SPECTRAL SUPPRESSION FOR ANNOTATION-FREE BIAS MITIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks trained with Empirical Risk Minimization (ERM) often rely on spurious correlations rather than robust features, particularly when multiple bias attributes are strongly correlated with the target label. Existing debiasing methods typically require expensive per-sample bias annotations, which limits their practicality in real-world settings. We introduce **Spectral Alignment Suppression (SAS)**, a novel annotation-free debiasing method. SAS identifies the dominant spectral subspace of the centered feature matrix and attenuates classifier feature coordinates with high leverage in that subspace. This is combined with a cosine classifier that makes the logits invariant to uniform feature scaling and per-class prototype scaling. Across Biased-MNIST, CelebA, and MS-COCO, SAS improves substantially over ERM, surpasses annotation-free two-step baselines (JTT, LfF) in the majority of evaluated settings, and remains competitive with supervised debiasing methods without using bias annotations. A balanced spectral audit shows dataset-dependent geometry: bias dominates the leading subspace on Biased-MNIST, co-locates with target information on CelebA, and is strongly decodable together with the target in a compact top-32 subspace on MS-COCO. On unbiased benchmarks (CIFAR-10, STL-10, and MNIST), SAS maintains accuracy comparable to ERM.

1 INTRODUCTION

Deep networks trained with ERM frequently exploit spurious correlations rather than causal features. For example, a gender classifier may rely on visual cues such as wearing heavy makeup or lipstick, while a bird classifier may associate certain bird species with a water background. This is worse under *multi-attribute bias*, where several spurious attributes simultaneously correlate with the label, degrading accuracy on bias-conflicting examples.

We model the problem as follows. A model $f = h \circ g$ pairs an encoder $g\colon \mathcal{X} \to \mathbb{R}^d$ with a classifier $h\colon \mathbb{R}^d \to \mathcal{Y}$. Under ERM with spurious correlations between label y and attributes $\{b_1, \ldots, b_m\}$, the bias attributes can dominate the learned decision rule, causing high error where they misalign with y.

Dominant bias-mitigation methods – BC-BB (Hong & Yang, 2021), EnD (Tartaglione et al., 2021), FLAC (Sarridis et al., 2025), FairKL (Barbano et al., 2023), GMBM (Dwivedi et al., 2025) – all require explicit per-attribute annotation, impractical when the number and identity of confounds is unknown a priori. JTT (Liu et al., 2021) and LfF (Nam et al., 2020) avoid bias annotation but use a *two-step* procedure: an auxiliary biased model is trained first, and its per-sample errors (or generalized cross-entropy loss) proxy bias-conflicting status for reweighting a second-stage model. This proxy is indirect and brittle when the auxiliary model fails to isolate the bias signal cleanly.

Rather than enumerating bias attributes or training a separate proxy model, we ask what *geometric structure* characterizes multi-attribute biased representations, and whether it can be exploited directly, in a single pass, without per-sample bias labels. Our central observation concerns the centered batch feature matrix $\mathbf{Z} \in \mathbb{R}^{N \times d}$:

Dataset-dependent concentration of bias information in the dominant spectral subspace. Bias information is frequently decodable from a compact leading subspace, but its enrichment relative to target information depends on the dataset. On Biased-MNIST, bias probes dominate label

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.

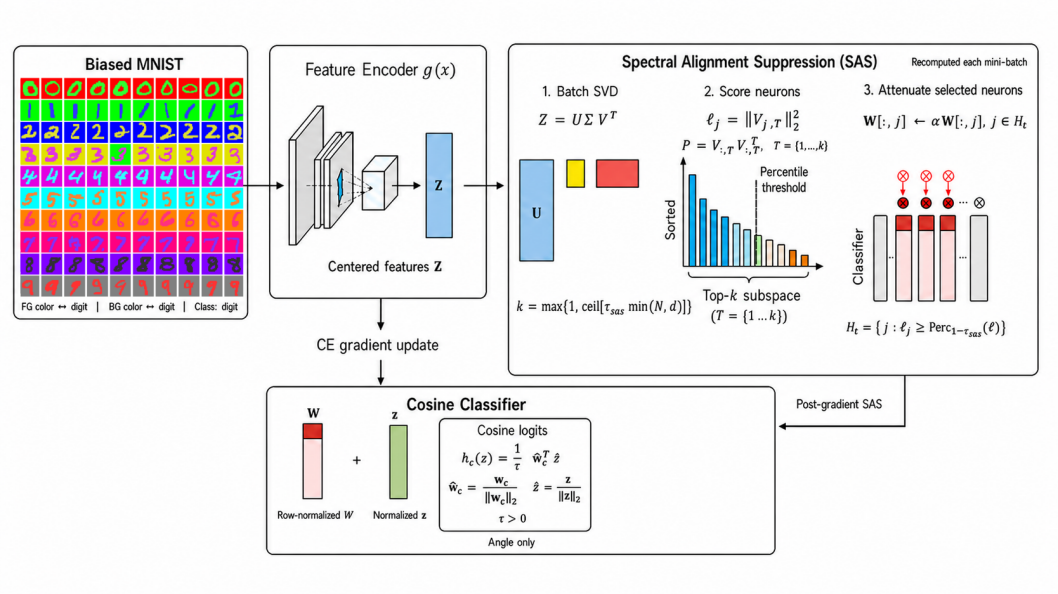

Figure 1: **Overview of Spectral Alignment Suppression (SAS).** An encoder $g(\mathbf{x})$ extracts centered features $\mathbf{Z}$, decomposed via SVD to find dominant spectral directions. Each feature coordinate gets a *spectral leverage score* from its overlap with the top-k right singular vectors. Coordinates above a percentile threshold are attenuated in the classifier during training. A cosine classifier makes predictions depend on angular alignment and removes scalar norm-based compensation. SAS is annotation-free and integrates into standard end-to-end training.

probes and top-direction ablation selectively removes bias-predictive signal (Section 3.2.2). On CelebA, label and bias information co-locate in the leading direction (Table 12). On MS-COCO, balanced probes show that the target dominates the earliest directions, while Sports and Kitchen information becomes increasingly decodable in the broader top-16/top-32 subspace (Table 11). Thus, spectral dominance is a soft, dataset-dependent *bias-risk prior*, not an automatic bias identifier.

We validate this spectral structure in Section 3.2 using balanced linear probes, AUROC, per-direction statistics, and directional ablations. The balanced controls are important on MS-COCO because raw probe accuracy is confounded by attribute imbalance.

SAS is a geometry-aware, annotation-free regularizer: after each gradient update it computes the SVD of the centered batch feature matrix, identifies the dominant subspace, and multiplicatively attenuates classifier feature coordinates with high spectral leverage in that subspace. A cosine classifier on ℓ_2-normalized features and class prototypes removes scalar norm-based compensation, while the coordinate-wise suppression changes classifier direction. Unlike JTT/LfF, SAS needs no auxiliary model, no second stage, and no error-based proxy: the dominant subspace is estimated directly from the model's own features within a single training run.

SAS outperforms ERM throughout the biased benchmarks and surpasses the annotation-free baselines JTT/LfF in the majority of evaluated settings. It is also competitive with several supervised baselines – notably achieving the best unbiased accuracy on CelebA HeavyMakeup without a single bias label at training time. As a sanity check, CIFAR-10, STL-10, and MNIST show accuracy comparable to ERM (Appendix K.6).

Contributions:

- We show empirically that multi-attribute spurious information is often concentrated in a compact dominant subspace, while establishing that its enrichment relative to target information is dataset-dependent. This distinction explains why SAS uses adaptive attenuation rather than hard subspace deletion.

- We introduce **SAS**, an annotation-free, single-step, geometry-aware regularizer that attenuates classifier feature coordinates with high leverage in this subspace after each gradient step – in contrast to JTT/LfF, which need an auxiliary biased model and a two-stage procedure.

2 Related Work

Supervised bias mitigation. Mutual-information methods (Hong & Yang, 2021) penalize dependence between predictions and a known bias attribute. EnD (Tartaglione et al., 2021) disentangles bias/target representations via entropy objectives; FLAC (Sarridis et al., 2025) builds bias-aware contrastive pairs; FairKL (Barbano et al., 2023) minimizes a KL-divergence fairness objective; GMBM (Dwivedi et al., 2025) combines geometric feature matching with group-balanced sampling. All require annotation of every bias attribute at training time.

Annotation-free two-step mitigation. JTT (Liu et al., 2021) trains an identification model with ERM, upweights its misclassified samples as a bias-conflicting proxy, then retrains. LfF (Nam et al., 2020) jointly trains a deliberately biased model (generalized cross-entropy) and a debiased model reweighted by their relative loss. Both are annotation-free in the same sense as SAS – no per-sample bias labels – but both need an auxiliary model, a separate stage or joint optimization, and a loss-based proxy for group membership. SAS instead runs in a single pass on a single model, using the batch feature geometry itself, with no auxiliary model or explicit bias-conflicting identification step. We compare directly against JTT and LfF throughout Section 5.

Multi-attribute bias mitigation. Methods for co-occurring spurious attributes typically extend supervised frameworks, with annotation burden growing per attribute. JTT/LfF sidestep per-attribute labels but were designed and evaluated mainly under single-attribute bias; their loss-based proxy does not distinguish which of several co-occurring attributes drives a sample's high loss. We instead target a unified geometric structure that need not be counted per attribute, evaluated explicitly under multi-attribute bias (Biased-MNIST with two simultaneous attributes; MS-COCO with two candidate bias categories).

Spectral methods and worst-group generalization. The singular spectrum of feature matrices shapes generalization and robustness (Bartlett et al., 2017), and spectral techniques inform transferability in domain adaptation (Blanchard et al., 2011). Our work is motivated by worst-group generalization (Sagawa et al., 2020): GroupDRO needs explicit group labels and scales exponentially with the number of attributes, whereas SAS targets the shared geometric structure of spurious information directly, operationalizing the observation that top singular directions under ERM tend to encode multi-attribute spurious correlations into a simple post-hoc suppression regularizer.

3 Methodology

3.1 Problem Formulation

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ with unobserved spurious attributes $\{b_i^{(1)}, \ldots, b_i^{(m)}\}$, each correlated with y_i in training but not necessarily at evaluation; neither m nor the attributes' identity is known. Given mini-batch features $\mathbf{Z} \in \mathbb{R}^{N \times d}$ (rows $\mathbf{z}_i = g(\mathbf{x}_i) - \bar{\mathbf{z}}$):

What "annotation-free" means here. SAS never uses per-sample bias labels during training – suppression operates purely on the geometry of $\mathbf{Z}$, with no bias label entering the loss, gradient, or suppression rule. Bias labels are used only for post-hoc evaluation (UA, BCA, SBA) and, for Biased-MNIST, validation-time selection via the known correlation parameter q. For MS-COCO, validation and checkpoint selection use only ordinary target-label accuracy; no object-category, co-occurrence, group, or bias information is used. The model is fully supervised on y; only the bias-mitigation mechanism is annotation-free. JTT and LfF (Section 2) satisfy this same definition but achieve it via an auxiliary model and a two-stage procedure rather than single-model geometry; we call them *annotation-free two-step* baselines to distinguish the mechanism from SAS's single-step approach.

3.2 Spectral Dominance: Geometric Signature and Per-Direction Validation

When an encoder relies on multiple spurious cues simultaneously, samples align along a small number of dominant feature directions, inflating the leading singular values: $\sum_{i=1}^{k} \sigma_i^2(\mathbf{Z})/\|\mathbf{Z}\|_F^2 \gg k/d$. We call the subspace spanned by the top-k right singular vectors the *dominant spectral subspace*.

3.2.1 Per-Direction Analysis: Top Singular Directions Encode Spurious Attributes

For CelebA and Biased-MNIST, we report linear-probe accuracy on each of the top-5 scalar projections $\mathbf{Z}\mathbf{v}_r$, together with KSG mutual information and η^2 against the label and both bias attributes (bias labels are used only for post-hoc verification). Table 12 reports mean $\pm$ std over three seeds. Biased-MNIST shows clear bias-over-label enrichment, whereas CelebA shows co-location of label and bias information in the leading direction.

MS-COCO requires a class-balanced audit because its raw attribute-probe accuracies are affected by label prevalence. We therefore fit class-balanced probes on cumulative top-k subspaces using training features, select probe regularization on validation features, and evaluate balanced accuracy (BA) and AUROC on held-out test features. Table 11 shows a different but informative geometry: the target dominates the earliest directions, while Sports and Kitchen become increasingly decodable as k grows. At $k = 32$, a subspace containing only 6.25% of the 512 feature dimensions retains 95.2% and 86.9% of the full-representation AUROC lift above chance for Sports and Kitchen, respectively; it also retains 98.5% of the target lift. COCO therefore supports compact *shared* target–bias concentration, rather than a bias-only dominant subspace.

Because label and bias information can co-occur in the dominant subspace (as on CelebA and COCO), suppression risks attenuating task-relevant signal. SAS mitigates this by decaying rather than deleting highly aligned classifier coordinates (Eq. (5)), using the subspace estimate as a soft prior rather than a hard filter. The strong COCO target concentration makes this design choice directly testable: hard top-direction deletion collapses target performance, whereas SAS improves bias-conflicting accuracy while preserving ordinary predictive utility (Section 5).

3.2.2 Directional Ablation: Bias Is Localized in Top Directions

Figure 2: ablating the top-16 singular directions sharply drops bias-probe accuracy (32.38 pp for $b^{(1)}$, 43.99 pp for $b^{(2)}$), while ablating an equal number of random directions leaves accuracy essentially unchanged – direct evidence that bias is structurally concentrated in the dominant subspace.

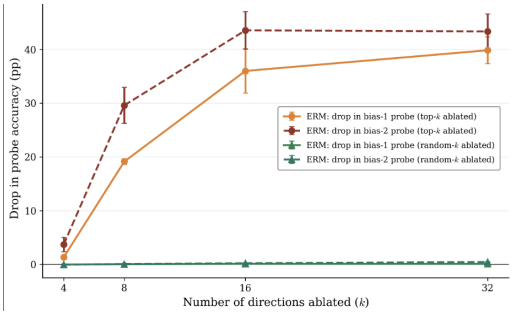

Figure 2: **Directional ablation (Biased-MNIST, $q = 0.95$).** Bias-probe accuracy after ablating the top-k singular directions (red, solid) vs. k random directions (blue, dashed). Single representative seed; qualitatively identical across seeds.

3.2.3 Aggregate Top-k Probe Analysis

Table 1 reports aggregate top-16 subspace statistics for Biased-MNIST and CelebA (mean over three seeds). On Biased-MNIST, SAS reverses the bias-over-label ordering by increasing label information and sharply reducing both bias probes. On CelebA, label and bias share the dominant

Table 1: **Aggregate Top-16 Singular Subspace Analysis** on Biased-MNIST and CelebA (mean over 3 seeds).

Dataset	Method	Probe Acc. (%)			Mutual Info.		
		Label	Bias1	Bias2	Label	Bias1	Bias2
Biased-MNIST (q=0.99)	ERM	25.13	99.87	95.95	0.0397	0.3574	0.2557
	SAS (Ours)	**33.72**	44.80	37.07	**0.1117**	0.1718	0.1273
CelebA	ERM	98.58	90.80	85.91	0.0604	0.0444	0.0356
	SAS (Ours)	**99.15**	88.07	81.78	**0.0858**	0.0632	0.0496

Algorithm 1 Spectral Alignment Suppression (SAS) + Cosine Classifier Training

Require: Training data $\mathcal{D}$, encoder $g(\cdot;\theta)$, cosine classifier $h(\cdot;\mathbf{W})$, batch size N, suppression ratio τ_{sas}, decay factor α, temperature $\tau_{\cos}$

Initialize g from pretrained ResNet-18 (or SimpleConvNet); initialize $\mathbf{W}$ with Xavier uniform
for each training iteration **do**
  Sample mini-batch $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N} \sim \mathcal{D}$
  $\mathbf{z}_i \leftarrow g(\mathbf{x}_i;\theta)$ *// raw penultimate features*
  $\mathbf{Z} \leftarrow [\mathbf{z}_1, \ldots, \mathbf{z}_N]^\top \in \mathbb{R}^{N\times d}$ *// batch feature matrix*
  Cosine Classifier Forward Pass:
  $\hat{\mathbf{z}}_i \leftarrow \mathbf{z}_i/\|\mathbf{z}_i\|_2, \quad \hat{\mathbf{w}}_c \leftarrow \mathbf{w}_c/\|\mathbf{w}_c\|_2 \quad \forall c = 1, \ldots, C$
  $\hat{\mathbf{y}}_i \leftarrow \frac{\hat{\mathbf{W}}\hat{\mathbf{z}}_i}{\tau_{\cos}}$ *// full logit vector for sample i*
  $\mathcal{L} \leftarrow \text{CrossEntropy}(\{\hat{\mathbf{y}}_i\}_{i=1}^{N}, \{y_i\}_{i=1}^{N})$
  Update parameters: $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}$, $\mathbf{W} \leftarrow \mathbf{W} - \eta\nabla_{\mathbf{W}}\mathcal{L}$
  Spectral Alignment Suppression (post-gradient):
  $\tilde{\mathbf{Z}} \leftarrow \mathbf{Z} - \bar{\mathbf{z}}$ *// center the batch features*
  $\mathbf{U}, \boldsymbol{\sigma}, \mathbf{V}^\top \leftarrow \text{SVD}(\tilde{\mathbf{Z}})$
  $k \leftarrow \max\big(1, \lceil \tau_{\text{sas}} \cdot \min(N, d) \rceil\big)$ *// or* $\text{rank}(\tilde{\mathbf{Z}})$
  Compute leverage scores: $\ell_j \leftarrow \|\mathbf{V}_{j,1:k}\|_2^2$ for each feature coordinate $j = 1, \ldots, d$
  $\mathcal{H} \leftarrow \{j \mid \ell_j \geq \text{percentile}(\boldsymbol{\ell}, 100(1-\tau_{\text{sas}}))\}$
  Attenuate selected coordinates: $\mathbf{W}_{:,j} \leftarrow \alpha \cdot \mathbf{W}_{:,j} \quad \forall j \in \mathcal{H}$
end for

Ensure: Debiased model $f = h \circ g$

region; SAS increases label decodability while modestly lowering bias-probe accuracy. The class-balanced COCO audit is reported separately in Table 11 because ordinary accuracy is not comparable across imbalanced targets.

On CelebA, the top-20 most-suppressed feature coordinates show higher mean bias-prediction accuracy than label accuracy (Appendix H). On Biased-MNIST, a random feature-coordinate-suppression control reaches only 26.2% unbiased accuracy at $q = 0.99$ (vs. 34.89% for SAS), confirming that SVD-guided selection drives most of the gain. Layer-level CKA between ERM and SAS on MS-COCO drops from 0.822 ± 0.015 in early-to-mid layers to 0.483 ± 0.065 in the deepest layers (Table 13), indicating spectral regularization primarily reshapes higher-level, task-specific representations while leaving low-level features relatively more aligned.

3.3 Algorithm

3.4 Spectral Alignment Suppression with Cosine Classification

3.4.1 Cosine Classifier

We use a cosine classifier to prevent the model from compensating for spectral suppression simply by increasing feature or class-prototype norms. Accordingly, we replace the linear head with a cosine classifier. Each class prototype $\mathbf{w}_c \in \mathbb{R}^d$ (for $c = 1, \ldots, C$) is independently ℓ_2-normalized, so the logit for class c depends only on the angle between the normalized feature and the normalized

prototype:

$$h_c(\mathbf{z}) = \frac{1}{\tau} \cdot \hat{\mathbf{w}}_c^\top \hat{\mathbf{z}}, \qquad \hat{\mathbf{w}}_c = \frac{\mathbf{w}_c}{\|\mathbf{w}_c\|_2}, \qquad \hat{\mathbf{z}} = \frac{\mathbf{z}}{\|\mathbf{z}\|_2}, \tag{1}$$

with temperature $\tau > 0$.

Let $\mathbf{W} \in \mathbb{R}^{C \times d}$ be the classifier weight matrix whose rows are the class prototypes $\mathbf{w}_c^\top$. Each row is normalized independently, while the feature vector is also normalized; consequently, the logits depend on angular alignment and are invariant to uniform rescaling of the feature vector and to positive rescaling of any individual class-prototype row. This removes scalar norm-based compensation, but does not rule out coordinate-wise or anisotropic adaptation by the encoder. The ablation in Appendix A.1 shows that the cosine classifier adds 3.29–4.04 points over spectral suppression alone.

3.4.2 Spectral Alignment Suppression Regularizer

After each gradient update, SAS computes the SVD of the centered mini-batch feature matrix and attenuates classifier weights associated with feature dimensions that align strongly with the dominant spectral directions. Unlike conventional regularizers, SAS operates as a *post-gradient spectral regularizer*: it leaves the optimization objective unchanged and instead directly modifies the classifier weights after each parameter update.

Let $\mathbf{Z} \in \mathbb{R}^{N \times d}$ denote the centered feature matrix extracted from a mini-batch of size N, where d is the feature dimension. Its economy SVD is

$$\mathbf{Z} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \qquad k = \max(1, \lceil \tau_{\text{sas}} \min(N, d) \rceil), \tag{2}$$

where $\tau_{\text{sas}} \in (0, 1)$ controls the fraction of dominant singular directions retained. For the economy SVD, $\mathbf{V} \in \mathbb{R}^{d \times r}$ with $r = \min(N, d)$, and its columns are ordered by decreasing singular value.

Each feature coordinate $j \in \{1, \ldots, d\}$ is assigned its spectral leverage score in the top-k singular subspace,

$$\ell_j = \|\mathbf{V}_{j,\mathcal{T}}\|_2^2 = \mathbf{e}_j^\top \mathbf{P} \mathbf{e}_j, \qquad \mathbf{P} = \mathbf{V}_{:,\mathcal{T}} \mathbf{V}_{:,\mathcal{T}}^\top, \qquad \mathcal{T} = \{1, \ldots, k\}, \tag{3}$$

where $\mathbf{P}$ is the projector onto the top-k right-singular subspace. Larger values of ℓ_j indicate that coordinate axis $\mathbf{e}_j$ has greater overlap with the dominant spectral subspace (the unsquared score $s_j = \sqrt{\ell_j}$ gives the same ranking).

Let

$$\mathcal{H}_t = \{ j \ : \ \ell_j \geq \text{Percentile}\left(\boldsymbol{\ell}, 100(1 - \tau_{\text{sas}})\right) \}, \tag{4}$$

denote the set of highly aligned feature dimensions. Since the cosine classifier weight matrix is $\mathbf{W} \in \mathbb{R}^{C \times d}$, where each row corresponds to a class prototype and each column corresponds to one feature dimension, SAS suppresses the selected feature dimensions by attenuating the corresponding columns of $\mathbf{W}$,

$$\mathbf{W}_{:,j} \leftarrow \alpha \, \mathbf{W}_{:,j}, \qquad \forall \, j \in \mathcal{H}_t, \tag{5}$$

where $\alpha \in (0, 1)$ is the decay factor. The set $\mathcal{H}_t$ is recomputed independently for every mini-batch after the gradient update, allowing the selected feature coordinates to adapt dynamically throughout training. Importantly, this operation does not alter gradient computation or back-propagation through the encoder; it only modifies the classifier weights after the optimizer step.

The training objective therefore remains the standard cross-entropy loss,

$$\mathcal{L}^{\text{SAS}} = \mathcal{L}_{\text{CE}}\big(h(g(\mathbf{x})), y\big), \tag{6}$$

followed by the post-gradient spectral suppression step defined in Eq. (5).

Adaptivity. The dominant subspace and its leverage scores are re-estimated from every mini-batch, so $\mathcal{H}_t$ need not be constant or nested across steps. Theorem E.5 gives a perturbation condition under which both the estimated subspace and the selected coordinate set are stable.

Soft rather than hard suppression. The dominant subspace may contain both semantic and spurious information. SAS therefore attenuates, rather than deletes, selected classifier coordinates – a soft geometric prior that encourages the classifier to reduce reliance on coordinates concentrated in the dominant subspace, without guaranteeing that semantic information is unaffected.

3.5 Properties of Spectral Alignment Suppression

The formal properties below describe the operation implemented by Algorithm 1. Let $\mathbf{P}_t = \mathbf{V}_{t,k}\mathbf{V}_{t,k}^\top$ be the projector onto the top-k right-singular subspace and let $\ell_{t,j} = \mathbf{e}_j^\top \mathbf{P}_t \mathbf{e}_j$ be the leverage score of feature coordinate j. SAS selects the coordinates with the largest leverage scores and applies a diagonal post-gradient update to the classifier matrix $\mathbf{W} \in \mathbb{R}^{C\times d}$.

This update has four properties that match the implemented method. First, the selected set maximizes coordinate-axis coverage of the estimated dominant subspace among all sets of the same size. Second, SAS exactly contracts both the selected classifier columns and a leverage-weighted classifier-energy functional; under bounded optimizer updates, persistent selection keeps a coordinate norm below a finite floor. Third, the cosine classifier is invariant to uniform feature scaling and independent positive scaling of each class-prototype row. This removes scalar norm-based compensation, but not all possible anisotropic adaptation. Fourth, when the reference top-k subspace has a nonzero spectral gap, Wedin's theorem controls mini-batch subspace error and therefore leverage-score error. If that stable dominant subspace contains spurious signal, SAS contracts classifier coordinates with high coverage of a bias-risk region; if it is specifically bias-enriched, SAS is conditionally bias-targeting. Geometry alone, however, cannot determine that every high-variance direction is biased or separate bias from target information when they share a subspace. Full statements and proofs are given in Appendix E. Appendix F.2 evaluates the stability condition directly on mini-batches of real trained features.

4 Experimental Setup

Datasets. **Biased-MNIST:** digit class y is spuriously correlated with both foreground color b_1 and background color b_2 at correlation strengths $q \in \{0.90, 0.95, 0.99\}$.

CelebA (Liu et al., 2015): We use the CelebA dataset containing over 200K face images. The target classification task is *Gender*, while *WearingLipstick* and *HeavyMakeup* are treated as spurious (bias) attributes to evaluate robustness against attribute correlations.

MS-COCO (Lin et al., 2014): MS-COCO 2017, binary gender classification from image captions (Female vs. Male). Sports-object and kitchen-object co-occurrences define post-hoc evaluation groups only. These object categories are not used by SAS during training, validation splitting, hyperparameter selection, or checkpoint selection. Models are saved using ordinary average validation accuracy computed from the target labels only.

Baselines. **ERM** (lower bound); **ERM + Cosine** (cosine classifier, no spectral regularization); **Random Feature-Coordinate Suppression** (same fraction of feature coordinates suppressed at random; App. G).

Annotation-free two-step: **JTT** (Liu et al., 2021) trains an identification model with ERM, upweights misclassified samples, and retrains. **LfF** (Nam et al., 2020) jointly trains a deliberately biased model (generalized cross-entropy) and a debiased model reweighted by the biased model's relative per-sample loss. Neither observes bias labels, matching SAS's setting, but both need an auxiliary model and a two-stage (or dual-model) procedure rather than a single-pass geometric criterion.

Supervised: BC-BB (Hong & Yang, 2021), EnD (Tartaglione et al., 2021), FLAC (Sarridis et al., 2025), FairKL (Barbano et al., 2023), GMBM (Dwivedi et al., 2025). We do not claim uniform superiority over all supervised approaches.

All results for SAS, JTT, LfF, and the supervised baselines follow an identical protocol - same backbone, epoch budget, augmentation, optimizer, and splits - implemented within the official GMBM codebase.

Architecture and Implementation. Biased-MNIST uses a 4-layer CNN (16-32-64-128 channels) with global average pooling (128-d features). CelebA and MS-COCO use ResNet-18 (He et al., 2016) pretrained on ImageNet (512-d penultimate features) with a cosine classifier. We train with Adam (lr $= 10^{-4}$, weight decay 10^{-4}) for 80 epochs (Biased-MNIST), 50 (MS-COCO), 20 (CelebA). The exact SVD of the centered batch feature matrix is computed after every gradient step, no rank truncation. Full details and overhead are in Appendix B.

Table 2: **Biased-MNIST.** Unbiased test accuracy (%), reported as mean ± standard deviation over three seeds. **Bold**: best annotation-free result. Underline: best supervised result.

Setting	Method	$q = 0.90$	$q = 0.95$	$q = 0.99$
None	ERM (Vanilla)	81.5 ± 2.3	52.9 ± 1.04	23.5 ± 3.3
	ERM + Cosine	83.21 ± 2.10	59.47 ± 1.50	26.38 ± 1.50
	Random Suppression	82.80 ± 2.40	55.60 ± 1.60	26.20 ± 1.70
Annot.-free (2-step)	JTT (Liu et al., 2021)	85.42 ± 1.67	62.00 ± 4.36	33.90 ± 0.17
	LfF (Nam et al., 2020)	87.58 ± 0.25	64.23 ± 1.12	$\mathbf{38.48 \pm 1.06}$
Annot.-free (1-step)	SAS (Ours)	$\mathbf{88.22 \pm 0.40}$	$\mathbf{67.63 \pm 1.98}$	34.89 ± 0.92
Supervised	BC-BB (Hong & Yang, 2021)	80.90 ± 1.42	66.00 ± 2.11	40.90 ± 1.87
	EnD (Tartaglione et al., 2021)	82.50 ± 1.08	57.50 ± 1.76	25.70 ± 1.34
	FLAC (Sarridis et al., 2025)	84.40 ± 0.95	63.10 ± 1.53	32.40 ± 1.22
	FairKL (Barbano et al., 2023)	87.60 ± 0.81	61.60 ± 1.47	42.00 ± 1.09
	GMBM (Dwivedi et al., 2025)	$\underline{90.30 \pm 1.03}$	$\underline{70.50 \pm 0.89}$	$\underline{48.60 \pm 4.50}$

Evaluation Metrics. **UA (Unbiased Accuracy):** accuracy uniformly averaged across all bias-attribute groups, then across attributes, preventing larger groups from dominating.

BCA (Bias-Conflicting Accuracy): primary metric – accuracy on bias-conflicting samples only, averaged across attributes.

SBA (Scaled Bias Amplification), adopted from Dwivedi et al. (2025), measures amplification of existing dataset bias (lower is better); full definition and all per-dataset results are consolidated in Appendix D.

Computing UA, BCA, SBA requires bias-attribute labels on the test set (standard practice, also used by all baselines). SAS never accesses these labels during training. All main results are mean ± std over three seeds.

5 Results

5.1 Biased-MNIST

As bias severity increases, ERM collapses to 23.5% at $q = 0.99$. SAS reaches 88.22% at $q = 0.90$ (+6.72 over ERM), beating BC-BB, EnD, FLAC, and both annotation-free two-step baselines JTT (85.42%) and LfF (87.58%); it leads all annotation-free methods at $q = 0.95$ (67.63% vs. 62.00%/64.23%). Random feature-coordinate suppression (82.80% at $q = 0.90$) confirms the gain stems from SVD-guided regularization, not generic coordinate shrinkage.

At $q = 0.99$, LfF's dual-model mechanism reaches 38.48%, exceeding SAS (34.89%, +11.39 over ERM) and JTT (33.90%): when correlation is nearly deterministic, explicitly training against an intentionally biased proxy can better isolate the residual bias-conflicting signal than a single-pass geometric criterion, at the cost of a second stage and auxiliary model. SAS remains competitive with JTT here and still beats ERM, EnD, and FLAC.

GMBM shows the advantage of explicit bias supervision, exceeding all annotation-free methods at every severity tested. Corresponding SBA results for all three datasets are in Appendix D.

5.2 MS-COCO

SAS achieves the strongest overall performance on both biases, outperforming GMBM and the two-step baselines. It attains 83.77%/83.51% on Sports and 84.41%/82.60% on Kitchen, improving roughly 2–3 UA points over JTT and LfF. Relative to GMBM, SAS gains +1.35 UA on Sports and +0.42 UA on Kitchen while remaining competitive on BCA. The balanced spectral audit shows that the top-32 subspace retains most of the AUROC lift (95.2%/86.9%) and 98.5% of the target lift (Table 11), indicating that SAS softens high-leverage coordinates rather than deleting the dominant subspace. Corresponding SBA results are in Appendix D.

Table 3: **MS-COCO.** UA and BCA (%) for Sports and Kitchen biases, reported as mean ± standard deviation over three seeds. **Bold**: best annotation-free result. Underline: best supervised result.

		Sports Bias		Kitchen Bias	
Setting	**Method**	UA	BCA	UA	BCA
None	ERM	70.86±1.98	64.61±2.32	73.20±1.02	67.36±4.20
Annot.-free (2-step)	JTT (Liu et al., 2021)	80.50±0.95	77.15±1.26	83.17±1.35	81.25±2.15
	LfF (Nam et al., 2020)	80.19±0.24	77.01±2.25	82.19±0.83	81.28±1.23
Annot.-free (1-step)	SAS (Ours)	**83.77±0.74**	**83.51±1.37**	**84.41±0.63**	**82.60±1.01**
Supervised	FairKL (Barbano et al., 2023)	76.32±1.11	67.11±1.23	74.35±0.89	76.90±1.60
	EnD (Tartaglione et al., 2021)	81.77±1.85	79.45±0.52	82.47±1.00	81.23±2.21
	FLAC (Sarridis et al., 2025)	80.02±1.23	77.31±0.80	81.62±0.79	79.95±0.99
	GMBM (Dwivedi et al., 2025)	82.42±1.20	82.24±0.28	83.99±1.20	82.48±0.98

Table 4: **CelebA.** UA and BCA (%) for WearingLipstick and HeavyMakeup, reported as mean ± standard deviation over three seeds. **Bold**: best annotation-free result. Underline: best supervised result. (†) indicates the use of bias-attribute labels during training.

		WearingLipstick		HeavyMakeup	
Setting	**Method**	UA	BCA	UA	BCA
None	ERM	90.05±3.27	81.01±6.41	87.19±3.23	75.30±6.23
Annot.-free (2-step)	JTT (Liu et al., 2021)	92.32±0.74	86.40±0.98	88.56±1.02	82.89±1.39
	LfF (Nam et al., 2020)	93.11±0.15	88.48±1.74	90.06±1.71	84.28±4.31
Annot.-free (1-step)	SAS (Ours)	**94.95±0.47**	**90.72±0.90**	**92.52±0.95**	**85.82±1.87**
Supervised †	FairKL (Barbano et al., 2023)	82.70±1.34	74.70±1.89	84.40±1.21	77.90±1.76
	BC-BB (Hong & Yang, 2021)	91.60±0.82	85.80±1.14	89.70±0.93	81.80±1.27
	EnD (Tartaglione et al., 2021)	93.65±0.58	89.89±0.74	90.30±0.69	85.30±0.88
	FLAC (Sarridis et al., 2025)	95.40±0.43	89.60±0.62	89.98±0.57	87.20±0.71
	GMBM (Dwivedi et al., 2025)	95.38±0.11	90.84±0.07	89.60±0.22	82.40±0.30

5.3 CelebA

SAS attains 94.95%/90.72% on WearingLipstick and 92.52%/85.82% on HeavyMakeup without using bias labels. On WearingLipstick, SAS is within 0.45 UA points of the best supervised result from FLAC and 0.12 BCA points of the best supervised result from GMBM. On HeavyMakeup, SAS achieves the highest UA overall, while FLAC leads BCA by 1.38 points. SAS also surpasses both annotation-free two-step baselines on every metric: JTT reaches 92.32%/86.40% and 88.56%/82.89%, while LfF reaches 93.11%/88.48% and 90.06%/84.28%. SBA values are reported in Appendix D.

6 Conclusion

We introduced **Spectral Alignment Suppression** (SAS), a simple annotation-free regularizer that attenuates high-leverage coordinates in the dominant spectral subspace of each batch via SVD. No bias labels, auxiliary models, or multi-stage training are required. Experiments on Biased-MNIST, CelebA, and MS-COCO show substantial gains over ERM and consistent improvement over two-step baselines (JTT, LfF), with competitive or superior results against strong supervised methods on several subgroups. Dominant subspaces can be bias-enriched or shared; SAS therefore treats spectral dominance as a soft risk prior. While supervised methods retain an edge at extreme bias, SAS offers a lightweight, geometrically principled alternative for annotation-free debiasing.

ETHICS STATEMENT

SAS is annotation-free with respect to bias attributes during training. Bias labels are used only for evaluation and (on Biased-MNIST) validation via the known parameter q. For MS-COCO, selection uses only average target accuracy. Deployment in high-stakes settings still requires domain-specific auditing.

REPRODUCIBILITY STATEMENT

All experiments use public datasets and standard architectures. Pseudocode is given in Algorithm 1; proofs appear in Appendix E. Results are averaged over three seeds with standard deviations. Hyperparameters, baseline reimplementations, and code will be released upon acceptance.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, LaTeX formatting, and consistency checking. All AI-assisted changes were reviewed by the authors, who take responsibility for the final content of the paper, including its text, claims, and artifacts. This statement should be expanded before submission if generative AI tools were used for any additional research tasks.

REFERENCES

Carlo Alberto Barbano, Benoit Dufumier, Enzo Tartaglione, Marco Grangetto, and Pietro Gori. Unbiased supervised contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Peter L. Bartlett, Dylan J. Foster, and Matus J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Rajeev Ranjan Dwivedi, Ankur Kumar, and Vinod Kumar Kurmi. Multi-attribute bias mitigation via representation learning. In *ECAI 2025*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pp. 1221–1228. IOS Press, 2025. doi: 10.3233/FAIA250935.

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 (CVPR)*, pp. 770–778, June 2016.

Youngkyu Hong and Eunho Yang. Unbiased classification through bias-contrastive and bias-balanced learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 26449–26461. Curran Associates, Inc., 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

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.

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 (ICCV)*, pp. 3730–3738, December 2015.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Debiasing classifier from biased classifier. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20673–20684. Curran Associates, Inc., 2020.

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.

Ioannis Sarridis, Christos Koutlis, Symeon Papadopoulos, and Christos Diou. FLAC: Fairness-aware representation learning by suppressing attribute-class associations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):1148–1160, February 2025. doi: 10.1109/TPAMI.2024.3487254.

Enzo Tartaglione, Carlo Alberto Barbano, and Marco Grangetto. EnD: Entangling and disentangling deep representations for bias correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13508–13517, June 2021. doi: 10.1109/CVPR46437.2021.01330.