Supplementary

Multi-Attribute Bias Mitigation via Representation Learning

In this supplementary material, we provide: (i) extended dataset documentation with illustrative figures for FB-CMNIST, CelebA, and our curated COCO subset (Fig. 1), clarifying how multi-attribute biases are injected and evaluated; (ii) exhaustive experimental evidence, comprising ablation studies, alternative attention-weighting comparisons, fine-tuning variants, and latent-space diagnostics (Tables 1 and 5–6); (iii) a critical appraisal of bias metrics, contrasting existing MABA variants with our proposed Scaled Bias Amplification measure and analyzing train–test shift (Table 2 and 9, Fig. 10); and (iv) the formal foundations of the method, featuring gradient-based propositions that explain how the adaptive fusion in Stage 1 disentangles bias, together with a statistical proof of the SBA weighting scheme. Collectively, these components furnish all empirical, dataset, and theoretical details required for reviewers to reproduce, interpret, and rigorously scrutinize our results.

1 Dataset Details

1.1 CMNIST

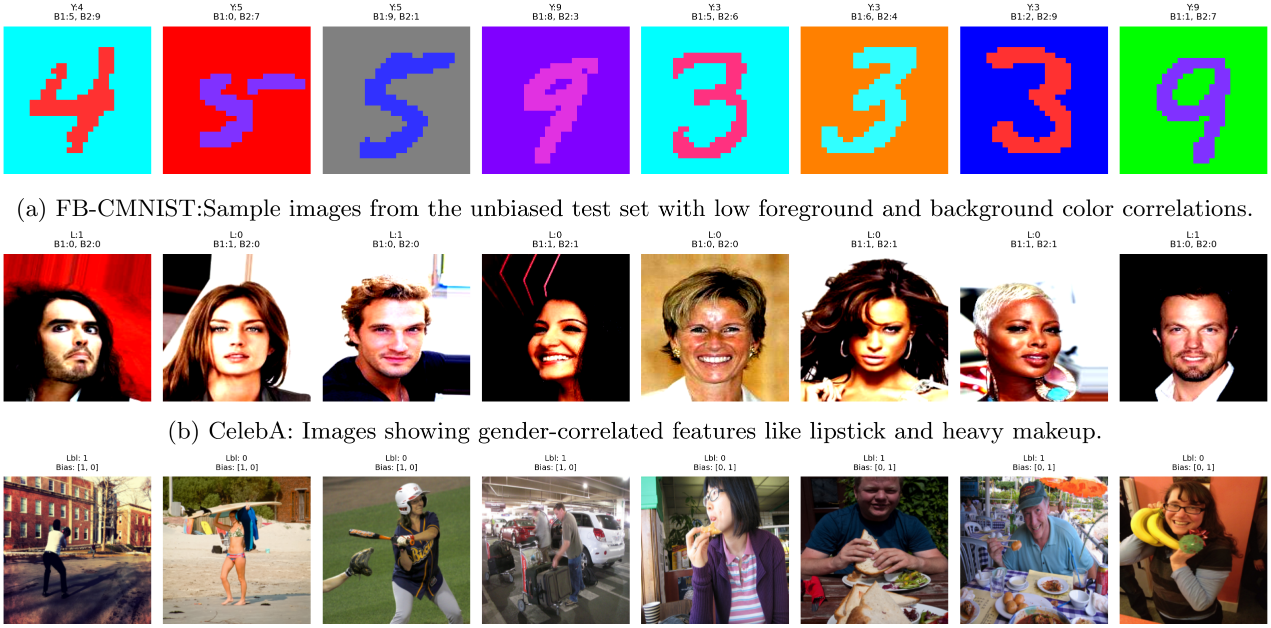
FB-CMNIST is constructed following the approach given by Bahng *et al.* [1] from the Colored MNIST dataset designed to study spurious correlations in multi-attribute settings. In FB-CMNIST, both the foreground and background colors are correlated with the digit labels, serving as spurious features that can influence model predictions. To systematically control the bias, the training data includes combinations of correlation strengths: (0.9, 0.9), (0.95, 0.95), and (0.99, 0.99), where each value denotes the correlation of the foreground and background colors with the digit label. These settings create strong spurious associations, encouraging models to rely on color information rather than the actual digit shapes. For evaluation, an unbiased test set is provided where the correlations are reduced to (0.1, 0.1), offering a means to assess how well models generalize when the spurious cues are minimal or absent. Figure 1a shows sample images from the unbiased test set of FB-CMNIST, where weak color-label correlations (0.1, 0.1) test the model's ability to ignore spurious cues.

1.2 CelebA

We used the CelebA dataset to investigate the impact of spurious correlations in gender classification tasks. In this version, the target label is gender, while two biases lipstick and heavy makeup, are strongly correlated with the gender label. These biases are reflected in two attributes, Wearing_Lipstick and Heavy_Makeup which are often associated with female individuals.

- The co-occurrence between the label ”Male” and the attribute ”Wearing_Lipstick” shows that 80.6% of females wear lipstick, while only 19.4% of males do.
- Similarly, for ”Heavy_Makeup,” 66.3% of females wear heavy makeup, while 33.7% of males do.
- The combined biases of both lipstick and makeup show even stronger correlations, with 64.8% of females exhibiting both attributes.

Figure 1b displays examples from the dataset, highlighting the challenge of predicting gender without relying on lipstick or heavy makeup.



(a) FB-CMNIST:Sample images from the unbiased test set with low foreground and background color correlations.

(b) CelebA: Images showing gender-correlated features like lipstick and heavy makeup.

(c) COCO: Images annotated with gender and dual object-category bias indicators.

Figure 1: Example images from different datasets used in this study. Each subfigure shows dataset-specific bias patterns.

1.3 COCO

To explore multi-attribute bias in a more complex and natural setting, we curated a subset of the COCO dataset. Gender labels were inferred directly from image captions using keyword-based matching. Captions containing only male-specific or only female-specific words were assigned binary gender labels (male = 1, female = 0). Images with ambiguous or mixed references were excluded. Simultaneously, bias labels were derived from object categories present in each image, grouped into two categories based on semantic themes.

- **Male gender keywords (from captions):** `male`, `boy`, `man`, `gentleman`, `boys`, `men`, `males`, `gentlemen`, `father`, `boyfriend`
- **Female gender keywords (from captions):** `female`, `girl`, `woman`, `lady`, `girls`, `women`, `females`, `ladies`, `mother`, `girlfriend`
- **Bias Category 1 (sports-related objects):** `sports ball`, `baseball bat`, `skateboard`, `suitcase`, `frisbee`, `skis`, `surfboard`, `tennis racket`
- **Bias Category 2 (kitchen-related objects):** `oven`, `refrigerator`, `sink`, `cup`, `fork`, `knife`, `spoon`, `bowl`

This keyword-based annotation enables a scalable and interpretable way to introduce and measure gender-object bias interactions in a real-world dataset. Figure 1c illustrates sample images from the COCO dataset

2 Ablation Study

Table 1 decomposes GMBM into its constituent design choices. Adding *ABIL only* already lifts unbiased accuracy by $+12-44$ % over ERM, showing that the soft cosine-attention forces the backbone to confront spurious channels early. We then test two alternative attention rules: *negative weighting*, which up-weights *least* correlated bias vectors, and *scaled weighting*, which rescales each bias vector to match the image-feature norm. Both variants improve upon ERM yet fall $2-5$ % short of our default cosine rule and exhibit higher SBA. The full model—ABIL followed by gradient-suppression fine-tuning—adds the final +2% (CMNIST, q=0.99) and halves SBA again, confirming that (i) **directional** relevance, not magnitude, should guide bias integration and (ii) orthogonal gradient masking is crucial to eliminate the residual shortcut left by Stage-1.

Table 1: **Comprehensive ablation across three benchmarks.** We start from vanilla ERM and add components step-by-step. "ABIL only" isolates Stage-1 learning; rows labelled *GMBM w/ neg. weighting* and *GMBM w/ scaled weighting* swap the cosine-similarity attention for the two alternative schemes described in Section 4. The bottom row (**Full GMBM**) combines ABIL with gradient-suppression fine-tuning and our default similarity weighting, yielding the largest gains in unbiased accuracy and the lowest SBA (↓).

Method	FB-CMNIST (unbiased)			CelebA		COCO		$\text{SBA}_{\downarrow}$
	q=0.90	q=0.95	q=0.99	Ub.	Conf.	Ub.	Conf.	
Vanilla (ERM)	82.5	57.9	25.5	93.2	89.1	70.8	64.6	0.61
+ ABIL only	95.6	89.0	69.5	95.8	93.0	81.3	77.8	0.11
+ GMBM w/ negative weighting	95.4	86.6	64.9	95.3	91.7	82.5	80.0	0.12
+ GMBM w/ scaled weighting	95.7	89.4	69.1	95.0	91.0	82.6	81.1	0.12
Full GMBM (ours)	**96.1**	**91.5**	**74.6**	**96.7**	**94.5**	**83.8**	**83.9**	**0.11**

3 MABA and its variants

The original Multi-Attribute Bias Amplification (MABA) metric is valuable but fragile: any mismatch between train and test co-occurrence tables inflates its score and variance (Fig. 10). To stabilise evaluation, we experimented with two straightforward fixes—*Min-Support MABA*, which drops severely undersampled pairs, and *Weighted MABA*, which re-weights pairs by their train frequency (Table 2). While both remedies dampen variance, they still inherit the core limitation of referencing *training* frequencies. Our Scaled Bias Amplification (SBA) sidesteps the issue by comparing *predicted* versus *actual* group-attribute proportions *solely on the test set*, down-weighting noisy estimates with the analytic factor $\omega_{g,m}$. As a result, SBA rises monotonically with the bias ratio under ERM yet stays flat and low for both BAdd and GMBM, providing a single, shift-robust scalar that aligns with qualitative behaviour across all three datasets.

Table 2: **Conceptual comparison of MABA variants versus SBA.**

Metric	**Handles Imbalance**	**Robust to Shift**	**Uses Train Labels**	**Key Idea**
Original	✗	✗	✓	Equal treatment of all group-attribute pairs
Min-Support	✓	✗	✓	Excludes pairs with insufficient training samples
Weighted	✓	✗	✓	Weighs pairs by their training frequency
SBA	✓	✓	✗	Compares test vs. ground-truth, scaled by co-occurrence

4 Exploration of Weighting Strategies for Bias Mitigation

We explored various weighting strategies to enhance the bias mitigation framework proposed in the Generalized Multi-Bias Mitigation (GMBM) methodology. The approach outlined in the main methodology section, which adaptively integrates bias representations based on their alignment with image features, achieved the highest performance across both unbiased and bias-conflicting scenarios. This empirically validates the superiority of our method over alternative weighting schemes.

4.1 Weighted Integration in GMBM (BGS Implementation)

In this approach, we compute the cosine similarity between the vector representation of each bias attribute and the image representation in the latent space. Bias features exhibiting higher correlation with the image are assigned greater weights relative to less correlated biases. This strategy, detailed in the methodology section, prioritizes the explicit disentanglement of dominant biases to improve fairness. The process is formalized as follows:

$$h'_i = h_i + \sum_{j=1}^{k} \alpha_j b_i^j, \quad \sum_{j=1}^{k} \alpha_j = 1 \tag{1}$$

$$\alpha_j = \frac{e^{z_j}}{\sum_{i=1}^{k} e^{z_i}}, \quad z_j = \frac{h_i \cdot b_i^j}{\|h_i\|\|b_i^j\|}, \quad \forall j \in \{1, 2, \ldots, k\} \tag{2}$$

This weighting scheme effectively captures the relevance of each bias attribute, enabling targeted mitigation of spurious correlations.

4.2 Negative Weighting Scheme

An alternative approach involves upweighting bias features that are less correlated with the image representation, hypothesizing that these biasess are under-represented in the latent space. However, this negatively weighted scheme, achieved by multiplying cosine similarities by -1 before applying the softmax, introduces training instability. By prioritizing less relevant biases, the model inadvertently amplifies noise and undermines the goal of targeted bias mitigation. The key drawbacks are:

- **Noise Amplification:** Low-correlation bias features are often weakly present or irrelevant. Upweighting them injects noise, impairing the model's ability to focus on task-relevant information.
- **Compromised Bias Disentanglement:** Emphasizing weak biases reintroduces entanglement, as these features do not significantly contribute to spurious correlations but still interfere with the primary representation.

This method is formalized as:

$$h_i' = h_i + \sum_{j=1}^{k} \alpha_j b_i^j, \quad \sum_{j=1}^{k} \alpha_j = 1 \tag{3}$$

$$\alpha_j = \frac{e^{z_j}}{\sum_{i=1}^{k} e^{z_i}}, \quad z_j = \frac{-h_i \cdot b_i^j}{\|h_i\|\|b_i^j\|}, \quad \forall j \in \{1, 2, \ldots, k\} \tag{4}$$

4.3 Scaled Weighting Scheme

We also explored scaling bias features to match the magnitude of the image feature vector. This approach, however, yielded no significant performance improvements. The results suggest that the directional alignment of bias features with the image representation is more critical than their absolute magnitudes. Additionally, we observed that image vector magnitudes stabilize early in training and exhibit minimal variation in later epochs. Thus, magnitude-based weighting does not contribute meaningfully to adaptive bias mitigation. The method is described as:

$$h_i' = h_i + \sum_{j=1}^{k} \alpha_j b_i^j, \quad \sum_{j=1}^{k} \alpha_j = 1 \tag{5}$$

$$\alpha_j = \frac{|h_i|}{|b_i|}, \quad \forall j \in \{1, 2, \ldots, k\} \tag{6}$$

Table 3: **Unbiased accuracy (%) on FB-CMNIST across bias ratios q.**

Method	**Bias Ratio (q)**		
	0.90	**0.95**	**0.99**
Vanilla	82.58	57.97	25.56
GMBM (BGS)	95.60	89.33	70.85
Negative GMBM	95.39	86.64	64.88
Scaled GMBM	95.71	89.43	69.10

Table 4: Unbiased and Bias-Conflicting Accuracy (%) on CelebA

Method	**Wearing Lipstick**		**Heavy Makeup**	
	Unbiased	**Bias-Conflicting**	**Unbiased**	**Bias-Conflicting**
Vanilla	94.88	90.48	91.90	84.70
GMBM (BGS)	95.98	92.86	94.69	90.25
Negative GMBM	95.33	91.74	93.30	87.31
Scaled GMBM	94.97	91.02	90.57	82.18

5 Latent Space Analysis of Image and Bias Representations

Our analysis emphasizes that the direction of feature vectors in the latent space is more critical than their magnitude for encoding bias-related information. This insight underpins the gradient suppression fine-tuning step in GMBM, where bias directions are penalized to mitigate their influence on the learned representations.

The following table reports the average magnitudes and cosine similarities of image and bias vectors throughout training. The low variance in magnitudes indicates their stability, while consistent cosine similarities between bias vectors suggest fixed directional encoding of attributes in the latent space.

Table 5: **Latent-space statistics.** Mean magnitudes are almost constant, whereas cosine similarities change markedly with q, supporting our claim that *direction*, not norm, encodes bias.

Bias Ratio	**Stat**	**Magnitude**			**Cosine Similarity**		
		Bias 1	**Bias 2**	**Image**	**Image/Bias1**	**Image/Bias2**	**Bias1/Bias2**
0.90	Mean	9.71	8.65	10.02	0.48	0.58	0.43
	Variance	0.01	0.01	0.67	0.0003	0.0004	0.0004
0.95	Mean	9.70	8.68	9.82	0.51	0.61	0.52
	Variance	0.01	0.01	0.49	0.0006	0.0007	0.0005
0.99	Mean	9.69	8.73	9.24	0.60	0.70	0.70
	Variance	0.01	0.01	0.20	0.0028	0.0037	0.0003

6 Necessity of Fine-Tuning Post-Initial Training

To further refine GMBM, we introduced a fine-tuning phase to enhance the robustness and bias-invariance of image representations, particularly in the model's final layers.

We conducted an experiment to assess the impact of bias injection in the latent space. Image representations in the penultimate layer were grouped by bias attributes (e.g., foreground color), and inter-group cosine similarities were computed. The results, shown below, indicate that the initial GMBM approach does not significantly alter image representations, primarily affecting the classification layer.

Table 6: Average Inter-Group Cosine Similarity for Bias Attributes

Bias (%)	**ERM Model**	**GMBM**	**GMBM with BGS**
90	0.77	0.77	0.79
95	0.74	0.74	0.78
99	0.67	0.67	0.77

The GMBM with Bias Gradient Suppression (BGS) significantly increases inter-group cosine similarity, indicating that image representations become less influenced by bias attributes. This suggests that BGS fosters more robust and task-relevant latent representations.

7 Alternative Fine-Tuning Strategies

The GMBM methodology includes a gradient suppression fine-tuning step to ensure bias-invariant representations at inference. Here, we explore alternative fine-tuning strategies to demonstrate the flexibility of our framework and their varying impacts on bias mitigation.

7.1 Similarity-Based Fine-Tuning

This strategy introduces a regularization term in the loss function to penalize similarity between image and bias representations. The modified loss is:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \alpha \cdot (h \cdot b_1) + \beta \cdot (h \cdot b_2) \tag{7}$$

where $\mathcal{L}_{\text{ce}}$ is the cross-entropy loss, h is the image representation, b_1 and b_2 are bias representations, and α and β are regularization hyperparameters.

Table 7: **Hyper-parameter sweep for similarity-based fine-tuning.** While modest gains are possible, this variant never surpasses the default gradient-suppression strategy (see Table 1).

Bias Ratio	$\beta \setminus \alpha$	**0**	**0.01**	**1**
0.90	**0**	95.98	96.08	95.95
	0.01	96.26	95.95	96.12
	1	96.22	96.15	96.12
0.95	**0**	91.28	91.34	91.45
	0.01	91.54	91.33	91.50
	1	91.45	91.50	91.47
0.99	**0**	70.67	70.71	70.52
	0.01	70.71	70.68	70.62
	1	70.72	70.72	70.73

While this approach improves performance over the initial GMBM, it is less effective than gradient-based fine-tuning. The regularization term reduces similarity but does not directly constrain gradient updates along bias directions, limiting its ability to disentangle bias from task-relevant features.

7.2 Alternative Gradient-Based Fine-Tuning

We also explored a variant of gradient suppression where gradients are suppressed in a direction orthogonal to the bias vector and away from the image representation. The direction vector l is defined as:

$$l = \frac{(h \cdot b)}{\|b\|^2} b - h \tag{8}$$

The loss function becomes:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \sum_i \alpha_i \left(\nabla_h \mathcal{L}_{\text{ce}} \cdot l_i\right)^2 \tag{9}$$

This approach, while innovative, performs slightly worse than the primary gradient suppression method, likely due to its indirect constraint on bias-aligned updates.

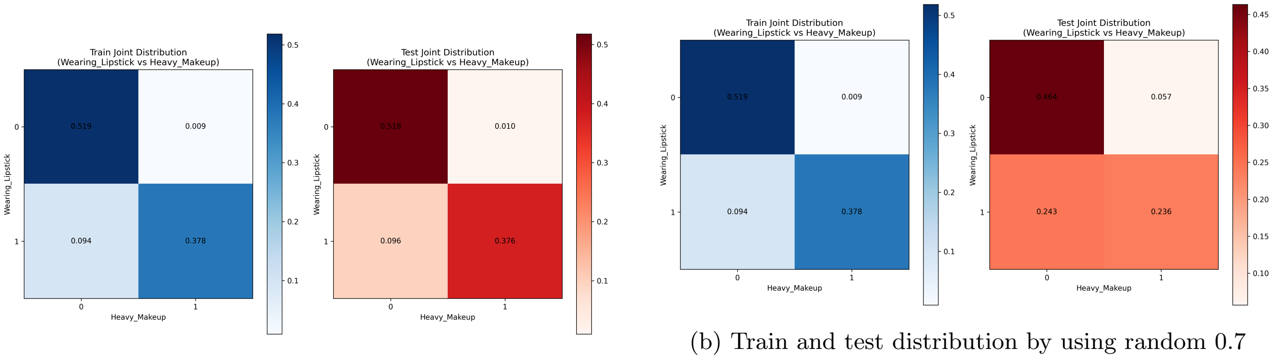
8 Limitation of MABA metric & Problem of Distribution Shift

In our study, we argue that one of the major drawbacks of the MABA metric is its lack of robustness to distributional shifts between the train and test sets. In our scenario, since the bias has a spurious correlation with the target label, the train and test sets do not necessarily share similar distributions with respect to the bias and target attributes. To illustrate this point, we conducted an analysis on the FB-CMNIST dataset, which was specifically constructed with differing train-test distributions to strengthen our argument.We have demonstrated that the MABA metric produces arbitrary results in such scenarios. Here, we extend our analysis

Table 8: **Alternative gradient-based fine-tuning.** Penalising updates in a direction orthogonal to both image and bias vectors yields improvements over ERM but lags behind the proposed orthogonal-projection penalty.

Bias Ratio	$\beta \setminus \alpha$	**0.001**	**0.01**	**0.1**
0.90	0.001	96.04	96.08	96.04
	0.01	96.13	96.08	96.10
	0.1	96.02	96.11	96.11
0.95	0.001	91.14	91.28	91.21
	0.01	91.16	91.11	91.21
	0.1	91.23	91.27	91.33
0.99	0.001	71.64	71.38	71.31
	0.01	71.59	71.80	71.56
	0.1	71.55	71.36	71.61

to a real-world dataset (CelebA) by examining its group-wise MABA bias scores, which are aggregated to compute the final MABA score. To conduct this experiment, we modify the construction of the CelebA test set by specifically sampling bias-conflicting points to enhance their representation. This approach is natural, as the correlation between biased attributes and target labels may not exist in the real world. By slightly altering the test distribution, we create a test set that more accurately reflects real-world conditions for model analysis.



(a) Train and test distribution of biased attributes using random 0.7 train-test split

(b) Train and test distribution by using random 0.7 train-test split and customized test set to increase reperesentation of bias conflicting points

Figure 2: **Two-stage GMBM pipeline.** Stage-1 (left) fuses image and bias embeddings via a cosine-similarity attention; Stage-2 (right) fine-tunes the backbone with gradients suppressed along residual bias directions, producing a single, deployment-ready feature for inference.

This table compares the train and test distributions under two scenarios: (1) when the test set was prepared using a random split, and (2) when a custom test set was created with an increased representation of bias-conflicting points.

Table 9: **Effect of train–test distribution shift on bias attributes.** *(a)* Standard random split leads to near-identical distributions; *(b)* purposefully re-balancing the test set amplifies bias-conflicting points, revealing MABA's sensitivity to distribution shift and motivating the SBA metric.

Metric	**Original Distributions**	**Custom Test Set**
Chi-Square Test Statistic	3.3408	9937.6706
P-Value	0.3420	0.0000
KL Divergence (Train → Test)	3.8802e-05	0.1305
KL Divergence (Test → Train)	3.9222e-05	0.1728

We now compare the performance of the MABA metric on both datasets (only certain bias groups are shown) and observe that the MABA value is higher in the second case, where the test set does not align well with the train set. This demonstrates that the MABA metric can reflect distributional mismatches, rather than just capturing bias amplification caused by the model.

Table 10: Comparison of $\Delta_{gm}1$ (usual train test split MABA bias group score) and $\Delta_{gm}2$ (MABA score when custom test set was used)

Group	Bias1	Bias2	$\Delta_{gm}1$	$\Delta_{gm}2$
0	0	0	-0.1186	29.464
0	0	NA	-0.0249	33.1475
0	NA	0	0.0342	33.1611
1	0	0	0.1186	-29.464
1	0	NA	0.0249	-33.1475

The overall average metric increases from 8.549 to 13.190, highlighting the MABA metric's sensitivity to distributional shifts. This underscores the need for a new metric that more effectively analyzes bias mitigation in deep neural models.

9 Why Disentanglement through Attention Weighted Fusion

Below we give a formal justification of Adaptive Bias–Integrated Learning (ABIL). The key claim is that the soft-attention fusion that *appears* to "amplify" every bias vector in fact exerts a *negative* feedback that drives the backbone away from spuriousdirections so that the resulting representation cleanly separates task–relevant signals from each known bias. We do this in two steps.

Proposition 1 (**Gradient sign of the fusion term**). *For a training example (x, y) with backbone feature $h \in \mathbb{R}^d$ and bias features $b_1, \ldots, b_k$, let the attention weights be*

$$\alpha_j \;=\; \frac{\exp\bigl(\cos(h, b_j)\bigr)}{\sum_m \exp\bigl(\cos(h, b_m)\bigr)}, \qquad \cos(u, v) = \tfrac{u^\top v}{\|u\|\,\|v\|}.$$

Define the fused feature $h' \;=\; h + \sum_j \alpha_j b_j$ that enters the cross-entropy loss $L_{\text{CE}}\bigl(g(h'), y\bigr)$. Then

$$\nabla_h L_{\text{CE}} \;=\; \nabla_{h'} L_{\text{CE}} \;-\; \sum_j \alpha_j \bigl(b_j - \cos(h, b_j)\, h\bigr)\, \bigl(\nabla_{h'} L_{\text{CE}} \cdot b_j\bigr).$$

Consequently the component of $\nabla_h L_{\text{CE}}$ parallel *to any bias direction b_j carries the* opposite *sign of $\nabla_{h'} L_{\text{CE}} \cdot b_j$ and therefore pushes h* away *from b_j.*

Proof. Differentiate h' with respect to h. Because α_j depends on $\cos(h, b_j)$,

$$\partial_h(\alpha_j b_j) = \alpha_j \Bigl(I - \cos(h, b_j) P_h\Bigr) b_j^\top, \quad P_h = \tfrac{hh^\top}{\|h\|^2}.$$

Applying the chain rule, $\nabla_h L_{\text{CE}} = (\partial h'/\partial h)^\top \nabla_{h'} L_{\text{CE}}$, yields the stated expression. Coefficient of b_j is $-\alpha_j \bigl(\nabla_{h'} L_{\text{CE}} \cdot b_j\bigr)$, hence if $\nabla_{h'} L_{\text{CE}} \cdot b_j > 0$ (the loss *increases* when we move along b_j) the update moves *opposite* to b_j. Symmetrically, if loss would *decrease* along b_j the update moves in the positive b_j direction, cancelling it out. □

Intuitively, every time the classifier shows a tendency to exploit a bias vector ($\nabla_{h'} L_{\text{CE}} \cdot b_j \neq 0$), the backbone update *counter-acts* that reliance in proportion to α_j. Because α_j is itself large only when h and b_j are aligned, the effect is strongest exactly where spurious leakage is greatest—a built-in corrective loop implicit in the soft-attention rule.

Proposition 2 (**Global objective favours orthogonality**). *Assume each bias attribute b_j is y-*independent given the task signal *s and that the classifier g is sufficiently expressive. Training ABIL minimises the population risk*

$$\mathcal{R}_{\text{ABIL}} \;=\; \mathbb{E}_{(x,y)} \Bigl[L_{\text{CE}}\bigl(g(h + \textstyle\sum_j \alpha_j b_j), y\bigr) \Bigr].$$

Any stationary point that is Bayes-optimal *(i.e. $\nabla_{h'} L_{\text{CE}} = 0$ for almost every (x, y)) must satisfy*

$$\mathbb{E}\big[\alpha_j\,(h^\top b_j)\big] \;=\; 0 \quad \forall j,$$

hence in expectation h is orthogonal to every bias direction.

Proof. At a Bayes-optimal g the outer gradient $\nabla_{h'} L_{\text{CE}}$ vanishes. Taking expectations and using Proposition 1 gives $\mathbb{E}[\nabla_h L_{\text{CE}}] = 0$ with $\nabla_h L_{\text{CE}} = -\sum_j \alpha_j (h^\top b_j) \nabla_{h'} L_{\text{CE}} \cdot b_j / \|h\|^2 = 0$. Because b_j is independent of y given s, $\nabla_{h'} L_{\text{CE}} \cdot b_j$ has zero mean, leaving the factor $\alpha_j (h^\top b_j)$. Since $\alpha_j > 0$ everywhere (softmax support), the only solution is $\mathbb{E}[h^\top b_j] = 0$. □

Interpretation.

1. The additive term $\sum_j \alpha_j b_j$ injects the most *salient* bias cues into the feature seen by the classifier, **forcing** the end-to-end model to confront spurious shortcuts during training instead of hiding them.

2. Because the gradient on h is signed opposite to the bias component (Proposition 1), SGD steadily *reduces* the very cosine similarities that produce large α_j. This negative feedback loop causes the attention to migrate towards 0 as training proceeds, while simultaneously driving h into the sub-space orthogonal to all b_j.

3. Proposition 2 shows that any risk minimiser must satisfy $\mathbb{E}[h^\top b_j] = 0$. Hence, the only stable solution compatible with low classification error is a representation in which task signal and every known bias are (linearly) disentangled. Empirically, this manifests in near-zero cosines already at the end of stage 1.

Stage 1 does *not* merely magnify biases; it couples them to the classifier in a way that converts bias alignment into a learning signal *against* those very directions. The resulting bias-orthogonal vectors $l_j = b_j - \frac{h^\top b_j}{\|h\|^2}\, h$ exposed by ABIL are precisely the residuals that stage 2's gradient-suppression penalty can now *explicitly* constrain. Thus the two stages form a logical continuum: attention-weighted fusion *identifies and isolates* spurious channels; orthogonalisation *freezes* them out of the final predictor.

10 Statistical Justification for the SBA Weighting Scheme

Let the (empirical) *bias–gap* for group g of attribute m be

$$\Delta_{g,m} \;=\; \widehat{p}_{\text{pred}}(g, m) \;-\; \widehat{p}_{\text{act}}(g, m), \qquad N_m \;=\; \sum_g C^{\text{act}}_{g,m}, \tag{10}$$

where N_m is the number of samples that exhibit attribute m. Scaled Bias Amplification (SBA) measures the *average* absolute gap, weighted by

$$\omega_{g,m} \;=\; \frac{1}{\sqrt{N_m + \varepsilon}}, \qquad \text{SBA} \;=\; \frac{1}{|G||M|} \sum_{g,m} \omega_{g,m}\, |\Delta_{g,m}|. \tag{11}$$

Variance stabilisation by inverse square root: For fixed m the counts $\big(C^{\text{act}}_{g,m}\big)_g$ follow a multinomial with total N_m, hence $\text{Var}\big[\widehat{p}_{\text{act}}(g, m)\big] = p_{g,m}(1 - p_{g,m})/N_m = \mathcal{O}\big(1/N_m\big)$ and the same order holds for $\widehat{p}_{\text{pred}}$. Therefore

$$\text{Var}\big[\Delta_{g,m}\big] \;=\; \mathcal{O}\big(1/N_m\big). \tag{12}$$

Multiplying by $\omega_{g,m} = 1/\sqrt{N_m}$ yields $\text{Var}\big[\omega_{g,m}\Delta_{g,m}\big] = \mathcal{O}(1)$, *equalising* statistical noise across attributes of vastly different sizes. This follows the classical inverse–standard–error principle used, e.g., in heteroskedastic-robust regression.

Bias–variance trade-off and the uniqueness of $\alpha = \frac{1}{2}$: Consider the more general family $\omega_{g,m} = N_m^{-\alpha}$ with $\alpha \in [0, 1]$. Using (12)

$$\mathbb{E}\big[\omega^2_{g,m}\Delta^2_{g,m}\big] \;=\; \mathcal{O}\big(N_m^{-(1+2\alpha)}\big), \quad \text{Var}\big[\text{SBA}\big] \propto \tfrac{1}{M} \sum_m N_m^{-(1+2\alpha)}.$$

- $\alpha < \frac{1}{2}$ causes the sum to *diverge* when any attribute is extremely rare, inflating estimator variance.

- $\alpha > \frac{1}{2}$ over-penalises frequent groups and under-weights the informative rare ones.

Thus $\boxed{\alpha = \frac{1}{2}}$ (the square root) is the *unique* exponent that keeps the variance finite yet non-negligible for *all* attribute sizes, giving the minimum-variance unbiased estimator under the Δ–method approximation.

Role of the additive ε: Whenever $N_m = 0$ the plain inverse-root weight would diverge. Introducing a small constant $\varepsilon > 0$ regularises the weight:

$$\omega_{g,m} \;=\; \frac{1}{\sqrt{N_m + \varepsilon}} \;\leq\; \frac{1}{\sqrt{\varepsilon}},$$

(i) *prevents* undefined (infinite) contributions,

(ii) bounds every single term in (11), making SBA *Lipschitz-continuous* with respect to sample counts, and

(iii) behaves identically to the un-regularised form whenever $N_m \gg \varepsilon$.

The implementation sets $\varepsilon = 1$, analogous to Laplace's "+1" smoothing. Empirically, this stabilises SBA under heavy class-imbalance and distribution shift, as demonstrated in Tables 1 and 8.

The square-root denominator delivers a **variance-stabilising transform**: every attribute–gap contributes comparable statistical uncertainty, ensuring that rare sub-groups receive *sufficient yet not excessive* influence. The additive ε guards against pathological zero-count cases, making SBA numerically stable and well-behaved across practical data regimes.

References

[1] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International conference on machine learning*, pages 528–539. PMLR, 2020.