1)$ have $\mathrm{DPE} = 2\rho(1 - \rho)$, $n_+ = \rho^2$, $n_- = (1 - \rho)^2$, and $C_{\mathrm{coll}} = 2(1 - \rho)^2$. As $\rho \to 1$, direct coverage decreases linearly and complementary coverage quadratically in $1 - \rho$. Random pairing uses the fraction $2\rho(1 - \rho)/[2(1 - \rho)] = \rho$ of its fixed-marginal DPE budget: 98% at $\rho = 0.98$. Its low absolute DPE therefore primarily reflects scarce minority-polarity endpoint mass.

3.2 TOPOLOGY CONSTRAINTS AND IDENTIFICATION

The noiseless conditional squared-loss risks are

$$\mathcal{R}_{\mathrm{SD}} = r^2 + \kappa_\alpha t^2, \qquad \mathcal{R}_{\mathrm{DS}} = \kappa_\alpha r^2 + t^2, \tag{17}$$

$$\mathcal{R}_{\mathrm{SS},s} = (r + st)^2, \qquad \mathcal{R}_{\mathrm{DD},s} = \kappa_\alpha (r + st)^2, \quad s \in \{-1, +1\}. \tag{18}$$

Independent isotropic endpoint noise adds $\beta_\alpha \sigma^2 \|w\|^2$ to every topology. SD and DS place different weights on r and t. Positive- and negative-polarity SS/DD support constrain $r + t$ and $r - t$, respectively, jointly implying $r = t = 0$. Let $B_+ = m_{\mathrm{SS},+} + \kappa_\alpha m_{\mathrm{DD},+}$ and $B_- = m_{\mathrm{SS},-} + \kappa_\alpha m_{\mathrm{DD},-}$.

Proposition 2 (Polarity support and noiseless signal identification)**.** *Let H_{mix} and H_{end} be the quadratic-form matrices of the noiseless Mixup and endpoint signal risks in coordinates (r, t). For every finite $\alpha > 0$,*

$$H_{\mathrm{end}} = \begin{pmatrix} 1 & 1 - 2e_- \\ 1 - 2e_- & 1 \end{pmatrix}, \qquad \kappa_\alpha H_{\mathrm{end}} \preceq H_{\mathrm{mix}} \preceq H_{\mathrm{end}}, \tag{19}$$

where $U \preceq V$ means $V - U$ is positive semidefinite. Both signal risks have the unique coefficient minimizer $(r, t) = (0, 0)$ if and only if

$$0 < e_- < 1, \qquad \textit{equivalently} \qquad \mathrm{DPE} + C_{\mathrm{coll}} > 0. \tag{20}$$

At zero DPE, the eigenvalues of H_{mix} are $2B_+$ and $2B_-$.

Endpoint-polarity support determines noiseless identifiability; topology changes curvature. SD/DS supply both polarities within pairs, whereas zero-DPE SS/DD support can supply them across pairs. The proposition concerns only the two signal coefficients: directions orthogonal to u, v do not enter the noiseless risk.

3.3 CONTROLLED ALLOCATION OF DIRECT AND COMPLEMENTARY SUPPORT

Let $\delta \in [0, 1]$ be the direct-pair probability. Direct draws select SD or DS equally; non-DPE draws select SS or DD equally and use negative polarity with probability $q \in [0, 1/2]$. Within each selected anchor polarity, both label signs are equally likely. Then

$$\mathrm{DPE} = \delta, \qquad C_{\mathrm{coll}} = 2(1 - \delta)q, \qquad e_- = \frac{\delta}{2} + (1 - \delta)q. \tag{21}$$

The group-position exposures are $E^{\mathrm{pos}}_{++} = E^{\mathrm{pos}}_{--} = (1 - e_-)/2$ and $E^{\mathrm{pos}}_{+-} = E^{\mathrm{pos}}_{-+} = e_-/2$, where signs index (Y, A). The population Mixup squared loss is

$$\mathcal{R}_{\mathrm{mix}} = \beta_\alpha \left[r^2 + t^2 + 2(1 - 2e_-)rt + \sigma^2 \|w\|^2 \right]. \tag{22}$$

The bracketed expression is the endpoint squared loss. Thus, within this family, both unregularized population minimizers depend on e_-, not separately on δ, q. At $q = 1/2$, every endpoint group has exposure $1/4$, and the population objective is unchanged as DPE varies. Experiments use this path with binary cross-entropy rather than squared loss. For this balanced family with $\sigma > 0$, the noisy optimum has $w^\top u = \mu_y/(\mu_y^2 + \sigma^2)$ and $w^\top v = 0$: it is core-only but shrunk relative to the noiseless target. Appendix C.5 gives the solution for general endpoint-polarity exposure.

3.4 Independent random-pairing equivalence

For the endpoint residual $e = f_w(X) - Y$ and independent copies e_i, e_j,

$$\mathbb{E}[(\lambda e_i + (1 - \lambda)e_j)^2] = \beta_\alpha \mathbb{E}[e^2] + (1 - \beta_\alpha)(\mathbb{E}[e])^2. \tag{23}$$

If $\mathbb{E}[Y] = \mathbb{E}[A] = 0$, then $\mathbb{E}[e] = 0$ for the linear predictor, and independent random Mixup and endpoint training share their unregularized population squared-loss minimizers in this setting. Appendix C provides the full constraint matrix, conditioning results, and population WGA formula.

4 Experimental design

We vary endpoint exposure, pair topology, and interpolation space using the endpoint references in Section 2. Implementation details appear in Appendix D.

4.1 Pairing policies and controls

Vanilla Mixup uses empirical-distribution anchors and ordinary random pairing. CS-Mix mixes SD, DS, and random branches with probabilities $(p_{\mathrm{inv}}, p_{\mathrm{sens}}, p_{\mathrm{rand}})$; the selective relations follow LISA (Yao et al., 2022). Our empirical-anchor configurations are inv08 $= (0.8, 0, 0.2)$, sens08 $= (0, 0.8, 0.2)$, only-inv $= (1, 0, 0)$, and only-sens $= (0, 1, 0)$. Analyses use achieved pair frequencies, not requested branch probabilities.

Group-balanced sampling trains on clean examples with uniformly sampled groups. Group-balanced Mixup independently samples both endpoint groups uniformly, then samples examples within groups; its policy-matched reference is the group-balanced endpoint method. GB-SS, GB-SD, GB-DS, and GB-DD instead use uniform-group anchors and the specified topology. Each topology map is a permutation, so all four have uniform group-position marginals. We compare GB-SD with GB-SS and GB-DS with GB-DD to retain both group exposure and the same-label or different-label relation.

High-DPE uses group-balanced anchors, selecting SD/DS equally; balanced Low-DPE equally selects SS/DD. Empirical-support Low-DPE restricts pairs to SS/DD, retaining source polarity imbalance. The synthetic family (Equation (21)) varies δ and the non-DPE negative-polarity probability q. For CelebA and Waterbirds marginal-matched empirical-policy comparisons, endpoint groups are independently drawn from $\widehat{E}^{\mathrm{pos}}$, the finite-run estimate of the originating policy's pooled exposure.

4.2 Datasets and training protocol

Synthetic data follows Equation (16), with source probabilities $\pi_{+1,+1} = \pi_{-1,-1} = \rho/2$ and $\pi_{+1,-1} = \pi_{-1,+1} = (1 - \rho)/2$. Labels and attributes are individually balanced, with agreement probability ρ and correlation $\mathrm{Corr}(Y, A) = 2\rho - 1$. We train affine classifiers $w^\top x + b$ at $\rho \in \{0.70, 0.80, 0.90, 0.95, 0.98\}$. The analysis uses zero intercept; Appendix C.6 gives group accuracy with an intercept. **Colored MNIST** uses digit parity and color at $\rho = 0.95$, a two-layer CNN, and input Mixup (LeCun et al., 1998; Arjovsky et al., 2019). **Biased CIFAR** uses animals versus vehicles and a binary image-patch attribute at $\rho = 0.95$, a four-layer CNN, and input or penultimate-feature Mixup (Krizhevsky, 2009). **Waterbirds** uses bird type and background (Sagawa et al., 2020); **CelebA** uses Blond Hair and Male (Liu et al., 2015). Both use ImageNet-pretrained ResNet-18 and penultimate-feature Mixup (Deng et al., 2009; He et al., 2016).

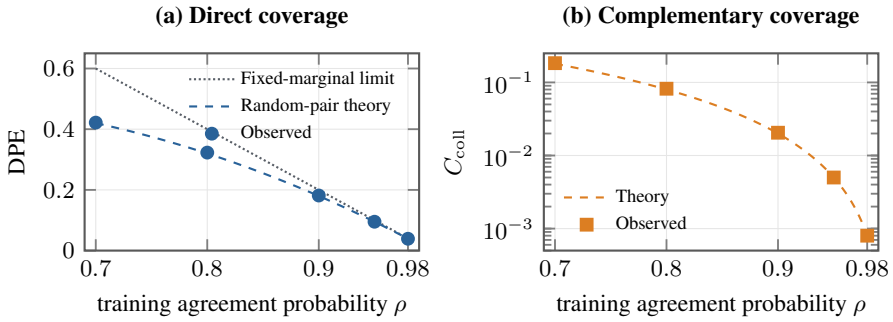

Figure 3: **Random pairing under fixed endpoint marginals.** Curves show population values at the nominal training agreement probability, and markers show means over 20 seeds. As ρ increases, random pairing uses an increasing fraction of the DPE permitted by its fixed polarity marginals, while both the absolute DPE budget and complementary coverage shrink.

All empirical models use binary cross-entropy with logits; main Mixup runs use $\lambda \sim \text{Beta}(0.4, 0.4)$. Controlled validation/test sets are group-balanced, whereas real benchmarks use standard splits. Synthetic and Colored MNIST use final checkpoints; biased CIFAR, Waterbirds, and CelebA select by validation WGA. Group-aware sampling and WGA-based selection use training and validation group metadata, respectively.

We report held-out WGA throughout, computing WGA within seed before aggregation. Table 15 additionally reports Avg and GBA for CelebA. Controlled datasets also measure swap sensitivity. For each held-out x_n, x_n^{swap} changes only the designated spurious attribute while retaining label-defining content and all other available input components: the signed spurious component, digit color, or patch assignment. For N_{eval} held-out examples,

$$\text{Swap} = \frac{100}{N_{\text{eval}}} \sum_{n=1}^{N_{\text{eval}}} \mathbf{1}[\widehat{Y}(x_n^{\text{swap}}) \neq \widehat{Y}(x_n)]. \tag{24}$$

Thus Swap is the percentage of predictions changed by the intervention. Synthetic experiments use 20 seeds, Colored MNIST uses 10, and main natural-image comparisons use five, except Waterbirds GB-SD versus GB-SS, which uses ten matched seeds. Absolute accuracies are reported as means with sample standard deviations. Matched differences are computed within seed and summarized by two-sided 95% Student's t confidence intervals. The intervals are descriptive rather than equivalence tests, and no multiple-comparison correction is applied.

5 Results

5.1 Random pairing and the fixed-marginal coverage budget

Figure 3 compares achieved coverage with the population values at the nominal source agreement probability. At $\rho = 0.98$, observed DPE is 0.0390, close to the nominal fixed-marginal limit 0.0400, and population random pairing uses 98% of that limit. Random-pair Mixup reaches $23.65 \pm 2.31\%$ WGA, compared with $37.69 \pm 2.89\%$ for empirical risk minimization (ERM). Group-balanced sampling and group-balanced Mixup reach $89.69 \pm 1.12\%$ and $90.27 \pm 0.72\%$, respectively (Table 5). Group balancing changes endpoint-polarity mass and yields substantially higher WGA in this setting. The following experiments vary direct and complementary support separately.

5.2 Robustness with zero direct pair exposure

At $\rho = 0.98$ and zero DPE, increasing non-DPE negative-polarity support from $q = 0$ to $q = 1/2$ raises synthetic Mixup WGA from 3.07% to 90.41%, and endpoint WGA from 8.17% to 89.83% (Figure 4b). Balanced SS/DD pairs alone attain high WGA in this experiment. With balanced endpoints, Mixup WGA remains between 90.34% and 90.59% as DPE varies from zero to one, and endpoint WGA follows the same stable pattern (Figure 4c). The settings $(\delta, q) = (0, 0.25)$ and

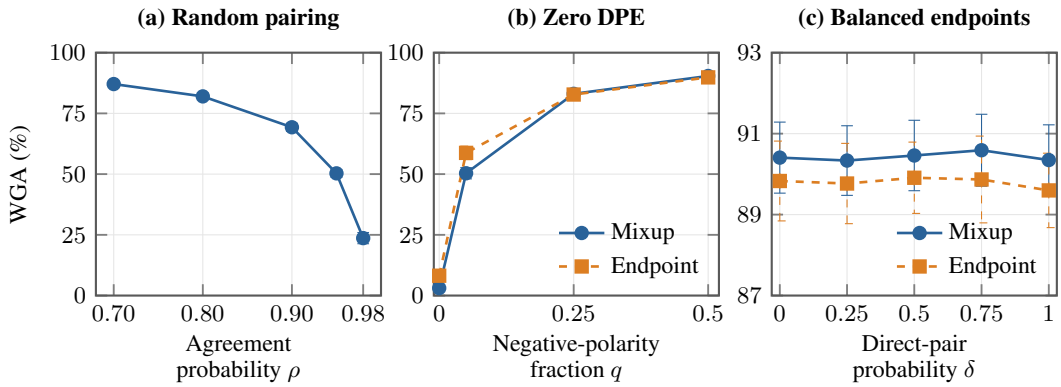

Figure 4: **Synthetic WGA across the controlled allocation family.** Panel (a) varies the source agreement probability. Panel (b) fixes $\delta = 0$ and adds negative-polarity SS/DD support. Panel (c) balances all endpoint groups and varies DPE from zero to one, with the vertical axis restricted to 87–93%. Error bars show sample standard deviations over 20 seeds.

Table 1: **Paired decomposition of policy contrasts.** Each cell reports a mean WGA difference in percentage points, with its two-sided 95% Student's t confidence interval below. Positive values favor the first policy. The endpoint column uses policy-matched references. Waterbirds GB-SD versus GB-SS uses ten matched seeds; the remaining comparisons use five.

Dataset	Contrast	Mixup	Endpoint	Differential residual
CelebA	only-sens vs. only-inv	+12.24 [+7.38, +17.10]	+19.31 [+13.78, +24.84]	−7.07 [−15.47, +1.34]
	sens08 vs. inv08	+11.95 [+3.95, +19.95]	+16.06 [+6.88, +25.23]	−4.11 [−9.94, +1.72]
Waterbirds	GB-SD vs. GB-SS	+2.53 [+0.15, +4.91]	+1.83 [+0.07, +3.58]	+0.70 [−2.54, +3.95]
	GB-DS vs. GB-DD	+0.61 [−0.56, +1.77]	+1.30 [−0.79, +3.39]	−0.69 [−3.60, +2.22]
	inv08 vs. sens08	+4.44 [−1.25, +10.12]	+5.40 [+0.15, +10.66]	−0.97 [−6.69, +4.75]

$(0.5, 0)$ both have $e_- \approx 0.25$, while allocating their support through complementary and direct pairs, respectively. Their endpoint WGA values are 82.77% and 82.94%. Colored MNIST shows the same qualitative pattern (Tables 10 and 11). At zero DPE, restoring balanced polarity support raises WGA from $0.00 \pm 0.00\%$ to $92.38 \pm 2.50\%$; the matched endpoint reference reaches $90.97 \pm 2.69\%$. The support intervention changes mean WGA by 92.38 points, while the Mixup–endpoint difference at balanced support is 1.41 points. Balanced SS/DD training attains high WGA at zero DPE in both controlled settings.

5.3 Policy contrasts and endpoint references

Table 1 applies the within-seed decomposition in Equation (15). The Mixup and policy-matched endpoint point estimates have the same direction in all five comparisons. On CelebA, the endpoint contrasts are 19.31 points for only-sens versus only-inv and 16.06 points for sens08 versus inv08, compared with Mixup contrasts of 12.24 and 11.95 points. On Waterbirds, GB-SD versus GB-SS fixes uniform group exposure and the same-label relation: its Mixup and endpoint contrasts are 2.53 and 1.83 points, respectively. The five differential-residual point estimates range from −7.07 to +0.70 points, and their 95% confidence intervals all include zero.

The CelebA exposure measurements in Table 17 show why equal DPE need not imply equal training exposure. Only-inv and only-sens both have DPE 1, but their blond-male endpoint exposures

Table 2: **Absolute WGA under Mixup and endpoint training.** Percentages are means ± sample standard deviations over five seeds. Mixup uses penultimate-feature interpolation.

Dataset	Policy	Mixup	Policy-matched endpoint	Marginal-matched endpoint
CelebA	only-inv	75.78 ± 2.03	68.14 ± 2.27	61.78 ± 3.37
	only-sens	88.02 ± 2.60	87.44 ± 2.56	74.00 ± 3.67
	inv08	73.67 ± 4.24	70.78 ± 7.61	64.00 ± 7.04
	sens08	85.61 ± 2.61	86.83 ± 3.24	71.11 ± 4.63
Waterbirds	inv08	78.05 ± 4.30	75.67 ± 1.21	76.17 ± 2.95
	sens08	73.61 ± 3.86	70.27 ± 3.93	74.45 ± 1.77

are 0.0745 and 0.2097, respectively. Only-inv retains the source label marginal, whereas only-sens balances labels within each attribute. Their Mixup WGA values are $75.78 \pm 2.03\%$ and $88.02 \pm 2.60\%$, with average accuracies of 94.63% and 91.36%, respectively (Table 15).

Table 2 complements the paired contrasts in Table 1. Across all three procedures, mean WGA favors sensitivity-oriented policies on CelebA and inv08 over sens08 on Waterbirds. The endpoint implementations differ in absolute performance: policy-matched versus marginal-matched WGA is 87.44% versus 74.00% for CelebA only-sens, and 70.27% versus 74.45% for Waterbirds sens08.

5.4 Interpolation space, label relation, and endpoint support

Interpolation space. On biased CIFAR, vanilla feature Mixup reaches $71.10 \pm 1.74\%$ WGA, compared with $63.57 \pm 2.29\%$ for input Mixup, giving a 7.53-point difference under the same pair statistics (Table 12). For the matched-topology comparison, GB-DS exceeds GB-DD by 0.70 points in input space and 0.36 points in feature space (Table 9). The larger vanilla-Mixup contrast is associated with interpolation space rather than a change in pair coverage.

Label relation. On Colored MNIST, GB-DS and GB-DD achieve $93.82 \pm 1.12\%$ and $93.48 \pm 1.76\%$ WGA, respectively (Table 9). Both use different-label pairs, while only GB-DS fixes the spurious attribute within each pair. GB-SD reaches $90.61 \pm 2.62\%$ under Mixup and $91.25 \pm 2.04\%$ under endpoint training.

Endpoint support. On Waterbirds, balanced Low-DPE exceeds empirical-support Low-DPE by 5.79 points, with both policies having zero DPE (Table 13). Balanced Low-DPE and High-DPE exceed their endpoint references by 3.40 and 1.51 points, respectively. Taken together, these comparisons show substantial variation across endpoint distributions, interpolation spaces, and pair relations.

6 Discussion and conclusion

Selective Mixup changes both endpoint sampling and interpolation. Under the population assumptions of Equation (5), pooled group-position exposure determines the clean-endpoint objective. The binary coverage identity further separates cross-polarity DPE from complementary SS/DD support. In the linear squared-loss model, positive DPE provides sufficient polarity support, but the same signal-identification condition can also be satisfied by SS/DD pairs that represent both polarities. Balanced SS/DD sampling attains high WGA at zero DPE on synthetic data and Colored MNIST. On Waterbirds and CelebA, policy contrasts also appear under clean-endpoint training, while the Mixup–endpoint differences vary across pair relations and interpolation spaces. These results show that endpoint sampling is an important component of selective Mixup and that DPE alone does not characterize the training distribution. The theoretical conclusions concern the stated binary population model, and the empirical endpoint references retain different finite-training implementations. Appendix G describes this scope in detail. We recommend reporting endpoint exposure together with pair topology, evaluating a clean-endpoint reference derived from the sampler, and specifying the interpolation layer. These controls distinguish robustness differences already visible under endpoint training from those observed after interpolation is introduced.

AI USE STATEMENT

ChatGPT was used for copyediting, LaTeX assistance, code-debugging support and graph making. Experimental observations and reported numerical results were produced by the authors' experimental pipeline and reviewed against the underlying records.

ETHICS STATEMENT

This work studies robustness to spurious correlations using synthetic data and established public benchmarks. CelebA demographic attributes are used for controlled sampling, model selection, robustness evaluation, and representation diagnostics. The study evaluates learning procedures for robustness analysis and does not develop a demographic-classification or deployment system. Dataset access and redistribution follow the applicable licenses.

REPRODUCIBILITY STATEMENT

The appendix provides analytical derivations, dataset and model descriptions, training settings, sampler definitions, endpoint objectives, checkpoint criteria, and statistical procedures. The anonymous supplement provides configurations, preprocessing and training code, per-seed records, and scripts for regenerating the reported tables and figures.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. URL https://arxiv.org/abs/1907.02893.

Luigi Carratino, Moustapha Cissé, Rodolphe Jenatton, and Jean-Philippe Vert. On mixup regularization. *Journal of Machine Learning Research*, 23(325):1–31, 2022. URL https://jmlr.org/papers/v23/20-1385.html.

Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: Rebalanced mixup. In *Computer Vision – ECCV 2020 Workshops*, volume 12540 of *Lecture Notes in Computer Science*, pp. 95–110. Springer, 2020. doi: 10.1007/978-3-030-65414-6_9. URL https://doi.org/10.1007/978-3-030-65414-6_9.

Elliot Creager, Joern-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2189–2200. PMLR, 2021. URL https://proceedings.mlr.press/v139/creager21a.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Hongyu Guo, Yongyi Mao, and Richong Zhang. MixUp as locally linear out-of-manifold regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3714–3722, 2019. doi: 10.1609/aaai.v33i01.33013714. URL https://doi.org/10.1609/aaai.v33i01.33013714.

Ethan Harris, Antonia Marcu, Matthew Painter, Mahesan Niranjan, Adam Prügel-Bennett, and Jonathon Hare. FMix: Enhancing mixed sample data augmentation. *arXiv preprint arXiv:2002.12047*, 2020. URL https://arxiv.org/abs/2002.12047.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 336–351. PMLR, 2022. URL https://proceedings.mlr.press/v177/idrissi22a.html.

Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5275–5285. PMLR, 2020. URL https://proceedings.mlr.press/v119/kim20b.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5815–5826. PMLR, 2021. URL https://proceedings.mlr.press/v139/krueger21a.html.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. URL https://doi.org/10.1109/5.726791.

Evan Z. Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6781–6792. PMLR, 2021. URL https://proceedings.mlr.press/v139/liu21f.html.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015. doi: 10.1109/ICCV.2015.425. URL https://doi.org/10.1109/ICCV.2015.425.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20673–20684, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/eddc3427c5d77843c2253f1e799fe933-Abstract.html.

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*, 2020. URL https://arxiv.org/abs/1911.08731.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A. Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 13811–13820, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/36ad8b5f42db492827016448975cc22d-Abstract.html.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6438–6447. PMLR, 2019. URL https://proceedings.mlr.press/v97/verma19a.html.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. URL https://authors.library.caltech.edu/records/cvm3y-5hh21.

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25407–25437. PMLR, 2022. URL https://proceedings.mlr.press/v162/yao22b.html.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6023–6032, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Yun_CutMix_Regularization_Strategy_to_Train_Strong_Classifiers_With_Localizable_Features_ICCV_2019_paper.html.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. doi: 10.1109/TPAMI.2017.2723009. URL https://doi.org/10.1109/TPAMI.2017.2723009.

A Additional related work

Interpolation and selective augmentation. Mixup trains on convex combinations of examples and targets (Zhang et al., 2018). Its extensions include representation-space interpolation (Verma et al., 2019), spatial mixing operators (Yun et al., 2019; Harris et al., 2020; Kim et al., 2020), and imbalance-aware target weighting (Chou et al., 2020). LISA uses label- and environment-conditioned selective augmentation (Yao et al., 2022). Our selective SD and DS branches use the corresponding pair relations when the environment variable is the spurious attribute. The endpoint references evaluate the training distribution induced by these pair samplers separately from interpolation.

Learning under spurious correlations. Group robustness has been studied through group-risk optimization (Sagawa et al., 2020), invariant learning (Arjovsky et al., 2019; Krueger et al., 2021), environment inference (Creager et al., 2021), and learning procedures that identify or emphasize difficult examples (Nam et al., 2020; Liu et al., 2021). Balancing and last-layer retraining provide further evidence that robustness can improve through changes to sampling or classifier fitting (Idrissi et al., 2022; Kirichenko et al., 2023). Waterbirds, CelebA, and WILDS provide widely used evaluation settings (Sagawa et al., 2020; Liu et al., 2015; Koh et al., 2021). Our controls connect these sampling-based observations to the evaluation of selective Mixup.

B Pairing identities and endpoint controls

B.1 Polarity coverage and fixed-marginal budgets

Let $M_{g,h} = \mathbb{P}(G_i = g, G_j = h)$. The position-specific and pooled group exposures are

$$E_g^{(i)} = \sum_h M_{g,h}, \qquad E_g^{(j)} = \sum_h M_{h,g}, \qquad E_g^{\mathrm{pos}} = \frac{1}{2}\left(E_g^{(i)} + E_g^{(j)}\right).$$

These quantities count endpoint participation, not the number of unique examples encountered.

For $S = YA$,

$$S_i S_j = (Y_i Y_j)(A_i A_j).$$

Thus polarity changes exactly when one, but not both, of the label and attribute changes:

$$\mathrm{DPE} = \mathbb{P}(S_i \neq S_j). \tag{25}$$

Writing

$$n_+ = \mathbb{P}(S_i = S_j = +1), \qquad n_- = \mathbb{P}(S_i = S_j = -1),$$

gives

$$\mathrm{DPE} + n_+ + n_- = 1, \qquad e_- = \frac{\mathrm{DPE}}{2} + n_-. \tag{26}$$

Every cross-polarity pair contributes one negative endpoint; every negative-polarity non-DPE pair contributes two. By contrast, the probability that a pair contains at least one negative endpoint is

$$E_-^{\mathrm{pair}} = \mathrm{DPE} + n_-. \tag{27}$$

Since $C_{\mathrm{coll}} = 2\min(n_+, n_-)$,

$$\begin{aligned} \mathrm{DPE} + C_{\mathrm{coll}} &= 1 - |n_+ - n_-| \\ &= 1 - |1 - 2e_-| = 2\min(e_-, 1 - e_-). \end{aligned} \tag{28}$$

Equivalently, with $I_{\mathrm{coll}} = |n_+ - n_-|$,

$$\mathrm{DPE} + C_{\mathrm{coll}} + I_{\mathrm{coll}} = 1.$$

This proves Proposition 1. Polarity balance is coarser than four-group balance: it does not specify how mass is divided between the two groups within each polarity.

For position-specific negative-polarity masses η_i, η_j, let $c = \mathbb{P}(S_i = S_j = -1)$. Then

$$\mathrm{DPE} = \eta_i + \eta_j - 2c, \qquad \max(0, \eta_i + \eta_j - 1) \leq c \leq \min(\eta_i, \eta_j),$$

which yields Equation (11).

Under independent random pairing with $\rho = \mathbb{P}(S = +1) \geq 1/2$,

$$\mathrm{DPE} = 2\rho(1-\rho), \qquad n_+ = \rho^2, \qquad n_- = (1-\rho)^2, \qquad C_{\mathrm{coll}} = 2(1-\rho)^2. \tag{29}$$

Hence $\mathrm{DPE} + C_{\mathrm{coll}} = 2(1-\rho)$. For $\rho < 1$, random pairing uses the fraction ρ of the largest DPE permitted by its fixed polarity marginals. These identities do not require balanced labels. Under the additional symmetric group model

$$\pi_{++} = \pi_{--} = \frac{\rho}{2}, \qquad \pi_{+-} = \pi_{-+} = \frac{1-\rho}{2},$$

the direct components satisfy

$$p_{\mathrm{SD}} = p_{\mathrm{DS}} = \rho(1-\rho).$$

B.2 Alternative coverage summaries

The pooled statistic allows SS and DD support to complement one another. A stricter topology-separated version is

$$\begin{aligned} C_{\mathrm{coll}}^{\mathrm{sep}} &= 2\min(m_{\mathrm{SS},+}, m_{\mathrm{SS},-}) \\ &\quad + 2\min(m_{\mathrm{DD},+}, m_{\mathrm{DD},-}) \leq C_{\mathrm{coll}}. \end{aligned} \tag{30}$$

The inequality follows from $\min(a,b) + \min(c,d) \leq \min(a+c, b+d)$. For CelebA, inv08 has pooled/separated coverage $0.0806/0.0784$, and sens08 has $0.0803/0.0782$.

A coefficient-weighted diagnostic is

$$\mathrm{wDPE} = \mathbb{E}[4\lambda(1-\lambda)\mathbf{1}[\mathrm{SD \text{ or } DS}]]\,. \tag{31}$$

For an independent $\lambda \sim \mathrm{Beta}(\alpha, \alpha)$,

$$\mathrm{wDPE} = \frac{2\alpha}{2\alpha+1}\mathrm{DPE}.$$

The factor $4\lambda(1-\lambda)$ summarizes interpolation centrality; the full quadratic constraint strength is characterized by the matrices in Appendix C.

B.3 Selective and balanced samplers

For each anchor, CS-Mix requests SD, DS, or random pairing with probabilities $(p_{\mathrm{inv}}, p_{\mathrm{sens}}, p_{\mathrm{rand}})$. Partners are sampled uniformly from the requested candidate set. The selective operations coincide with LISA-style pair construction when the environment variable is the spurious attribute (Yao et al., 2022).

Using the signed coding $y, a \in \{-1, +1\}$, the topology maps are

$$\begin{aligned} F_{\mathrm{SS}}(y,a) &= (y,a), \qquad F_{\mathrm{SD}}(y,a) = (y,-a), \\ F_{\mathrm{DS}}(y,a) &= (-y,a), \quad F_{\mathrm{DD}}(y,a) = (-y,-a). \end{aligned} \tag{32}$$

A balanced single-topology policy has

$$M_{g,h}^{\mathrm{GB}\text{-}\tau} = \frac{1}{4}\mathbf{1}[h = F_\tau(g)]. \tag{33}$$

All four maps are permutations, so every policy has $E_g^{\mathrm{pos}} = 1/4$. GB-SS and GB-DD have $(\mathrm{DPE}, C_{\mathrm{coll}}) = (0,1)$; GB-SD and GB-DS have $(1,0)$. Independent uniform-group pairing has $M_{g,h} = 1/16$, with $p_{\mathrm{SD}} = p_{\mathrm{DS}} = 1/4$ and $C_{\mathrm{coll}} = 1/2$.

High-DPE uses group-balanced anchors and selects SD and DS equally. Balanced Low-DPE uses group-balanced anchors and selects SS and DD equally. Empirical-support Low-DPE retains the source polarity imbalance while restricting pairs to SS or DD.

Write

$$\pi_{y,a} = \mathbb{P}_{\mathcal{D}}(Y = y, A = a), \qquad \pi_Y(y) = \sum_{a \in \{-1,+1\}} \pi_{y,a}, \qquad \pi_A(a) = \sum_{y \in \{-1,+1\}} \pi_{y,a}.$$

With source-distributed anchors and no fallback, only-SD and only-DS induce

$$E_{y,a}^{\mathrm{pos,SD}} = \frac{1}{2}\big(\pi_{y,a} + \pi_{y,-a}\big) = \frac{\pi_Y(y)}{2}, \qquad E_{y,a}^{\mathrm{pos,DS}} = \frac{1}{2}\big(\pi_{y,a} + \pi_{-y,a}\big) = \frac{\pi_A(a)}{2}. \tag{34}$$

Thus equal DPE does not imply equal four-group exposure.

Fallback. Let T_{achieved} be the achieved pair topology, and let $k \in \{\text{SS}, \text{SD}, \text{DS}, \text{DD}\}$. For a requested selective topology $\tau \in \{\text{SD}, \text{DS}\}$, let r_τ be the fallback probability and $f_{k|\tau}$ the probability of achieving topology k conditional on that fallback. Let p_k^R be the achieved-topology probability in the ordinary random branch. Then

$$\mathbb{P}(T_{\text{achieved}} = k) = \sum_{\tau \in \{\text{SD},\text{DS}\}} p_\tau^{\text{req}} \left[(1 - r_\tau)\mathbf{1}[k = \tau] + r_\tau f_{k|\tau}\right] + p_{\text{rand}} p_k^R, \tag{35}$$

where $p_{\text{SD}}^{\text{req}} = p_{\text{inv}}$ and $p_{\text{DS}}^{\text{req}} = p_{\text{sens}}$. Terms multiplied by zero fallback probability are set to zero. Fallback need not match the ordinary random branch: if an SD request fails because no SD partner exists in the candidate pool, resampling from that pool gives $f_{\text{SD}|\text{SD}} = 0$.

Without fallback,

$$\text{DPE}_{\text{inv}} = p_{\text{inv}} + p_{\text{rand}} p_{\text{SD}}^R, \qquad \text{DPE}_{\text{sens}} = p_{\text{sens}} + p_{\text{rand}} p_{\text{DS}}^R.$$

If selective branches contribute no non-DPE mass, then

$$C_{\text{coll}}^{\text{CS,pop}} = p_{\text{rand}} C_{\text{coll}}^{\text{R,pop}},$$

where the superscripts denote population quantities for CS-Mix and random pairing, respectively. All experiments use achieved counts rather than nominal branch probabilities.

B.4 Endpoint objectives

For a minibatch of B ordered pairs, the policy-matched endpoint reference minimizes

$$\widehat{\mathcal{L}}_{\text{end}} = \frac{1}{B} \sum_{b=1}^{B} \left[\lambda_b \ell\big(f_\theta(x_{i_b}), y_{i_b}^{\text{tar}}\big) + (1 - \lambda_b)\ell\big(f_\theta(x_{j_b}), y_{j_b}^{\text{tar}}\big)\right]. \tag{36}$$

Both terms are backpropagated in one optimizer update. Matched Mixup and endpoint runs regenerate the same seeded anchor, partner, and coefficient sequences. Statistics attached to endpoint rows describe this retained sampler.

For policies that preserve the within-group example distributions, including uniform within-group sampling, an independent symmetric coefficient gives

$$\begin{aligned}\mathcal{L}_{\text{end}} &= \frac{1}{2} \sum_g E_g^{(i)} R_g(f) + \frac{1}{2} \sum_g E_g^{(j)} R_g(f) \\ &= \sum_g E_g^{\text{pos}} R_g(f).\end{aligned} \tag{37}$$

This identity characterizes the additive population objective. At finite sample, the implementation also determines minibatch composition, gradient covariance, and normalization-state updates.

The marginal-matched reference instead draws

$$G_i, G_j \overset{\text{i.i.d.}}{\sim} \text{Categorical}(\widehat{E}^{\text{pos}}) \tag{38}$$

and uniformly samples an example from each selected group. Its loss is

$$\widehat{\mathcal{L}}_{\text{marg}} = \frac{1}{2B} \sum_{b=1}^{B} \left[\ell\big(f_\theta(x_{i_b}), y_{i_b}^{\text{tar}}\big) + \ell\big(f_\theta(x_{j_b}), y_{j_b}^{\text{tar}}\big)\right]. \tag{39}$$

This reference targets the originating policy's pooled group exposure using independently sampled endpoint groups and equal endpoint weights. The reported finite-run exposure is computed from the realized draws.

B.5 Finite-run coverage computation

For run s, let T_s be the total number of ordered-pair occurrences over the scheduled training run. Let $N_{g,h,s}$ count occurrences with anchor group g and partner group h. Define

$$\widehat{M}_{g,h,s} = \frac{N_{g,h,s}}{T_s}, \qquad \widehat{E}_{g,s}^{\text{pos}} = \frac{1}{2}\left(\sum_h \widehat{M}_{g,h,s} + \sum_h \widehat{M}_{h,g,s}\right).$$

For $\tau \in \{\mathrm{SS}, \mathrm{DD}\}$ and $\xi \in \{-1, +1\}$, let $N_{\tau,\xi,s}$ count occurrences of topology τ with both endpoint polarities equal to ξ, and set

$$\widehat{m}_{\tau,\xi,s} = \frac{N_{\tau,\xi,s}}{T_s}.$$

Also define

$$\widehat{n}_{\xi,s} = \widehat{m}_{\mathrm{SS},\xi,s} + \widehat{m}_{\mathrm{DD},\xi,s}, \qquad \widehat{e}_{-,s} = \sum_{(y,a):\, ya=-1} \widehat{E}^{\mathrm{pos}}_{(y,a),s}.$$

The empirical DPE is the fraction of pair occurrences with topology SD or DS. The logged statistic `C_coll_pool` is

$$\widehat{C}_{\mathrm{coll},s} = 2\min\big(\widehat{n}_{+,s}, \widehat{n}_{-,s}\big). \tag{40}$$

The denominator includes all pair occurrences, including direct pairs, repeated pairs, and self-pairs. If each of L epochs contains exactly N anchor occurrences and no final minibatch is dropped, then $T_s = LN$. Otherwise, T_s is the actual number of processed pair occurrences.

Reported coverage summarizes the full scheduled sampler, including runs for which validation selects an earlier checkpoint. All nonlinear summaries are computed within run before averaging. Every run satisfies

$$0 \le \widehat{C}_{\mathrm{coll},s} \le 1 - \widehat{\mathrm{DPE}}_s$$

and

$$\widehat{\mathrm{DPE}}_s + \widehat{C}_{\mathrm{coll},s} = 2\min(\widehat{e}_{-,s}, 1 - \widehat{e}_{-,s}).$$

For K runs, writing an overbar for the arithmetic mean gives

$$\overline{\widehat{\mathrm{DPE}}} + \overline{\widehat{C}_{\mathrm{coll}}} = 1 - \frac{2}{K}\sum_{s=1}^{K}\left|\widehat{e}_{-,s} - \frac{1}{2}\right|.$$

Consequently, applying the minimum to the mean exposure need not reproduce the mean coverage.

C Squared-loss derivations

C.1 Coefficient moments and topology risks

For $\lambda \sim \mathrm{Beta}(\alpha, \alpha)$,

$$\mathbb{E}[\lambda] = \frac{1}{2}, \qquad \mathbb{E}[\lambda^2] = \frac{\alpha+1}{2(2\alpha+1)}.$$

Consequently,

$$\kappa_\alpha = \frac{1}{2\alpha+1}, \qquad \beta_\alpha = \frac{\alpha+1}{2\alpha+1}, \qquad \mathbb{E}[4\lambda(1-\lambda)] = \frac{2\alpha}{2\alpha+1}. \tag{41}$$

At $\alpha = 0.4$, these values are $5/9$, $7/9$, and $4/9$.

Recall

$$X = \mu_y Y u + \mu_a A v + \varepsilon, \qquad r = \mu_y w^\top u - 1, \qquad t = \mu_a w^\top v.$$

The noiseless mixed residuals for SD and DS are respectively

$$Y_i r + A_i(2\lambda - 1)t, \qquad Y_i(2\lambda-1)r + A_i t.$$

For an SS pair of polarity s, the residual is $Y_i(r + st)$; for DD it is $Y_i(2\lambda - 1)(r + st)$. Taking coefficient expectations gives Equations (17)–(18).

Independent isotropic endpoint noise contributes

$$\mathbb{E}\Big[\big(w^\top[\lambda\varepsilon_i + (1-\lambda)\varepsilon_j]\big)^2\Big] = \beta_\alpha \sigma^2 \|w\|^2. \tag{42}$$

C.2 Constraint matrix and identification

Let

$$B_s = m_{\mathrm{SS},s} + \kappa_\alpha m_{\mathrm{DD},s}, \qquad s \in \{-1, +1\},$$

and write $H = H_{\mathrm{mix}}$. The signal risk is $\mathbf{c}^\top H \mathbf{c}$, where $\mathbf{c} = (r, t)^\top$, with

$$H = \begin{pmatrix} p_{\mathrm{SD}} + \kappa_\alpha p_{\mathrm{DS}} + B_+ + B_- & B_+ - B_- \\ B_+ - B_- & \kappa_\alpha p_{\mathrm{SD}} + p_{\mathrm{DS}} + B_+ + B_- \end{pmatrix}. \tag{43}$$

Proof of Proposition 2. The noiseless endpoint residual is $Yr + At$. Since the pooled endpoint expectation of YA is $1 - 2e_-$, the endpoint signal risk is

$$r^2 + t^2 + 2(1 - 2e_-)rt.$$

This gives the matrix

$$H_{\text{end}} = \begin{pmatrix} 1 & 1 - 2e_- \\ 1 - 2e_- & 1 \end{pmatrix}.$$

For $s \in \{-1, +1\}$, define

$$Q_s = \begin{pmatrix} 1 & s \\ s & 1 \end{pmatrix}.$$

Conditional on SS or DD at polarity s, the endpoint signal matrix is Q_s. The corresponding Mixup matrices are Q_s for SS and $\kappa_\alpha Q_s$ for DD. Conditional on either SD or DS, the endpoint signal matrix is the 2×2 identity matrix. The Mixup matrices are $\text{diag}(1, \kappa_\alpha)$ for SD and $\text{diag}(\kappa_\alpha, 1)$ for DS.

For every finite $\alpha > 0$, $0 < \kappa_\alpha < 1$. Each conditional Mixup matrix is therefore bounded below by κ_α times its endpoint counterpart and above by that counterpart in the positive-semidefinite order. Averaging over the pair distribution gives

$$\kappa_\alpha H_{\text{end}} \preceq H_{\text{mix}} \preceq H_{\text{end}},$$

which proves Equation (19).

The eigenvalues of H_{end} are $2e_-$ and $2(1 - e_-)$. If $0 < e_- < 1$, then $H_{\text{end}} \succ 0$, and the lower bound $H_{\text{mix}} \succeq \kappa_\alpha H_{\text{end}}$ implies $H_{\text{mix}} \succ 0$. If $e_- \in \{0, 1\}$, then H_{end} is singular; because $0 \preceq H_{\text{mix}} \preceq H_{\text{end}}$, every vector in the null space of H_{end} also lies in the null space of H_{mix}. Hence H_{mix} is positive definite if and only if $0 < e_- < 1$. By Proposition 1, this is equivalent to $\text{DPE} + C_{\text{coll}} > 0$. In this case, both noiseless signal risks have the unique coefficient minimizer $(r, t) = (0, 0)$.

At zero DPE, the Mixup signal risk reduces to

$$B_+(r + t)^2 + B_-(r - t)^2.$$

Its eigenvectors are proportional to $(1, 1)$ and $(1, -1)$, with eigenvalues $2B_+$ and $2B_-$, respectively.

More generally, writing $\lambda_{\min}(H)$ for the smallest eigenvalue of H, the decomposition into direct and same-polarity contributions gives

$$\begin{aligned} \lambda_{\min}(H) \geq &\min\{p_{\text{SD}} + \kappa_\alpha p_{\text{DS}}, \kappa_\alpha p_{\text{SD}} + p_{\text{DS}}\} \\ &+ 2\min(B_+, B_-). \end{aligned} \tag{44}$$

The model-specific effective complementary curvature is

$$C_{\text{coll}}^{\text{eff}} = 2\min(B_+, B_-).$$

These matrices describe the noiseless signal objective. The complete population objective additionally includes the observation-noise contribution $\beta_\alpha \sigma^2 \|w\|^2$.

C.3 Random-pair conditioning

Under the symmetric random-pairing model,

$$B_+ = \frac{1 + \kappa_\alpha}{2}\rho^2, \qquad B_- = \frac{1 + \kappa_\alpha}{2}(1 - \rho)^2.$$

Combining these terms with $p_{\text{SD}} = p_{\text{DS}} = \rho(1 - \rho)$ gives

$$H_{\text{random}} = \beta_\alpha \begin{pmatrix} 1 & 2\rho - 1 \\ 2\rho - 1 & 1 \end{pmatrix}. \tag{45}$$

Its eigenvalues and condition number are

$$\lambda_{\min} = 2\beta_\alpha(1 - \rho), \qquad \lambda_{\max} = 2\beta_\alpha\rho, \qquad \text{cond}(H_{\text{random}}) = \frac{\rho}{1 - \rho}. \tag{46}$$

Complementary non-DPE curvature vanishes quadratically in $1 - \rho$, whereas the smallest eigenvalue of the complete signal matrix vanishes linearly because of the direct SD/DS contribution.

C.4 Coverage-family and endpoint equivalence

The controlled family assigns

$$p_{\mathrm{SD}} = p_{\mathrm{DS}} = \frac{\delta}{2},$$
$$m_{\mathrm{SS},+} = m_{\mathrm{DD},+} = \frac{(1-\delta)(1-q)}{2},$$
$$m_{\mathrm{SS},-} = m_{\mathrm{DD},-} = \frac{(1-\delta)q}{2}. \tag{47}$$

Its negative-polarity endpoint mass is

$$e_- = \frac{\delta}{2} + (1-\delta)q.$$

With balanced signs within each polarity,

$$E^{\mathrm{pos}}_{++} = E^{\mathrm{pos}}_{--} = \frac{1-e_-}{2}, \qquad E^{\mathrm{pos}}_{+-} = E^{\mathrm{pos}}_{-+} = \frac{e_-}{2}. \tag{48}$$

Substitution into the topology risks yields

$$\mathcal{R}_{\mathrm{mix}} = \beta_\alpha \left[r^2 + t^2 + 2(1-2e_-)rt + \sigma^2 \|w\|^2 \right]. \tag{49}$$

The endpoint risk is the bracketed expression without β_α. Their unregularized population minimizers therefore depend on e_-, not separately on δ and q. At $q = 1/2$, every group has exposure $1/4$.

For i.i.d. endpoint residuals e_i, e_j, a related identity is

$$\mathbb{E}[(\lambda e_i + (1-\lambda)e_j)^2] = \beta_\alpha \mathbb{E}[e^2] + (1-\beta_\alpha)(\mathbb{E}[e])^2. \tag{50}$$

The centered symmetric model has $\mathbb{E}[e] = 0$, so independent random Mixup and endpoint training share their unregularized population squared-loss minimizers. This equivalence concerns the population objective analyzed here.

C.5 Population minimizer with observation noise

For the controlled coverage family, let

$$\gamma = 1 - 2e_-, \qquad w_u = w^\top u, \qquad w_v = w^\top v, \qquad w_\perp = w - w_u u - w_v v.$$

For $\sigma > 0$, the complete population squared loss is strictly convex in w. Its minimizer has $w_\perp = 0$, and the signal-plane coefficients satisfy

$$\begin{pmatrix} \mu_y^2 + \sigma^2 & \gamma\mu_y\mu_a \\ \gamma\mu_y\mu_a & \mu_a^2 + \sigma^2 \end{pmatrix} \begin{pmatrix} w_u \\ w_v \end{pmatrix} = \begin{pmatrix} \mu_y \\ \gamma\mu_a \end{pmatrix}.$$

Define the positive determinant

$$D_\sigma = (\mu_y^2 + \sigma^2)(\mu_a^2 + \sigma^2) - \gamma^2\mu_y^2\mu_a^2.$$

Then

$$w_u^\star = \frac{\mu_y\left[\sigma^2 + \mu_a^2(1-\gamma^2)\right]}{D_\sigma}, \qquad w_v^\star = \frac{\gamma\mu_a\sigma^2}{D_\sigma}. \tag{51}$$

Mixup and endpoint training share this minimizer because their unregularized population objectives differ by the positive factor β_α.

At balanced endpoint polarity, $\gamma = 0$, so

$$w_u^\star = \frac{\mu_y}{\mu_y^2 + \sigma^2}, \qquad w_v^\star = 0.$$

For imbalanced endpoint polarity, the noisy population optimum retains a nonzero spurious coefficient. This distinguishes identification in the noiseless signal objective from minimization of the complete noisy risk.

C.6 Population group accuracy

For $w \neq 0$, $\sigma > 0$, and a classifier $\widehat{Y} = \operatorname{sign}(w^\top x)$,

$$Y w^\top X \mid (Y = y, A = a) \sim \mathcal{N}\big(\mu_y w_u + ya\,\mu_a w_v,\, \sigma^2 \|w\|^2\big)\,,$$

where $w_u = w^\top u$ and $w_v = w^\top v$. Writing Φ for the standard normal CDF,

$$\mathrm{Acc}_{y,a}(w) = \Phi\left(\frac{\mu_y w_u + ya\,\mu_a w_v}{\sigma \|w\|}\right). \tag{52}$$

With all four groups present,

$$\mathrm{WGA}_{\mathrm{pop}}(w) = \Phi\left(\frac{\mu_y w_u - \mu_a |w_v|}{\sigma \|w\|}\right). \tag{53}$$

For an affine classifier with intercept b, the numerator becomes $\mu_y w_u - \mu_a |w_v| - |b|$.

When $w_u > 0$, the signal ratio $\mu_a |w_v| / (\mu_y w_u)$ captures one component of WGA. The core coefficient, total norm, signal amplitudes, and intercept also matter. In the synthetic configuration $\mu_y = 1, \sigma = 0.7$, a positive core-only classifier has accuracy

$$\Phi(1/0.7) \approx 0.9234.$$

D Experimental setup and statistical reporting

All empirical models use binary cross-entropy (BCE) with logits. Hard labels are encoded as $z = (Y + 1)/2 \in \{0, 1\}$, and mixed targets are convex combinations of these encoded labels. For logit s and target $z \in [0, 1]$,

$$\ell_{\mathrm{BCE}}(s, z) = \log(1 + \exp(s)) - zs. \tag{54}$$

Mixup applies this loss to a mixed input or representation and its mixed target. Squared loss is used only for the analysis.

D.1 Datasets and architectures

Synthetic data. Inputs follow Equation (16) in dimension 20, with $\mu_y = 1$, $\mu_a = 2$, and $\sigma = 0.7$. For a nominal training agreement probability ρ, each training group is sampled according to

$$\pi_{+1,+1} = \pi_{-1,-1} = \frac{\rho}{2}, \qquad \pi_{+1,-1} = \pi_{-1,+1} = \frac{1 - \rho}{2}.$$

Training, validation, and test sizes are $8{,}000$, $4{,}000$, and $8{,}000$, respectively. Validation and test groups are uniform. An affine classifier is trained with input Mixup. The random-pairing sweep uses $\rho \in \{0.70, 0.80, 0.90, 0.95, 0.98\}$. The support intervention evaluates

$$\delta \in \{0, 0.25, 0.50, 0.75, 1\}, \qquad q \in \{0, 0.05, 0.25, 0.50\},$$

at $\rho \in \{0.95, 0.98\}$. The source agreement probability determines the available training examples, and the intervention determines the sampled endpoints.

Colored MNIST. We use digit parity as the target and red/blue color as the spurious attribute (LeCun et al., 1998; Arjovsky et al., 2019). Training agreement probability is $\rho = 0.95$. The split contains $12{,}000/3{,}000/4{,}000$ training/validation/test examples, with balanced validation and test groups. A two-layer CNN uses input Mixup.

Biased CIFAR. The target is animals versus vehicles in CIFAR-10 (Krizhevsky, 2009). A 4×4 patch with opacity 0.8 is correlated with the target at $\rho = 0.95$. The split contains $20{,}000/5{,}000/8{,}000$ images. A four-layer CNN has a 256-dimensional penultimate representation. Input and feature Mixup use the same five seeds and optimization settings.

Table 3: Main training settings. All models use Adam (Kingma & Ba, 2015) with weight decay 10^{-4}; main Mixup runs use $\alpha = 0.4$.

Dataset	Learning rate	Batch	Epochs	Seeds	Space	Checkpoint
Synthetic	10^{-2}	256	35	20	Input	Final
Colored MNIST	10^{-2}	256	35	10	Input	Final
Biased CIFAR	10^{-3}	128	50	5	Input/feature	Validation WGA
Waterbirds	10^{-4}	64	20	5	Feature	Validation WGA
CelebA	10^{-4}	64	10	5	Feature	Validation WGA

Waterbirds and CelebA. Waterbirds uses bird type and background (Sagawa et al., 2020; Wah et al., 2011; Zhou et al., 2018); CelebA uses Blond Hair and Male (Liu et al., 2015). Both use ResNet-18 backbones pretrained on ImageNet (Deng et al., 2009; He et al., 2016) and penultimate-feature Mixup. Training images are resized to 256 pixels, randomly cropped to 224×224, and horizontally flipped. Evaluation uses a centered 224×224 crop. All images use ImageNet normalization.

CelebA includes empirical-policy endpoint controls but no balanced single-topology experiments. Waterbirds includes both empirical and balanced-topology policies.

D.2 Optimization, selection, and computation

Table 3 summarizes the main optimization settings, training durations, interpolation spaces, and checkpoint criteria.

The Waterbirds GB-SD versus GB-SS comparison uses ten matched seeds. Its exploratory $\alpha \in \{0.1, 1.0\}$ runs use seeds $0, 1, 2$. Group-aware samplers use training labels and spurious-attribute metadata; validation-WGA selection also uses validation group labels. ERM and vanilla Mixup do not use group metadata in their training samplers.

Synthetic and Colored MNIST models contain no batch normalization. For ResNet feature Mixup, the two endpoint minibatches pass separately through the backbone before mixing. The endpoint reference uses the same two backbone passes and one optimizer update. Batch-normalization statistics are updated during both passes.

Input Mixup evaluates B mixed inputs per optimizer update, whereas its endpoint reference evaluates the corresponding $2B$ clean endpoints. The paired comparisons use identical pair draws and numbers of optimizer updates. ERM and ordinary group-balanced sampling process one endpoint minibatch per update.

D.3 Statistics, diagnostics, and reproduction records

For each method and seed s, let $a_{g,s}$ be its test accuracy in group g. We compute

$$\mathrm{WGA}_s = \min_{g \in \mathcal{G}} a_{g,s}$$

before averaging across seeds. Absolute accuracy entries are reported as means and sample standard deviations unless a table specifies otherwise. Coverage and diagnostic columns without uncertainty summaries report means.

For two methods A and B evaluated with K matched seeds, define

$$\Delta_s = \mathrm{WGA}_{A,s} - \mathrm{WGA}_{B,s}, \qquad \overline{\Delta} = \frac{1}{K} \sum_{s=1}^{K} \Delta_s,$$

and

$$s_\Delta = \left[\frac{1}{K-1} \sum_{s=1}^{K} (\Delta_s - \overline{\Delta})^2 \right]^{1/2}.$$

Table 4: Synthetic random-pairing sweep over 20 seeds. Swap sensitivity is defined in Equation (24).

ρ	DPE	C_{coll}	WGA (%)	Swap (%)	$\lvert w_v\rvert/\lvert w_u\rvert$
0.70	0.4217	0.1835	87.08 ± 0.91	7.93	0.094
0.80	0.3229	0.0820	82.00 ± 1.23	14.43	0.168
0.90	0.1815	0.0205	69.29 ± 1.68	28.40	0.308
0.95	0.0951	0.0050	50.31 ± 1.62	48.20	0.490
0.98	0.0390	0.0008	23.65 ± 2.31	75.47	0.818

Table 5: Synthetic baselines at $\rho = 0.98$, over 20 seeds. Entries are WGA percentages reported as mean $\pm$ sample standard deviation.

Method	WGA (%)
ERM	37.69 ± 2.89
Random-pair Mixup	23.65 ± 2.31
Group-balanced sampling	89.69 ± 1.12
Group-balanced Mixup	90.27 ± 0.72

The two-sided 95% Student's t confidence interval is

$$\overline{\Delta} \pm t_{K-1,0.975} \frac{s_\Delta}{\sqrt{K}},$$

where $t_{\nu,p}$ is the p-quantile of the Student's t distribution with ν degrees of freedom. For the reporting decomposition, the differential residual is formed within seed before applying the same calculation. These intervals summarize variation across the matched runs and are reported without multiple-comparison correction.

Contrasts without intervals are differences of means. All means, contrasts, and paired intervals are computed from the unrounded per-seed records; paired intervals use the within-seed differences defined above.

For every pair-based run, we log anchor/partner groups, topology and coefficient draws, endpoint exposure, DPE, complementary coverage, fallback events, and per-seed evaluation metrics. Synthetic runs additionally log $|w^\top v|/|w^\top u|$. Controlled tasks compute the paired spurious-attribute intervention and swap sensitivity defined in Equation (24).

The accompanying release contains split identifiers and indices, attribute mappings and group counts, preprocessing and architectures, training/checkpoint configurations, candidate-pool and fallback rules, self-pair and replacement settings, endpoint implementations, marginal-matched sampling code, per-seed records, and scripts for all tables and figures.

E Complete benchmark and support results

Pair statistics attached to endpoint references describe their retained originating-policy samplers. Unless otherwise stated, accuracy values are percentages reported as mean $\pm$ sample standard deviation, and coverage values are fractions. Where used, boldface marks the highest mean WGA within the corresponding directly comparable column.

E.1 Synthetic random-pairing sweep

Observed coverage in Table 4 closely follows Equation (29). As source agreement probability grows, coverage declines, the spurious-to-core weight ratio increases, and WGA falls. Table 5 reports the corresponding baseline comparison at the strongest training agreement probability.

Table 6: Synthetic Mixup coverage grid, with 20 seeds per reported estimate. Columns vary the negative-polarity probability q in the non-DPE branch. When $\delta = 1$, the non-DPE branch is inactive; the four estimates in that row are repetitions of the same nominal pairing condition rather than distinct q settings.

ρ	δ	$q = 0$	$q = 0.05$	$q = 0.25$	$q = 0.50$
0.95	0.00	3.11 ± 0.59	50.56 ± 1.69	84.15 ± 1.10	91.02 ± 0.77
	0.25	73.90 ± 1.23	78.46 ± 1.44	86.75 ± 1.15	90.87 ± 0.82
	0.50	84.73 ± 1.03	85.68 ± 1.01	88.83 ± 0.94	90.86 ± 0.79
	0.75	89.06 ± 1.14	89.25 ± 0.95	90.33 ± 0.99	90.85 ± 0.90
	1.00	Repetitions 1–4: 91.01 ± 0.67, 90.97 ± 0.70, 90.90 ± 0.75, 90.95 ± 0.79			
0.98	0.00	3.07 ± 0.54	50.35 ± 2.30	83.05 ± 1.45	90.41 ± 0.88
	0.25	73.68 ± 2.01	77.98 ± 1.71	86.05 ± 1.09	90.34 ± 0.86
	0.50	84.16 ± 1.31	85.26 ± 1.31	87.99 ± 1.15	90.46 ± 0.87
	0.75	88.13 ± 0.97	88.57 ± 1.08	89.49 ± 1.00	90.59 ± 0.89
	1.00	Repetitions 1–4: 90.35 ± 0.87, 90.39 ± 0.70, 90.46 ± 0.73, 90.35 ± 0.87			

Table 7: Policy-matched endpoint references for the synthetic coverage grid, with 20 seeds per reported estimate. No interpolation is applied. When $\delta = 1$, the non-DPE branch is inactive; the four estimates in that row are repetitions of the same nominal pairing condition rather than distinct q settings.

ρ	δ	$q = 0$	$q = 0.05$	$q = 0.25$	$q = 0.50$
0.95	0.00	8.19 ± 1.32	59.71 ± 2.24	84.12 ± 1.37	90.86 ± 0.99
	0.25	75.38 ± 1.59	78.90 ± 1.54	86.46 ± 1.41	90.63 ± 0.94
	0.50	84.17 ± 1.29	84.97 ± 1.38	88.31 ± 1.18	90.72 ± 0.96
	0.75	88.37 ± 1.36	88.67 ± 1.24	89.86 ± 1.16	90.62 ± 0.90
	1.00	Repetitions 1–4: 90.74 ± 0.84, 90.83 ± 0.94, 90.77 ± 0.94, 90.83 ± 0.85			
0.98	0.00	8.17 ± 1.28	58.74 ± 2.78	82.77 ± 1.65	89.83 ± 0.99
	0.25	74.49 ± 2.17	78.06 ± 2.04	85.34 ± 1.41	89.77 ± 0.99
	0.50	82.94 ± 1.76	83.98 ± 1.72	87.02 ± 1.52	89.91 ± 0.88
	0.75	87.07 ± 1.38	87.43 ± 1.33	88.65 ± 1.37	89.87 ± 1.07
	1.00	Repetitions 1–4: 89.79 ± 0.86, 89.77 ± 0.80, 89.85 ± 0.93, 89.60 ± 0.92			

E.2 Full synthetic support grid

Tables 6 and 7 report every tested (δ, q) condition. Table 8 retains the achieved support diagnostics and Mixup–endpoint differences for selected conditions, without repeating their absolute WGA values.

The grids exhibit a common trend when indexed by $e_- = \delta/2 + (1 - \delta)q$: direct and complementary allocations produce similar WGA at similar endpoint-polarity mass, and the balanced allocation paths converge near 90% at $\rho = 0.98$. The controlled dimensions in this family are pair topology and endpoint polarity.

At $\delta = 1$, the non-DPE branch is inactive, so the four entries in each row are independent repetitions of the same induced pairing condition. The random-pairing sweep varies the source distribution, whereas the support grid varies endpoint allocation.

Table 8: Selected support diagnostics at $\rho = 0.98$, over 20 seeds. The last column is mean Mixup WGA minus mean endpoint WGA, in percentage points, computed from unrounded records. Absolute WGA values appear in the full grids.

δ	q	C_{coll}	e_-	Mixup minus endpoint
0.00	0.00	0.000	0.000	-5.10
0.00	0.05	0.100	0.050	-8.39
0.00	0.25	0.499	0.250	$+0.28$
0.00	0.50	0.999	0.500	$+0.58$
0.50	0.00	0.000	0.250	$+1.22$
0.50	0.25	0.250	0.375	$+0.96$
0.75	0.00	0.000	0.375	$+1.06$
1.00	0.00	0.000	0.500	$+0.56$

Table 9: Balanced single-topology controls with uniform group-position marginals. Entries are WGA percentages. Waterbirds GB-SS and GB-SD use ten matched seeds; its other rows use five.

	Synthetic, $\rho = 0.98$		Colored MNIST		Biased CIFAR		Waterbirds	
Policy	Mixup	Endpoint	Mixup	Endpoint	Input Mixup	Feature Mixup	Feature	Endpoint
GB-SS	89.81 ± 1.00	89.78 ± 1.02	92.07 ± 2.91	89.64 ± 3.26	86.75 ± 0.95	86.96 ± 0.95	77.53 ± 2.26	77.46 ± 2.24
GB-SD	89.69 ± 0.92	89.79 ± 1.02	90.61 ± 2.62	91.25 ± 2.04	86.37 ± 2.12	86.85 ± 1.07	80.05 ± 2.08	79.28 ± 1.93
GB-DS	90.51 ± 0.85	89.64 ± 1.05	93.82 ± 1.12	90.59 ± 2.35	88.09 ± 0.82	87.78 ± 0.89	82.14 ± 1.74	79.85 ± 1.82
GB-DD	90.36 ± 1.04	89.61 ± 1.06	93.48 ± 1.76	90.23 ± 4.59	87.39 ± 0.59	87.42 ± 0.78	81.54 ± 1.55	78.55 ± 1.86

E.3 Balanced single-topology controls

Table 9 shows modest topology differences on synthetic data and biased CIFAR. For Waterbirds, the GB-SD versus GB-SS point estimate is positive under both Mixup and endpoint training. The paired decompositions and confidence intervals are reported in Table 1. The GB-DS versus GB-DD residual is -0.69 pp, with 95% CI $[-3.60, 2.22]$.

E.4 Colored MNIST and its polarity intervention

Tables 10 and 11 show that High-DPE and Low-DPE both improve robustness despite their opposite direct-coverage values. At zero DPE, restoring negative-polarity support and thereby balancing the endpoint polarities raises WGA from zero to above 90%.

The balanced coverage-family sampler and the Low-DPE policy in Table 10 share the same population group-pair distribution but use different anchor schedules. Low-DPE uses exactly balanced anchor counts within each epoch, whereas the coverage-family implementation samples anchor groups independently.

E.5 Biased CIFAR

Table 12 shows that feature-space vanilla Mixup exceeds input-space vanilla Mixup by 7.53 pp. The balanced policies vary less across interpolation spaces, and their endpoint references reach approximately 86% WGA.

E.6 Waterbirds

Balanced Low-DPE exceeds empirical-support Low-DPE by 5.79 pp, despite both having zero DPE. Balanced Low-DPE and High-DPE exceed their endpoint references by 3.40 pp and 1.51 pp, respectively.

Table 10: Colored MNIST at $\rho = 0.95$, over 10 seeds. WGA and Swap are percentages; Swap is defined in Equation (24).

Method	DPE	C_{coll}	WGA	Swap
ERM	—		85.60 ± 3.17	12.00
Vanilla Mixup	0.095	0.005	82.22 ± 6.09	14.58
Group-balanced sampling	—		90.43 ± 1.50	7.00
Group-balanced Mixup	0.500	0.499	$\mathbf{94.78 \pm 1.83}$	3.07
High-DPE	1.000	0.000	93.10 ± 2.65	4.41
Low-DPE	0.000	1.000	91.84 ± 4.79	4.76

Table 11: Colored MNIST zero-DPE support intervention over 10 seeds. Negative-polarity pair exposure E_{-}^{pair} is the probability that a pair contains at least one negative-polarity endpoint, as defined in Equation (27).

Policy	DPE	C_{coll}	Pair exposure	WGA (%)
$\delta = 0,\ q = 0$	0	0	0	0.00 ± 0.00
$\delta = 0,\ q = 0.5$	0	0.999	0.500	92.38 ± 2.50
Matched endpoint	0	0.999	0.500	90.97 ± 2.69

E.7 Waterbirds coefficient sensitivity

Table 14 reports positive mean Mixup contrasts at all three coefficients. The three-seed estimates at $\alpha = 0.1$ and 1.0 have wide confidence intervals. At the main setting $\alpha = 0.4$, the differential residual is 0.70 percentage points, with 95% CI $[-2.54, 3.95]$.

E.8 CelebA

Sensitivity-oriented policies improve WGA while reducing average accuracy. Their advantage over invariance-oriented policies is also visible under policy-matched endpoint training, while the individual Mixup–endpoint increments vary across policies.

E.9 Marginal-matched controls and group exposure

The mean marginal-matched contrasts preserve the directions highlighted in the main text:

$$\text{CelebA only-sens} - \text{only-inv} = +12.22\,\text{pp},$$
$$\text{CelebA sens08} - \text{inv08} = +7.11\,\text{pp},$$
$$\text{Waterbirds inv08} - \text{sens08} = +1.72\,\text{pp}.$$

The two endpoint constructions yield different absolute WGA values and contrast magnitudes. For example, CelebA only-sens reaches $74.00 \pm 3.67\%$ under marginal matching and $87.44 \pm 2.56\%$ under policy matching.

All balanced GB topologies target uniform pooled exposure. Their marginal-matched Waterbirds reference reaches $79.28 \pm 1.67\%$ WGA, while the policy-matched references retain their topology-specific minibatch construction.

The two-digit table codes are (y_{01}, a_{01}). When an integer array index is needed, it is $k = 2y_{01} + a_{01} \in \{0, 1, 2, 3\}$; the mathematical group variable remains $G = (Y, A)$. Only-inv retains the source label marginal while balancing attributes within each label. Only-sens balances labels within each attribute. Their blond-male endpoint exposures differ by a factor of approximately 2.81, despite identical DPE.

F Representation and reliance diagnostics

The synthetic ratio $|w^\top v|/|w^\top u|$ summarizes the relative loading on the spurious and core signal directions and is invariant to a common rescaling of w. Equation (53) gives the corresponding

Table 12: Biased CIFAR WGA for the 4×4, opacity-0.8 patch setting, over five matched seeds.

Method	Input	Feature
ERM	68.72 ± 2.01	68.72 ± 2.01
Vanilla Mixup	63.57 ± 2.29	71.10 ± 1.74
Group-balanced sampling	86.36 ± 1.95	86.36 ± 1.95
Group-balanced Mixup	87.35 ± 1.48	86.17 ± 1.07
High-DPE	87.36 ± 2.44	86.07 ± 1.40
High-DPE endpoint	86.34 ± 2.29	86.34 ± 2.29
Low-DPE	$\mathbf{87.67 \pm 1.06}$	87.03 ± 1.23
Low-DPE endpoint	86.05 ± 2.02	86.05 ± 2.02

Table 13: Waterbirds under the 20-epoch protocol, over five seeds. Coverage columns report achieved mean fractions.

Method	$\mathrm{DPE}_{\mathrm{inv}}$	$\mathrm{DPE}_{\mathrm{sens}}$	C_{coll}	WGA (%)
ERM	—	—	—	70.59 ± 3.11
Vanilla Mixup	0.061	0.035	0.005	71.32 ± 3.28
Vanilla endpoint	0.061	0.035	0.005	68.57 ± 4.98
Group-balanced sampling	—	—	—	79.20 ± 1.96
Group-balanced endpoint	0.253	0.248	0.496	79.18 ± 2.21
Group-balanced Mixup	0.253	0.248	0.496	$\mathbf{81.25 \pm 1.95}$
Low-DPE empirical	0.000	0.000	0.100	73.46 ± 1.37
Low-DPE balanced	0.000	0.000	0.996	79.25 ± 1.85
Low-DPE endpoint	0.000	0.000	0.996	75.85 ± 2.44
High-DPE	0.499	0.501	0.000	81.03 ± 2.89
High-DPE endpoint	0.499	0.501	0.000	79.52 ± 1.98
CS-Mix inv08	0.813	0.007	≈ 0.001	78.05 ± 4.30
CS-Mix inv08 endpoint	0.813	0.007	≈ 0.001	75.67 ± 1.21
CS-Mix sens08	0.012	0.809	≈ 0.001	73.61 ± 3.86
CS-Mix sens08 endpoint	0.012	0.809	≈ 0.001	70.27 ± 3.93
Only-inv	1.000	0.000	0.000	78.47 ± 1.00
Only-sens	0.000	1.000	0.000	76.01 ± 3.32

WGA dependence on signal amplitudes, the noise-bearing norm, and the intercept. Swap sensitivity provides an intervention-based diagnostic on the controlled datasets.

For CelebA, a linear probe predicts Male from the frozen penultimate representations of each validation-WGA-selected classifier. The probe is fitted on representations from the standard validation split using sample weights that give the two attribute classes equal total weight, Adam with learning rate 10^{-2}, a fixed initialization, and 10 epochs. It is then evaluated once on representations from the standard test split. The test split is not used to fit the probe, select its hyperparameters, choose its training duration, or select a probe checkpoint. The original encoder and classifier remain frozen throughout probe fitting and evaluation.

The sample weights are used only in the validation-set probe-training loss. Ordinary probe accuracy is the unweighted test-set accuracy. After mapping the prediction back to $\widehat{A} \in \{-1, +1\}$, test balanced accuracy is

$$\mathrm{BAcc}_{A,\mathrm{test}} = \frac{1}{2}\left[\mathbb{P}_{\mathcal{D}_{\mathrm{test}}}(\widehat{A} = -1 \mid A = -1) + \mathbb{P}_{\mathcal{D}_{\mathrm{test}}}(\widehat{A} = +1 \mid A = +1)\right]. \tag{55}$$

Table 18 reports held-out test performance of the validation-fitted probes. Higher WGA does not consistently coincide with lower held-out linear attribute decodability. In particular, sensitivity-oriented policies can improve WGA while retaining Male information that generalizes to the test representations under a linear probe. These measurements quantify information accessible to this probe protocol; they do not establish that the target classifier causally relies on the attribute.

Table 14: Waterbirds GB-SD versus GB-SS sensitivity to the Beta coefficient. Panel (a) reports WGA percentages as mean ± sample standard deviation. Panel (b) reports paired mean differences in percentage points, with two-sided 95% Student's t confidence intervals below. The main setting $\alpha = 0.4$ uses ten matched seeds; $\alpha = 0.1$ and 1.0 use seeds $0, 1, 2$.

(a) Absolute WGA

α	Seeds	GB-SS		GB-SD	
		Mixup	Endpoint	Mixup	Endpoint
0.1	3	77.31 ± 4.13	78.89 ± 1.84	79.45 ± 1.74	79.39 ± 1.48
0.4	10	77.53 ± 2.26	77.46 ± 2.24	80.05 ± 2.08	79.28 ± 1.93
1.0	3	78.14 ± 3.05	77.73 ± 0.56	79.29 ± 5.50	76.75 ± 1.56

(b) GB-SD minus GB-SS

α	Mixup contrast	Endpoint contrast	Differential residual
0.1	$+2.14$ $[-12.01, +16.28]$	$+0.50$ $[-7.58, +8.58]$	$+1.64$ $[-4.45, +7.72]$
0.4	$+2.53$ $[+0.15, +4.91]$	$+1.83$ $[+0.07, +3.58]$	$+0.70$ $[-2.54, +3.95]$
1.0	$+1.15$ $[-14.91, +17.22]$	-0.98 $[-5.50, +3.55]$	$+2.13$ $[-12.85, +17.11]$

Table 15: CelebA results over five seeds with penultimate-feature Mixup. Accuracy values are percentages; coverage values are achieved mean fractions.

Method	$\mathrm{DPE}_{\mathrm{inv}}$	$\mathrm{DPE}_{\mathrm{sens}}$	C_{coll}	Avg.	GBA	WGA
ERM	—	—	—	95.79 ± 0.22	83.46 ± 1.06	48.56 ± 4.13
Vanilla Mixup	0.3638	0.1309	0.4031	95.56 ± 0.26	84.94 ± 0.52	55.11 ± 1.07
Vanilla endpoint	0.3638	0.1309	0.4031	95.71 ± 0.12	83.70 ± 0.94	50.33 ± 3.44
Group-balanced sampling	—	—	—	93.27 ± 1.13	90.61 ± 1.11	80.95 ± 6.24
Group-balanced Mixup	0.2495	0.2500	0.4994	93.81 ± 0.27	89.95 ± 0.87	77.00 ± 4.19
Group-balanced endpoint	0.2495	0.2500	0.4994	93.78 ± 0.54	89.70 ± 1.69	76.44 ± 7.02
Low-DPE	0.0000	0.0000	0.9993	93.45 ± 0.55	90.24 ± 0.82	79.44 ± 3.99
Low-DPE endpoint	0.0000	0.0000	0.9993	93.75 ± 0.34	89.39 ± 1.24	75.44 ± 5.81
High-DPE	0.5003	0.4997	0.0000	94.02 ± 0.70	90.07 ± 0.99	78.56 ± 4.24
High-DPE endpoint	0.5003	0.4997	0.0000	93.99 ± 0.56	90.03 ± 1.33	77.44 ± 5.93
CS-Mix inv08	0.8726	0.0262	0.0806	94.66 ± 0.49	88.74 ± 1.06	73.67 ± 4.24
CS-Mix inv08 endpoint	0.8726	0.0262	0.0806	94.96 ± 0.46	87.59 ± 2.04	70.78 ± 7.61
CS-Mix sens08	0.0727	0.8265	0.0803	92.02 ± 1.21	91.53 ± 0.55	85.61 ± 2.61
CS-Mix sens08 endpoint	0.0727	0.8265	0.0803	91.71 ± 1.00	91.63 ± 0.95	86.83 ± 3.24
Only-inv	1.0000	0.0000	0.0000	94.63 ± 0.20	88.66 ± 1.00	75.78 ± 2.03
Only-inv endpoint	1.0000	0.0000	0.0000	94.42 ± 1.00	85.56 ± 1.60	68.14 ± 2.27
Only-sens	0.0000	1.0000	0.0000	91.36 ± 1.26	91.93 ± 0.60	$\mathbf{88.02 \pm 2.60}$
Only-sens endpoint	0.0000	1.0000	0.0000	92.69 ± 0.99	91.78 ± 0.60	87.44 ± 2.56

G Limitations and broader impact

Analytical setting. The polarity analysis considers binary labels, one binary spurious attribute, and a shared linear predictor under population squared loss. The empirical study evaluates the same endpoint and topology concepts with binary cross-entropy and nonlinear models.

Endpoint procedures. Policy-matched references preserve the pair and coefficient draws while training on two clean endpoints. For input-space experiments, this procedure processes $2B$ clean

Table 16: Marginal-matched endpoint WGA over five seeds. Both endpoint groups are independently sampled from the originating pairing policy's achieved pooled group-position exposure.

Dataset	Originating policy	WGA (%)
CelebA	only-inv	61.78 ± 3.37
	only-sens	74.00 ± 3.67
	inv08	64.00 ± 7.04
	sens08	71.11 ± 4.63
Waterbirds	inv08	76.17 ± 2.95
	sens08	74.45 ± 1.77
	uniform GB	79.28 ± 1.67

Table 17: Achieved CelebA group-position exposure, averaged over sampled endpoints and training epochs. The column codes use $y_{01} = (Y+1)/2$ for Blond Hair and $a_{01} = (A+1)/2$ for Male. Codes $00, 01, 10, 11$ denote non-blond female, non-blond male, blond female, and blond male, respectively.

Method	E^{pos}_{00}	E^{pos}_{01}	E^{pos}_{10}	E^{pos}_{11}
Vanilla Mixup	0.4404	0.4104	0.1407	0.0085
Group-balanced Mixup	0.2501	0.2500	0.2499	0.2500
Low-DPE	0.2505	0.2499	0.2501	0.2495
High-DPE	0.2499	0.2498	0.2502	0.2501
CS-Mix inv08	0.4284	0.4225	0.0877	0.0613
CS-Mix sens08	0.3203	0.2498	0.2604	0.1695
Only-inv	0.4255	0.4255	0.0745	0.0745
Only-sens	0.2903	0.2097	0.2903	0.2097

inputs per update, compared with B mixed inputs for Mixup. Marginal-matched references use independent endpoint groups and equal endpoint weights, defining a separate training procedure with the same target pooled exposure.

Experimental coverage. Most natural-image comparisons use five seeds. The Waterbirds GB-SD versus GB-SS comparison uses ten matched seeds, and the exploratory coefficient settings $\alpha \in \{0.1, 1.0\}$ use three seeds. CelebA includes balanced mixture policies, while the single-topology controls are evaluated on Waterbirds. Waterbirds and CelebA use penultimate-feature Mixup, and biased CIFAR compares input and penultimate-feature interpolation.

Metadata and sample support. Group-aware sampling uses training-group metadata, and validation-WGA selection uses validation-group labels. Resampling changes the frequency with which available examples participate in training while leaving the underlying set of examples fixed.

BROADER IMPACT

The methodological goal is to separate augmentation effects from the training support induced by the sampler. This distinction can improve the attribution and reporting of robustness gains across groups.

The experiments use group and sensitive-attribute annotations for controlled training, model selection, and evaluation. In applied settings, such annotations may be unavailable, noisy, incomplete, or defined differently across contexts. The present study evaluates learning methodology rather than demographic classification or deployment policy.

Table 18: CelebA test-set linear-probe accuracy for Male. Each probe is fitted on representations from the standard validation split and evaluated on representations from the standard test split. Attribute-balanced sample weights are used only for the probe-training loss. Ordinary accuracy is unweighted, and balanced accuracy is the mean of the two test attribute-conditional accuracies. Values are percentages reported as means ± sample standard deviations over five backbone-training seeds.

Method	Probe accuracy	Balanced probe accuracy
ERM	89.17 ± 1.04	89.41 ± 0.95
Vanilla Mixup	87.27 ± 1.68	87.34 ± 1.65
Group-balanced sampling	91.42 ± 0.76	92.06 ± 0.51
Group-balanced Mixup	89.83 ± 1.23	90.44 ± 0.91
Low-DPE	89.55 ± 1.43	90.12 ± 0.97
High-DPE	91.25 ± 0.54	91.59 ± 0.64
CS-Mix inv08	88.88 ± 1.66	89.97 ± 1.22
CS-Mix sens08	92.05 ± 0.57	92.18 ± 0.63
Only-inv	93.01 ± 0.89	93.26 ± 0.70
Only-sens	93.18 ± 1.14	92.71 ± 1.25