
Supplementary Material: Language-Bias-Resilient Visual Question Answering via Adaptive Multi-Margin Collaborative Debiasing

Huanjia Zhu

Beijing Institute of Technology, Zhuhai
bvyih3@gmail.com

Shuyuan Zheng

University of Osaka
zheng@ist.osaka-u.ac.jp

Yishu Liu

Harbin Institute of Technology, Shenzhen
liuyishu@stu.hit.edu.cn

Sudong Cai*

Beijing Institute of Technology, Zhuhai
caisudong.ai@gmail.com

Bingzhi Chen*

Beijing Institute of Technology, Zhuhai
chenbingzhi@bit.edu.cn

1 Related Works

1.1 Language Biases.

Language bias in VQA manifests when models exploit question–answer priors instead of genuine visual signals [13, 1, 6, 15, 16, 3, 41, 29, 35]. Goyal et al. [13] first quantified this deficiency by constructing VQA v2, demonstrating dramatic performance drops once language shortcuts were removed. Agrawal et al. [1] then introduced VQA-CP, deliberately misaligning train/test answer distributions to expose distributional biases. More recently, GGE [15] showed through controlled experiments that distributional bias and shortcut bias constitute complementary facets of language bias. However, despite these advances, the formation mechanisms of language bias remain unexamined. In the following sections, we investigate these mechanisms in depth and develop a novel debiasing method informed by our findings.

1.2 Margin Learning.

Sophisticated margin losses have been widely explored across various domains such as deep face recognition [5, 12, 27, 36], class imbalance learning [7, 20], and VQA [14, 3, 41]. CosFace [36] introduces cosine loss in angular space to enhance discriminative features. ElasticFace [5] proposes an elastic penalty margin by sampling random margins from a normal distribution to achieve flexible class separability. Inspiring, AdaVQA [14] firstly incorporates adaptive cosine into VQA to effectively separate answer embeddings and suppress language biases. RMLVQA [3] further incorporates instance-level difficulty to adjust decision boundaries dynamically. Despite these advances, a fundamental question remains unaddressed: Why do margin mechanisms effectively mitigate bias? In the following sections, we unpack this question through theoretical analysis and empirical validation, and introduce a novel framework informed by our insights.

*Corresponding authors: Bingzhi Chen and Sudong Cai.

1.3 Difficulty Model.

Quantifying sample difficulty or mining hard samples has always been a highly anticipated challenge. Inspired by human cognitive learning, curriculum learning [4] and self-paced learning [18, 23] advocate training models from simple to complex samples. Focal loss [26] balances positive and negative samples by modifying the classical cross-entropy loss. Inspired by the human learning process, Yu et al. [39] designed an instance difficulty model based on learning speed. Following [39], our method innovatively integrates classification margins and employs a weighted update strategy to balance historical and real-time information. This enables a more refined and fine-grained estimation of sample difficulty, ultimately contributing to a more robust margin.

2 Why Use Spherical Space instead of European Space?

Despite the remarkable results of spherical space learning [36, 5, 27, 14, 3, 40, 41], previous work has simply interpreted the benefits as a direct optimization of “geodesic” distances on the hypersphere [14, 3], without the corresponding mathematical insight, which may raise concerns or doubts about the correctness of the technique. Rigorously, we provide a mathematical explanation for this in this subsection as a theoretical guarantee.

In spherical space learning, the model normalizes the joint feature vector \mathcal{R} to the unit sphere \mathbb{S}^{d-1} . Thus, the feature space naturally becomes a unit sphere. We consider it as a $d - 1$ dimensional Riemannian manifold, where the metric tensor g comes from the induced metric of the Euclidean space. Specifically, given a point $p \in \mathbb{S}^{d-1}$ on the sphere, its tangent space $T_p\mathbb{S}^{d-1}$ consists of vectors perpendicular to p through the origin. The metric g_p is defined as an inner product $g_p(Y, Z) = \langle Y, Z \rangle_{\mathbb{R}^d}$ for any tangent vectors $Y, Z \in T_p\mathbb{S}^{d-1}$. Under this Riemannian structure, the Levi-Civita connection ∇ (Please allow a little abuse of symbols) can be given by the projection of a vector field: for any tangent vector field Y and directional tangent vector X , $\nabla_X Y$ is equal to the projection of the directional derivative $D_X Y$ in Euclidean space onto the tangent space. In general terms, for the unit sphere embedded in \mathbb{R}^d , we have:

$$\nabla_X Y = D_X Y - \langle D_X Y, p \rangle p, \quad (1)$$

where $D_X Y$ is the directional derivative along X in \mathbb{R}^d , minus its projection along the radial direction p , resulting in vectors that still belong to the tangent space. This liaison ensures that the geodesic is equivalent to a circle on the sphere passing through the center of the unit sphere: the geodesic on the manifold satisfies $\ddot{\gamma}(t) + \gamma(t) = 0$, where $\gamma(t)$ denotes the curve on the sphere, and $\ddot{\gamma}(t)$ represents the ordinary acceleration of the curve in Euclidean space.

In this geometric perspective, the classification decision boundary can be modeled as a **geodesic hypersurface**. For example, for two categories i and j , the classification boundary is the set of points that satisfy equal scores:

$$\cos \theta_i = \cos \theta_j, \quad (2)$$

where θ_i and θ_j are the angles between the features and the classification vector, respectively. Since $\cos \theta_i = \cos \theta_j$ on the unit sphere if and only if the feature vectors are distributed equiangularly about the weight vectors of the two categories, this decision boundary is exactly the set of all points that form an equirectangular angle with the weight vectors, in other words, it is a line that passes through the center of the sphere and intersects the perpendicular plane in the weight vectors, i.e., the geodesic line. In other words, the model determines the category by angle (geodesic distance) in normalized space, and the decision boundary is the geodesic isometric surface. Thus, categorizing by angle makes training more stable and geometrically meaningful, yielding explicitly metric and stable decision surfaces.

3 Proof for Classifier Weight Direction Equalization

We define the *Gram* matrix:

$$G = \mathcal{W}^\top \mathcal{W}, \quad G_{ij} = \mathcal{W}_i^\top \mathcal{W}_j = \cos \phi_{ij}, \quad (3)$$

where ϕ_{ij} is the angle between the vectors \mathcal{W}_i and \mathcal{W}_j .

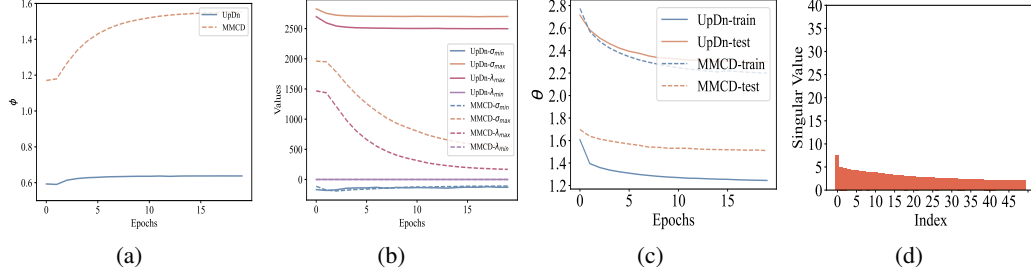


Figure 1: (a) Mean value of the angle between classifier weight vectors. (b) Upper bound λ_{\max} , lower bound λ_{\min} , maximum singular value σ_{\max} , and minimum singular value σ_{\min} of the classifier weight. (c) The mean value of the angle between the classifier weight vector and the normalized feature vector at different stages. (d) Classifier weight singular value spectra for our MMCD approach.

The Gershgorin disk theorem states that for any complex square matrix $A \in \mathbb{C}^{n \times n}$, its eigenvalues lie in the following concatenation of disks:

$$\bigcup_{i=1}^n \left\{ z \in \mathbb{C} \mid |z - A_{ii}| \leq \sum_{j \neq i} |A_{ij}| \right\}. \quad (4)$$

Each disk is centered on the diagonal element A_{ii} and has a radius that is the sum of the absolute values of the non-diagonal elements of the i -th row. Since the eigenvalues of the real symmetric matrix $G \in \mathbb{R}^{C \times C}$ are real (C is the number of candidate answers), the Gershgorin disk degenerates in the complex plane to an interval on the real axis. Then, each eigenvalue λ_i of G must be satisfied:

$$\lambda_i \in \bigcup_{k=1}^C [G_{kk} - R_k, G_{kk} + R_k], \quad (5)$$

where $R_k = \sum_{j \neq k} |G_{kj}|$. We define λ as an eigenvalue of G corresponding to an eigenvector v satisfying $Gv = \lambda v$. Expanding for the i -th row component:

$$\sum_{j=1}^C G_{ij} v_j = \lambda v_i. \quad (6)$$

Rewrite the equation as:

$$(\lambda - G_{ii})v_i = \sum_{j \neq i} G_{ij} v_j. \quad (7)$$

Take absolute values and apply trigonometric inequalities:

$$|\lambda - G_{ii}| \cdot |v_i| \leq \sum_{j \neq i} |G_{ij}| \cdot |v_j|. \quad (8)$$

We define the maximum component $|v_k| = \max_j |v_j|$, and substitute into the above equation to obtain:

$$|\lambda - G_{kk}| \cdot |v_k| \leq \sum_{j \neq k} |G_{kj}| \cdot |v_j| \leq |v_k| \sum_{j \neq k} |G_{kj}|. \quad (9)$$

Divide both sides by $|v_k|$ ($|v_k| > 0$) to get:

$$|\lambda - G_{kk}| \leq \sum_{j \neq k} |G_{kj}| = R_k. \quad (10)$$

That is, the eigenvalues λ are satisfied:

$$G_{kk} - R_k \leq \lambda \leq G_{kk} + R_k. \quad (11)$$

Considering the constructive properties of G : $G_{ii} = \|\mathcal{W}_i\|_2^2 = 1$ and $G_{ij} = \mathcal{W}_i^\top \mathcal{W}_j = \cos \phi_{ij}$, $i \neq j$, the interval for each eigenvalue λ_i is:

$$\lambda_i \in \left[1 - \sum_{j \neq i} |\cos \phi_{ij}|, 1 + \sum_{j \neq i} |\cos \phi_{ij}| \right]. \quad (12)$$

With the introduction of the margin m , the weight vectors of the different classes are forced to push away, and the pinch angle is forced to increase (see Fig. 1(a)):

$$\phi_{ij} \geq \phi'_{ij} \quad (i \neq j), \quad (13)$$

where ϕ'_{ij} is the pinch angle without margin, and ϕ_{ij} is the pinch angle with margin. From the monotonicity of the cosine function:

$$\cos \phi_{ij} \leq \cos \phi'_{ij}. \quad (14)$$

In other words, the non-diagonal elements of the Gram matrix satisfy:

$$G_{ij} = \cos \phi_{ij} \leq \cos \phi'_{ij} \quad (i \neq j). \quad (15)$$

In the case of insufficient discriminative classification, the pinch angle is less than $\pi/2$, and the optimization objective is to make the weight vectors orthogonal to each other, i.e., the pinch angle tends to $\pi/2$, so the actual range of values of the pinch angle is $[0, \pi/2]$. At this time, increasing the clamp angle ϕ_{ij} can obtain:

$$|\cos \phi_{ij}| \leq |\cos \phi'_{ij}| \quad (i \neq j). \quad (16)$$

The radius $R_i = \sum_{j \neq i} |\cos \phi_{ij}|$ of each row is then reduced, thus compressing the distribution interval of the eigenvalues:

$$\lambda_i \in [1 - R_i, 1 + R_i] \subset [1 - R'_i, 1 + R'_i], \quad (17)$$

where $1 + R'_i$ is the radius when there is no margin. With the introduction of margin, the eigenvalues λ_i are more densely distributed around 1. In the following, we explore the relationship between the eigenvalues of G and the singular values of the classifier weights \mathcal{W} . We perform a SVD of \mathcal{W} :

$$\mathcal{W} = U \Sigma V^\top, \quad (18)$$

where $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_r)$ ($r = \min(d, C)$) is the singular values of \mathcal{W} , d is the feature dimension. At this point, the eigenvalues of G are:

$$G = V \Sigma^\top U^\top U \Sigma V^\top = V \Sigma^2 V^\top. \quad (19)$$

Therefore, the eigenvalues of the Gram matrix $\lambda_i = \sigma_i^2$. The eigenvalue equalization of the Gram matrix corresponds directly to the singular value equalization of the classifier weight matrix \mathcal{W} . In this regard, we show that the margin equalizes the direction of the classifier weights.

We corroborate our claims with empirical measurements of spectral bounds. Let λ_{\max} and λ_{\min} denote the largest and smallest eigenvalues of the Gram matrix G . Let σ_{\max} and σ_{\min} denote the largest and smallest singular values of the classifier weight matrix \mathcal{W} . Likewise, G' and \mathcal{W}' have extremal values λ'_{\max} , λ'_{\min} , σ'_{\max} , and σ'_{\min} , respectively, for the UpDn baseline. Fig. 1(b) shows that the Gram spectrum of UpDn features an exceedingly large λ'_{\max} that remains essentially unchanged after a minor initial decline. σ'_{\max} parallels the behavior of λ'_{\max} . By contrast, MMCD rapidly reduces λ_{\max} to a much lower level during both warm-up and training. The foremost singular value σ_{\max} exhibits an identical trend, confirming that our method achieves pronounced equalization of the singular-value distribution.

4 Pseudo Codes of Our Model

In this section, we present the pseudo codes of our entire algorithm. In Algorithm 1, we show how the instance difficulty model is computed. For each mini-batch b at the t iteration consisting of some image-question-answer pairs (v, q, a) , we pass the image and the question to the model M , which returns the logits f and the joint representations \mathcal{R} (lines 2-5). We compute the prediction

Algorithm 1 Instance Difficulty Model

Input: Training set \mathcal{S} divided into B mini-batches, training epoch T , model M

Output: Instance difficulty model D

```
1: Initialize  $p_{i,0} \leftarrow \{\frac{1}{k}, \dots\}$ ,  $vu_{i,0} \leftarrow \{0, \dots\}$ ,  $vl_{i,0} \leftarrow \{0, \dots\}$ ,  $mu_{i,0} \leftarrow \{0, \dots\}$ ,  $ml_{i,0} \leftarrow \{0, \dots\}$  for each  $\mathcal{R}_i$  in  $\mathcal{S}$ 
2: for  $t$  in 1 to  $T$  do
3:   for  $b$  in  $B$  do
4:     for  $(v, q, a)$  in  $b$  do
5:        $f, \mathcal{R} = M(v, q)$ 
6:        $p_t \leftarrow \text{softmax}(f)$ 
7:        $D_t \leftarrow \text{Difficulty}(p_{t-1}, p_t, vu_{t-1}, vl_{t-1}, mu_{t-1}, ml_{t-1})$ 
8:     end for
9:   end for
10: end for
11: return  $D$ 
```

Algorithm 2 Train

Input: Training set \mathcal{S} divided into B mini-batches, training epoch T , model M to train, the initial frequency-aware margins \hat{m}_i^{qt} for each \mathcal{R}_i in \mathcal{S}

Output: Trained model D

```
1:  $M.\text{train}()$ .
2: for  $t$  in 1 to  $T$  do
3:   for  $b$  in  $B$  do
4:     for  $(v, q, a, \hat{m}_i^{qt})$  in  $b$  do
5:        $f, f_q, f_v, \mathcal{R}, \mathcal{R}_q, \mathcal{R}_v = M(v, q)$ 
6:        $\bar{m}^{qt} \leftarrow \mathcal{N}(\hat{m}^{qt}, \sigma)$ 
7:        $s_q \leftarrow \text{softmax}(\mathcal{W}_q \cdot e_q(q) + \frac{b}{2})[gt]$ ,  $s_v \leftarrow \text{softmax}(\mathcal{W}_v \cdot e_v(v) + \frac{b}{2})[gt]$ ,
8:        $w_q \leftarrow \frac{s_q}{s_q + s_v}$ ,  $w_v \leftarrow \frac{s_v}{s_q + s_v}$ ,
9:        $f_{uni} \leftarrow w_q \cdot f_q + w_v \cdot f_v$ 
10:       $\bar{m}_{conf} \leftarrow \text{softmax}(f/\tau) + \text{softmax}(f_{uni}/\tau)$ 
11:       $D_t \leftarrow \text{Difficulty}(\dots)$ 
12:       $m_{diff} \leftarrow 1 - \text{softmax}(D_t)$ 
13:       $\hat{\theta} \leftarrow \theta + 1 - \text{combine}(m^{qt}, m_{conf}, m_{diff})$ 
14:       $\mathcal{L}_{\text{TOTAL}} \leftarrow \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{MAM}}(\hat{\theta}) + \mathcal{L}_{\text{DCL}}(\mathcal{R}, a, D_t)$ 
15:       $\mathcal{L}_{\text{TOTAL}}.\text{backward}()$ 
16:     end for
17:   end for
18: return  $M$ 
```

Algorithm 3 Evaluate

Input: Test set divided into B mini-batches, trained model M to test

Output: Evaluation score s

```
1:  $M.\text{eval}()$ .
2: for  $(v, q, a)$  in  $b$  do
3:    $f, \mathcal{R} = M(v, q)$ 
4:    $p = \text{softmax}(f)$ 
5:    $s = \text{computeScore}(p, a)$ 
6: end for
7: return  $s$ 
```

probabilities p and package relevant variables to calculate instance difficulty D_t . Please refer to the Difficulty-aware Margins section in the main paper for the specific calculation process.

In Algorithm 2, we show how our model is trained. For each mini-batch b consisting of some images v , questions q , answers a and the frequency-aware margins \hat{m}_i^{qt} (computed for the question types of q), we pass the images and the questions to the model M , which returns the primary logits f , the question-only logits f_q , the image-only logits f_v , the multimodal feature \mathcal{R} , the question feature \mathcal{R}_q , and the image feature \mathcal{R}_v (lines 2-5). We compute the randomized frequency-aware margins in line 6, the confidence-aware margins (lines 7-9), and the difficulty-aware margins (lines 10-11). Then we integrate all the margins into angle θ in line 12. Finally, we compute the loss function using the two losses as mentioned in the main paper, and back-propagate through it to update the model (lines 13-14).

Algorithm 3 shows how the trained model is used for prediction on the test data. We simply get the prediction probabilities p and compute the accuracy score s .

5 Convergence of Our Difficulty Model

In this section, we prove the convergence of our difficulty model (i.e., Eqn. (20) in the main paper. First, we assume that the model is converged when $t \rightarrow \infty$. Therefore, for any instance representation \mathcal{R}_i , we have:

$$\therefore vu'_{i,t} \rightarrow 0, vl'_{i,t} \rightarrow 0. \quad (20)$$

$$\therefore vu_{i,t} \rightarrow 0, vl_{i,t} \rightarrow 0. \quad (21)$$

$$\therefore \frac{vu_{i,t} + c}{vl_{i,t} + c} \rightarrow 1. \quad \text{learning speed} \quad (22)$$

$$\therefore m_{i,t} = |p_{i,t}^{gt} - p_{i,t}^j| = C_1, \quad (23)$$

$$\therefore mu'_{i,t} = \log\left(\frac{1}{|\Psi|} \sum_{j \in \Psi} \exp(m_{i,t}^j)\right) = C_2, \quad (24)$$

$$ml'_{i,t} = \log\left(\frac{1}{|\Omega|} \sum_{j \in \Omega} \exp(m_{i,t}^j)\right) = C_3, \quad (25)$$

$$\therefore mu_{i,t} = \beta \cdot mu_{i,t-1} + (1 - \beta) \cdot C_2, \quad (26)$$

$$ml_{i,t} = \beta \cdot ml_{i,t-1} + (1 - \beta) \cdot C_3, \quad (27)$$

$$\therefore \beta \in (0, 1), \quad (28)$$

$$\therefore mu_{i,t} \rightarrow (1 - \beta) \cdot C_2 + \beta \cdot (1 - \beta) \cdot C_2 + \quad (29)$$

$$\beta^2 \cdot (1 - \beta) \cdot C_2 + \dots + \beta^{t-1} \cdot (1 - \beta) \cdot C_2 \quad (30)$$

$$= (1 - \beta) \cdot C_2 \cdot \frac{1 - \beta^t}{1 - \beta} = (1 - \beta^t) \cdot C_2 \rightarrow C_2, \quad (31)$$

$$mu_{i,t} \rightarrow (1 - \beta) \cdot C_3 + \beta \cdot (1 - \beta) \cdot C_3 + \quad (32)$$

$$\beta^2 \cdot (1 - \beta) \cdot C_3 + \dots + \beta^{t-1} \cdot (1 - \beta) \cdot C_3 \quad (33)$$

$$= (1 - \beta) \cdot C_3 \cdot \frac{1 - \beta^t}{1 - \beta} = (1 - \beta^t) \cdot C_3 \rightarrow C_3. \quad (34)$$

$$\therefore \frac{mu_{i,t} + c}{ml_{i,t} + c} \rightarrow \frac{C_2 + c}{C_3 + c}. \quad \text{classification margins} \quad (35)$$

$$\therefore D_{i,t} = \alpha \cdot 1 + \alpha \cdot \frac{C_2 + c}{C_3 + c}. \quad (36)$$

In conclusion, $D_{i,t}$ converges to a constant related to the predicted distribution at the time of stopping learning. C_1, C_2, C_3 are all constants.

6 Discussion on Angle between Classifier Weight Vector and Features

The margin mechanism directly influences the angle θ between the classifier weight vector \mathcal{W} and the feature representation \mathcal{R} . To assess this effect, we compute θ only on correctly classified examples. As shown in Fig. 1(c), for the UpDn baseline, training-phase angles remain narrowly concentrated while test-phase angles broaden markedly. This phenomenon indicates its overreliance on language shortcuts and poor generalization. Conversely, our MMCD approach maintains larger angles during training and tighter angles at test time. This behavior confirms that MMCD discourages spurious question–answer correlations, instead promoting decision boundaries aligned with genuine multimodal features, thereby enhancing robustness.

7 Effectiveness of Our Method

Fig.1 in the main paper demonstrates that our MMCD framework effectively mitigates all three identified deviations compared to the baseline UpDn: (1) Modality gradient optimization deviation. As shown in Fig. 6(b), the ratio of the gradient paradigms of the different modalities decreases under our method. (2) Fusion feature component deviation. As shown in Fig. 6(c), the contributions of visual and textual features to the fused feature become nearly equal, eliminating compositional dominance. (3) Classifier weight direction deviation. As shown in Fig. 1(d), the singular value spectrum of the classifier weight matrix is substantially flattened, indicating more balanced information accumulation across all directions. These results confirm that MMCD restores balanced gradient flow, feature fusion, and classifier geometry, thereby substantially improving robustness to language bias.

8 Language Bias on Different Architectures

In Section 3.1 in the main paper, we derived gradient expressions for the question and image encoders of baseline. More generally, for any multimodal fusion architecture, the encoder gradients can be written as:

$$\frac{\partial \mathcal{L}}{\partial \mathcal{W}_v} = \frac{\partial \mathcal{L}}{\partial \mathcal{R}} \frac{\partial \mathcal{R}}{\partial \mathcal{R}_v} \frac{\partial \mathcal{R}_v}{\partial \mathcal{W}_v}, \quad (37)$$

$$\frac{\partial \mathcal{L}}{\partial \mathcal{W}_q} = \frac{\partial \mathcal{L}}{\partial \mathcal{R}} \frac{\partial \mathcal{R}}{\partial \mathcal{R}_q} \frac{\partial \mathcal{R}_q}{\partial \mathcal{W}_q}. \quad (38)$$

Crucially, the Jacobians $\frac{\partial \mathcal{R}}{\partial \mathcal{R}_v}$ and $\frac{\partial \mathcal{R}}{\partial \mathcal{R}_q}$ typically involve cross-modal terms (attention weights or outer-product interactions) that amplify heterogeneity between \mathcal{R}_v and \mathcal{R}_q . As a result, even under distinct architectures, the gradient magnitudes $\|\frac{\partial \mathcal{L}}{\partial \mathcal{W}_v}\|$ and $\|\frac{\partial \mathcal{L}}{\partial \mathcal{W}_q}\|$ can diverge substantially, creating modality gradient optimization bias. Thus, across diverse fusion designs, multimodal data heterogeneity inherently risks inducing language bias, albeit to varying degrees depending on the specific cross-modal coupling mechanisms.

9 Discussion on Other Bias Types

Recent advances in addressing language bias have highlighted the need to examine additional bias modalities within VQA datasets. Dancette et al. [11] introduced VQA-CE to evaluate multimodal shortcuts, and Si et al. [33] proposed the VQA-VS benchmark, which categorizes biases into language, visual, and multimodal types. Language bias manifests as keyword reliance; visual bias arises from overemphasis on salient objects; and multimodal bias combines both shortcuts. Exhaustively analyzing each bias in isolation is both complex and resource-intensive. Instead, we extend prior language-bias studies to identify shared mechanisms and key distinctions across bias categories.

Commonalities. Despite their diversity, different bias types share two fundamental properties: (1) *Distributional Discrepancy*. Each bias is precipitated by a mismatch in joint data distributions between training and evaluation splits. (2) *Implicit Encoding*. Biases are not explicit labels but latent in the fused multimodal feature space produced by neural encoders.

Differences (1) *Modal heterogeneity*: Data from different modalities—or varying formats within a modality—exhibit unique statistical properties that influence model learning. (2) *Bias targets*: Biases can originate from question constructs, keywords, object prominence, or their combinations. We hypothesize that these distinct bias sources preferentially occupy different subspaces within the fused feature vector.

Motivated by these observations, we conjecture that mechanisms analogous to those driving language bias also underlie other bias types: (1) **Modality gradient optimization deviation**. Because heterogeneous modalities are fused, certain modalities (e.g., keywords, object features) may receive disproportionately large gradients, leading to overfitting on those modalities. (2) **Fusion feature component deviation**. Data heterogeneity and gradient imbalances induce dominance of specific feature components in the fused representation. (3) **Classifier weights direction deviation**. When dominant components correlate with high-frequency classes, the classifier’s decision boundary shifts toward these components, exacerbating existing biases.

To test these hypotheses, future work will involve:

- **Subspace Analysis**. Employing techniques such as PCA or canonical correlation analysis on fused features to detect modality-specific subspaces associated with different biases.
- **Gradient Monitoring**. Tracking per-modality gradient norms during training to quantify imbalance and correlate it with downstream bias metrics.
- **Controlled Perturbations**. Introducing synthetic perturbations in one modality (e.g., masking keywords or blurring objects) to observe shifts in the learned feature subspaces and classifier behavior.

Through this multi-pronged approach, we aim to generalize beyond language bias and develop principled debiasing strategies that address the full spectrum of multimodal shortcuts in VQA.

10 Inspiration for the Community

Our analysis in Section 3 in the main paper decomposes language bias into three interrelated phenomena: (1) **modality gradient optimization deviation**, whereby differential feature scale leads to uneven gradient contributions; (2) **fusion feature component deviation**, in which a subset of fused features (e.g., high-frequency word embeddings or salient object vectors) overwhelms other modalities; and (3) **classifier weight direction deviation**, where the decision boundary aligns preferentially with dominant feature directions, exacerbating shortcut usage. We posit that explicitly mitigating these mechanisms can yield more robust VQA models.

First, to address modality gradient optimization deviation, one can employ modality-aware preprocessing, such as per-modality batch-norm layers, gradient clipping tailored to each feature stream, or learnable scaling factors that equalize gradient norms without distorting semantic representations. Second, to counter fusion feature component deviation, subspace regularization or variance-penalizing constraints may be imposed on the fused feature vector, ensuring that no single component monopolizes the representation. Third, to correct classifier weight direction deviation, one could integrate orthogonality constraints or adversarial re-weighting on the final linear layer, discouraging alignment with bias-prone feature axes. Moreover, our findings in Section 4 in the main paper demonstrate the power of margin-based objectives in amplifying inter-class separation for bias-susceptible examples. Based on these findings, we invite the community to develop principled debiasing strategies that holistically address the full spectrum of shortcut behaviors in VQA.

11 Limitations and Future Works

Despite its advantages, our difficulty-aware margin framework exhibits three key limitations, which we discuss below alongside prospective research avenues.

Training speed. Our method computes per-sample difficulty across the entire training corpus, incurring substantial time and memory costs. This limits applicability in resource-constrained or real-time settings. To alleviate this, future work could group examples by frequency, gradient norm, or loss value and estimate difficulty at the cohort level, thereby reducing computational complexity

Table 1: Details of various VQA datasets.

Dataset	Data category	Training set	Validation set	Test set
VQA v2	Images	82783	40504	-
	QA pairs	443757	214354	-
VQA-CP v1	Images	118442	-	87400
	QA pairs	244547	-	125314
VQA-CP v2	Images	120932	-	98226
	QA pairs	438183	-	219928
VQA-CE	Images	82783	-	30424
	QA pairs	443757	-	63,298
GQA-OOD	Images	74256	9406	388
	QA pairs	14305356	51045	2796

without compromising estimation fidelity. How to estimate difficulty faster and more accurately is an interesting and challenging problem for future work.

Bias types. This study concentrates on language bias induced by question–answer frequency priors. Consequently, the proposed frequency-based margin may underperform in scenarios dominated by other bias modalities. However, incorporating confidence- and difficulty-based margins provides complementary robustness. We also envision extending frequency statistics beyond question–answer pairs to capture broader co-occurrence patterns, enabling adaptation to diverse bias conditions.

Dependence on the statistics of the dataset. Our multi-margin scheme leverages a priori knowledge of question types and answer distributions. In deployment, misestimation of these statistics under a distribution shift may degrade performance. Although confidence- and difficulty-based margins help mitigate this risk, designing margin strategies that adapt online, without explicit statistical priors, remains an open challenge. Future research should explore self-calibrating or dataset-agnostic mechanisms to ensure reliable robustness in dynamic environments.

12 More Experiments Results

12.1 Implementations Details

General implementation details We employ a popular VQA architecture UpDn [2] as the base network for all our experiments. Following previous works [6, 24, 31, 38, 42, 3], we use a Faster RCNN model [32] pretrained by [2] to extract the top 36 visual object feature vectors, each of dimension 2048. All the questions are tokenized into tokens of maximum question length. Each question word is encoded by Glove vectors [30] of dimension 300. These embeddings are passed on to a single layer GRU [10] to obtain the final question feature vector, which is of dimension 1024. All our implementations are in PyTorch.

Implementation details of MMCD We implemented our MMCD model in PyTorch with a single RTX 3090 GPU and used the AdamW optimizer with a weight decay of 0.001. The batch size B is set to 512. The learning rate is set to 0.002. We train MMCD for 20 epochs. We list the hyperparameters and their values used by our models below:

- Scaling parameter s : 16.
- Standard deviation σ for the randomization of frequency-aware margins: 0.1.
- Temperature τ_1 used for generating the confidence-aware margins: 0.2.
- Temperature τ_2 used for difficulty-aware contrastive learning mechanism: 1.0.
- Difficulty contribution balance factor α : 0.6.
- Difficulty update weights β : 0.5.
- Weight λ of difficulty-aware margins: 0.05.

Table 2: **Performance comparisons with the state-of-the-art methods on the VQA v2 and VQA-CP v2 datasets** with respect to different answer categories. The best performance in each category is highlighted in bold.

Methods	VQA v2			VQA-CP v2		Diff
	Y/N-CP	Num-CP	Others-CP	All	All	
UpDn [2]	81.18	42.14	55.66	63.48	39.74	23.74
CF-VQA [28]	81.13	43.86	50.11	60.94	55.05	5.89
AdaVQA [14]	47.78	34.13	51.14	46.98	54.02	7.04
COB [17]	81.36	43.30	55.86	63.80	57.53	6.27
MMCD	76.16	37.00	52.43	59.32	61.43	2.11

12.2 Datasets

We validate the robustness of the proposed MMCD method on various out-of-distribution (OOD) benchmarks, such as VQA-CP v1 [1], VQA-CP v2 [1], VQA-CE [11], and GQA-OOD [19]. Additionally, we also evaluate the in-distribution (ID) performance on the VQA v2 dataset. In the manuscript and appendix, we provide experimental results on VQA v2, VQA-CP v1, and VQA-CP v2 datasets. We also provide additional experimental results on VQA-CE and GQA-OOD datasets below. The details of the datasets used in this work are shown in Table 1.

VQA v2 improves upon the original VQA v1 dataset by balancing answer distributions through collecting visually similar image pairs that yield differing answers for identical questions. The dataset comprises over 204 K images from MS COCO, each annotated with three human-generated questions and ten answers per question, resulting in approximately 1.1 M question-answer pairs. Two evaluation tracks (open-ended and multiple-choice) are provided. This balanced design mitigates language bias and yields heavier-tailed answer distributions, challenging models to ground answers in visual content rather than exploit priors.

VQA-CP v1 reorganizes the VQA v1 dataset to create a changing-priors benchmark, deliberately altering the answer distribution across training (118 K images, 245 K questions, 2.5 M answers) and test (87 K images, 125 K questions, 1.3 M answers) splits. Each question type’s most frequent answers in training are inverted in testing, forcing models to rely on visual grounding rather than statistical biases. Questions and annotations follow the original VQA format, with splits derived via greedy grouping and redistribution to maximize concept coverage without overlap.

VQA-CP v2 extends the changing-priors paradigm to VQA v2 by reclustering 121 K training images (438 K questions, 4.4 M answers) and 98 K test images (220 K questions, 2.2 M answers) to invert answer distributions per question type. By leveraging the larger, more balanced VQA v2 base, VQA-CP v2 accentuates the challenge of bias mitigation, evaluating robustness to distributional shifts. Models are evaluated using standard VQA metrics, highlighting the degradation in performance when visual cues cannot be supplanted by language priors.

VQA-CE (VQA Counterexamples) introduces a counterfactual evaluation protocol atop VQA v2 by mining multimodal shortcuts and selecting image-question-answer triplets where trivial predictive rules fail. The Counterexamples subset consists of about ~ 63 K examples where all shortcuts lead to incorrect answers. The non-overlapping Easy subset has ~ 147 K examples where at least one shortcut points to the correct answer. The Hard subset includes ~ 3 K examples where no shortcuts exist or provide any clues for question answering. Simply, we consider the Counterexamples subset as the test set.

GQA-OOD is an out-of-distribution benchmark derived from the GQA dataset to evaluate reasoning under varying concept frequencies. The benchmark partitions GQA’s 113 K real-world COCO-based images and 22 M programmatically generated questions into head and tail splits according to question-answer pair rareness, providing separate accuracy measures for frequent and infrequent pairs. By contrasting performance on rare concepts against overall accuracy, GQA-OOD exposes biases toward common training patterns and better assesses compositional reasoning.

Table 3: Accuracy comparison on VQA-CE and GQA-OOD datasets.

Datasets	VQA-CE			GQA-OOD		
Methods	Overall	Counter	Easy	All	Tail	Head
CSS [8]	53.55	34.36	62.08	44.24	41.20	46.11
GENB [9]	57.87	34.80	68.15	49.43	45.63	51.76
RMLVQA [3]	58.05	35.01	68.21	49.07	44.50	51.88
CVIV [29]	-	36.12	-	49.36	-	-
Ours	67.86	38.69	80.83	52.25	47.41	55.22

Table 4: **Comparison between LXMERT and MMCD on four metrics on multiple datasets.** \uparrow indicates larger is better, whereas \downarrow indicates smaller is better.

Metric	Methods	VQA-CP v1	VQA-CP v2
C \downarrow	LXMERT	0.49	0.52
	MMCD	0.56	0.58
S \uparrow	LXMERT	1.25	1.24
	MMCD	1.33	1.31
D \uparrow	LXMERT	0.78	0.76
	MMCD	0.82	0.79
Acc. \uparrow	LXMERT	52.82	51.75
	MMCD	72.07	66.95

12.3 Experiment Results on VQA v2

In addition to reporting the performance of the various models in the literature on the two datasets, we also validate the robustness of the proposed MMCD on the in-domain dataset VQA v2 [13]. As shown in Table 2, MMCD outperforms all competitors in terms of robustness, with the lowest difference between the accuracy on VQA-CP v2 and VQA v2, demonstrating that our method is the most robust approach.

12.4 Evaluation on VQA-CE and GQA-OOD

To assess general applicability in diverse real-world scenarios, we evaluate the debiasing performance of our method on the GQA-OOD and VQA-CE datasets, as shown in Table 3. We adopt LXMERT as our baseline. Notably, the VQA-CE test set emphasizes false negative examples involving questions, visual information, and answers, requiring special attention to performance on counterfactual datasets. The experimental results reveal that our proposed method effectively mitigates the impact of spurious correlations and enhances generalization, with overall accuracy improvements of 2.57% on the VQA-CE Counterexamples subset and 2.82% on the GQA-OOD compared to the second-best models.

12.5 Quantitative Analysis

In margin learning and contrastive learning, the features are regularized to fall on a hypersphere [14, 21], and the loss directly optimizes the distance between instances in the feature space. Thus, we can leverage distance to evaluate the quality of learned representations. We propose several metrics to evaluate representations learned from multiple benchmark datasets and study why MMCD achieves better performance than the baseline LXMERT [34].

Intra-class Compactness. One optimization goal is to minimize the distance between positive samples. We define intra-class compactness **C** under the hyperspheres setting as the average distance from samples of the same class to the corresponding class center. Specifically, unlike previous works [37, 25] that used Euclidean distance measurement, we use geodesic distance [12, 22] $d(x_1, x_2) =$

Table 5: Ablation experiments for different modules of the MMCD model on the VQA-CP v2 dataset.

Methods	All	Y/N	Num	Others
Baseline	41.42	46.83	12.96	46.40
+ Frequency-aware Margins	59.44	86.43	50.41	47.78
+ Confidence-aware Margins	59.74	85.35	52.17	48.40
+ Difficulty-aware Margins	61.09	89.06	53.40	48.54
+ DCL	61.34	88.93	55.68	48.44

Table 6: Ablation experiments for different modules of the MMCD model on the VQA-CP v2 dataset. We remove each component one by one from the complete method to evaluate the corresponding marginal contribution.

Methods	All	Y/N	Num	Others
MMCD	61.34	88.93	55.68	48.44
– Frequency-aware Margins	54.08	86.43	18.88	46.79
– Confidence-aware Margins	60.44	89.51	51.07	47.77
– Difficulty-aware Margins	61.22	89.12	54.16	48.54
Baseline	41.42	46.83	12.96	46.40

$\arccos(\langle x_1, x_2 \rangle)$ where $\langle \cdot, \cdot \rangle$ is the standard inner product, which is more intuitive and reasonable on hyperspheres.

$$\mathbf{C} = \frac{1}{C} \sum_{i=1}^C \frac{1}{|F_i|} \sum_{x_{i,j} \in F_i} d(x_{i,j}, c_i), \quad (39)$$

where F_i is the set of features from class i , c_i is the center of samples from class i on the hypersphere:

$$c_i = \frac{\sum_{x_{i,j} \in F_i} x_{i,j}}{\|\sum_{x_{i,j} \in F_i} x_{i,j}\|_2}, \quad x_{i,j} \text{ is the } j\text{-th sample of class } i.$$

Intra-class Separation. Another optimization goal is to maximize the distance between negative samples. Similar to [25], we define inter-class separation \mathbf{S} under the hyperspheres setting as the average distances between different class centers:

$$\mathbf{S} = \frac{1}{C(C-1)} \sum_{i=1}^C \sum_{j=1, j \neq i}^C d(c_i, c_j). \quad (40)$$

Discriminative Decision Boundaries. Inspired by [25], we recognize that intra-class compactness and inter-class separation fail to measure the discriminative power of feature space accurately. Furthermore, we define the discriminative degree of the feature space as the difference between the distance from the sample to other class centers and the corresponding class center distance:

$$\mathbf{D} = \frac{1}{C(C-1)} \sum_{i=1}^C \frac{1}{|F_i|} \sum_{x_{i,j} \in F_i} \sum_{k=1, k \neq i}^C d(x_{i,j}, c_k) - d(x_{i,j}, c_i). \quad (41)$$

LXMERT vs. MMCD. In Table 4, we compare the compactness, separation, and discriminative power of feature space between LXMERT [34] and MMCD on two benchmark datasets. The results highlight that MMCD efficiently aggregates common class samples and separates heterogeneous samples, reshapes reasonable class boundaries, and achieves clearer and discriminative feature space structures compared to LXMERT, effectively alleviating the language biases challenge.

12.6 Further Studies on Ablations

Each component of MAM on VQA-CP v2. We conducted a systematic evaluation to isolate the contribution of each component to the overall performance of our model. The detailed comparative results are provided in Table 5. Our findings are as follows: **Effect on frequency-aware margins.** The substantial performance enhancement observed when incorporating frequency-aware margins

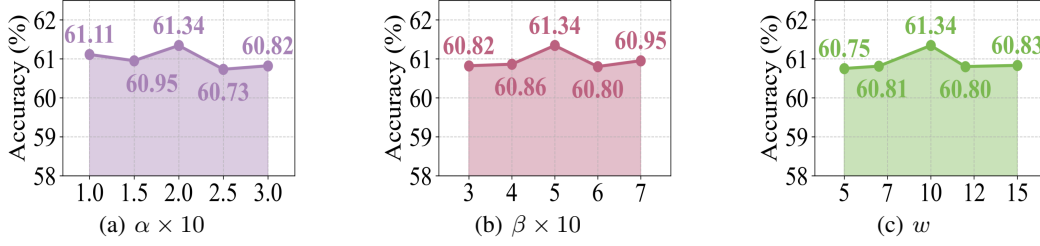


Figure 2: Comparison of Accuracy on the VQA-CP v2 dataset with different parameter configurations.

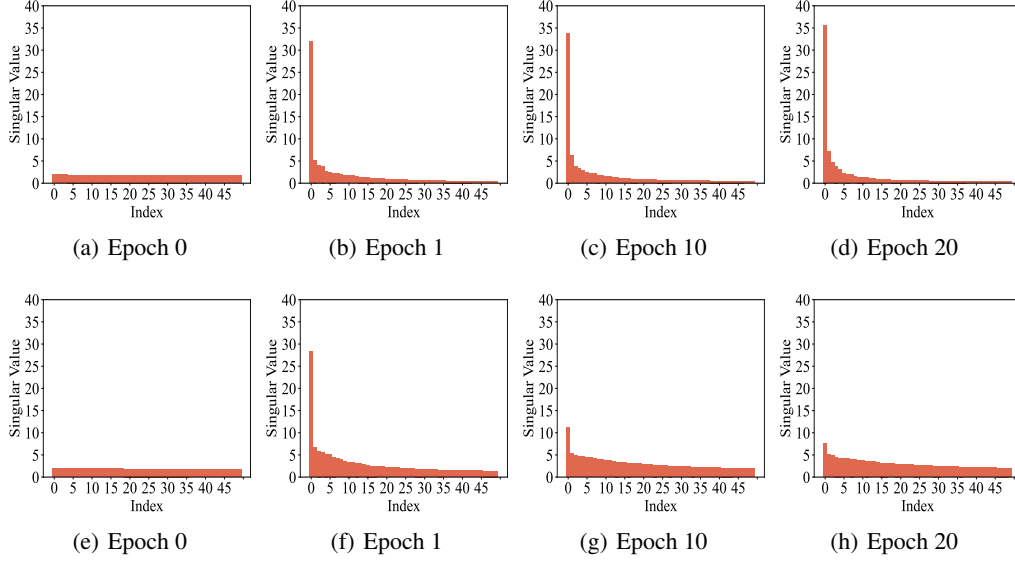


Figure 3: Trend of the top 50 singular values of the classifier weight matrices of UpDn and MMCD during training. (a), (b), (c), and (d) for UpDn and (e), (f), (g), and (h) for MMCD.

demonstrates their critical function in mitigating language biases resulting from class imbalance. **Effect on confidence-aware margins.** The simple multimodal logits strategy effectively utilizes the inherent sample complexity, resulting in the development of robust and discriminative feature spaces. **Effect on difficulty-aware margins.** The integration of a fine-grained instance difficulty model contributes significantly to forming more discriminative decision boundaries. **Effect on DCL.** Our experiments emphasize the pivotal role of the DCL mechanism in enhancing intra-class compactness and inter-class separation.

Another view of the ablations of MMCD. We conducted an ablation study by sequentially removing each component of the multi-grained adaptive margin. As summarized in Table 6, omitting any single margin degrades overall accuracy, confirming their complementary roles. In particular, the frequency-aware margins have the greatest impact: its removal incurs a 7.26% drop in accuracy, underscoring the importance of leveraging class-frequency priors to mitigate language bias. Furthermore, performance on “yes/no” and “other” questions remains largely unaffected, demonstrating MMCD’s consistent reasoning ability, whereas accuracy for “number” questions declines markedly, indicating heightened sensitivity to numerical responses.

12.7 More Parameter Analysis

As shown in Fig. 2, we conduct more parameter sensitivity analysis of MMCD with respect to three critical hyperparameters: α in Eqn. (20), β in Eqn. (22), and w in Eqn. (26) in the main paper. Across all combinations, accuracy fluctuates by no more than 0.61%. The results reveal that, regardless of the parameter combinations, the overall accuracy remains highly stable, with fluctuations

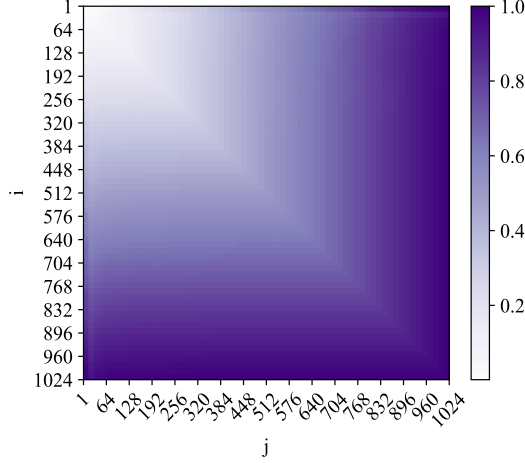


Figure 4: Subspace similarity between top- i column vectors of \mathcal{W} and top- j of \mathcal{W}' . For clearer illustrations, we present the analysis for $i, j \in [1, 1024]$.

limited to within 0.61%. This negligible variance highlights MMCD’s robustness, indicating that its effectiveness does not heavily depend on precise hyperparameter tuning. Such stability is particularly advantageous for real-world applications, where extensive grid search or parameter calibration is often impractical. Furthermore, the observed insensitivity underscores MMCD’s strong potential for transferability, enabling seamless adaptation to diverse VQA benchmarks and domain-specific tasks. In summary, the resilience to hyperparameter variations not only simplifies deployment and reduces computational cost but also affirms MMCD’s generalization strength in addressing language bias.

13 More Analysis on Classifier Weights

Let \mathcal{W} denote the classifier weight matrix of the baseline UpDn model, and \mathcal{W}' the analogous weights in our MMCD variant. Fig. 3 plots the evolution of their singular values during training. For UpDn, the leading singular value grows markedly larger than the rest, indicating a concentration of representational capacity along a single dominant direction. Conversely, MMCD yields a more uniform spectrum: the top singular value contracts while the second singular value expands. This spectrum equalization corroborates that MMCD attenuates classifier weight direction deviation, thereby mitigating language shortcuts.

To quantify subspace alignment, we perform singular value decomposition on \mathcal{W} and \mathcal{W}' , yielding right singular unitary matrices \mathbf{V} and \mathbf{V}' . We measure the normalized Grassmann distance-based subspace similarity

$$\psi(\mathbf{V}, \mathbf{V}', i, j) = \frac{\|\mathbf{V}_{:,i}^\top \mathbf{V}'_{:,j}^\top\|_F^2}{\min(i, j)} \in [0, 1]. \quad (42)$$

Here, $\psi(\cdot)$ ranges from 0 to 1, where 1 indicates a complete overlap of subspaces and 0 signifies total separation. \mathbf{V} and \mathbf{V}' represents the top- i and top- j column vectors of \mathbf{V} and \mathbf{V}' , respectively. As shown in Fig. 4, we make an *important observation*.

The directions corresponding to the top singular vectors exhibit minimal overlap between \mathbf{V} and \mathbf{V}' , while the others do not. This confirms that MMCD specifically disrupts the dominant weight direction, which is one of the key contributors to language bias, while preserving the remaining decision axes.

13.1 More Qualitative analysis.

To further evaluate MMCD’s qualitative performance, we examine attention maps across answer categories that are typically prone to language biases—such as “yes/no” queries, color recognition, and common-object identification. Fig. 5(c) reveals that, despite focusing on a minimal attention

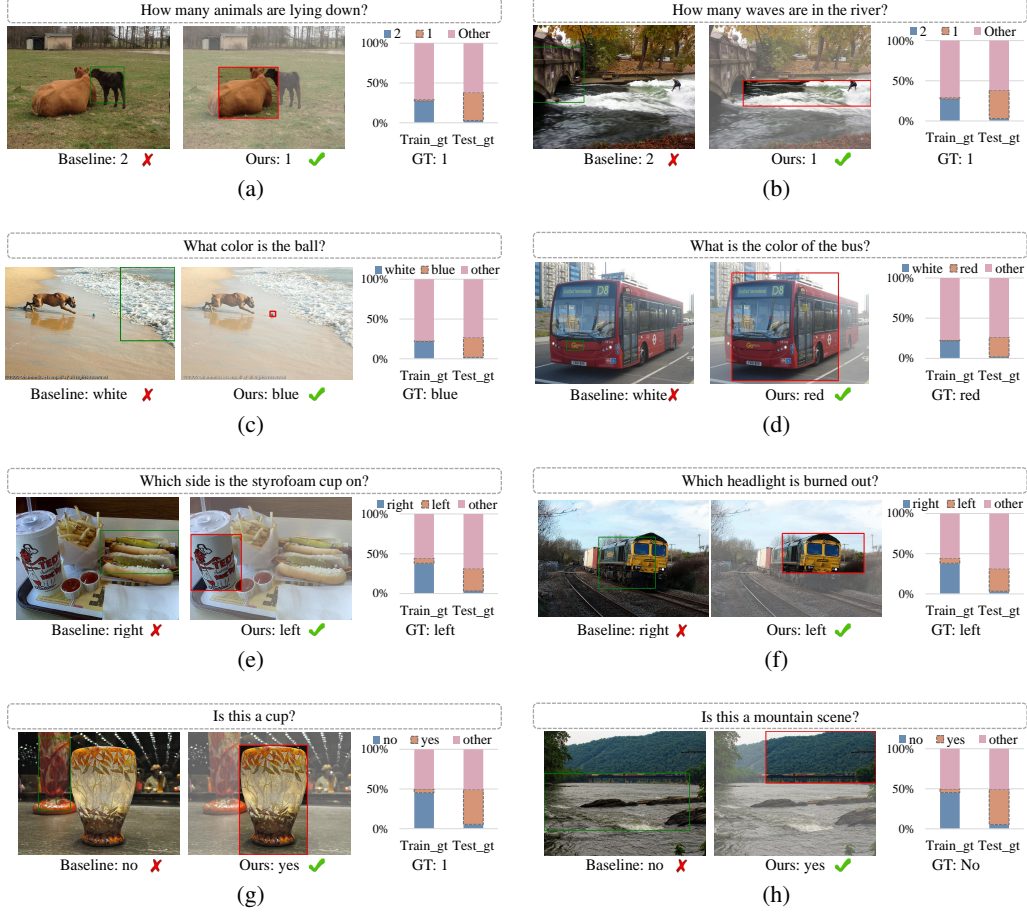


Figure 5: Visualization results of MMCD in robust reasoning and bias mitigation.

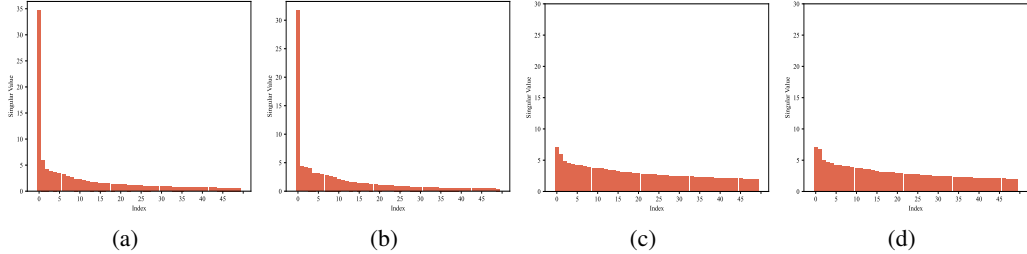


Figure 6: Analysis of classifier weight direction deviation on VQA-CE and GQA-OOD benchmarks. (a) and (b) for baseline UpDn on VQA-CE and GQA-OOD datasets. (c) and (d) for our MMCD on VQA-CE and GQA-OOD datasets.

footprint, MMCD consistently isolates the critical visual cues needed for accurate reasoning. This finding confirms that our method enhances multimodal inference by grounding predictions in both visual semantics and linguistic logic.

14 Classifier Weight Direction Deviation on More datasets

To investigate whether classifier-weight direction bias is a universal bias driver, we evaluate the singular-value spectra of the classifier on two additional benchmarks. As depicted in Fig. 6, both

Table 7: Comparison of FLOPs and Parameters on the VQA-CP v2 dataset.

Methods	Ref.	FLOPs (G)	Parameters (M)
UpDn	CVPR’18	0.18	12.94
RMLVQA	CVPR’23	0.17	12.94
GGD	TPAMI’23 0.19	33.98	
COB	WACV’23 0.16	35.18	
PHOH	AAAI’25	0.31	18.18
MMCD	Ours	0.19	35.9

datasets exhibit pronounced spectral skew: the leading singular value of the weight matrix is anomalously large compared to the second singular value, confirming a dominant decision axis. Notably, VQA-CE was explicitly designed to probe multimodal shortcuts, yet the persistence of an inflated top singular value indicates that classifier-weight direction bias underlies diverse bias phenomena across tasks.

15 Computational Efficiency

In our work, all experimental configurations conform to standard VQA practice and entail a modest, operationally acceptable computational footprint. Specifically, the proposed MAM and DCL operate exclusively during training as loss-level regularizers that shape the representation space, without invoking any auxiliary branches or contrastive operations. In the test phase, the model follows the standard VQA forward path without invoking any auxiliary branches or contrastive operations. Consequently, inference-time latency, FLOPs, and memory usage remain effectively unchanged relative to the baseline, ensuring predictable and efficient deployment. As shown in Table 7, the training FLOPs increase from only marginally 0.18G (COB) to 0.19G (MMCD) while the parameter count is listed as 35.18M (COB) vs. 35.9M (MMCD). Compared with the existing baselines, MMCD delivers robustness gains with negligible inference-time overhead and an acceptable training-time cost.

16 Broader Impact

By revealing the formation process of language bias challenges in visual question answering and the efficacy of margin mechanisms, we believe that this work will have a positive impact.

References

- [1] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi. Don’t just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4971–4980, 2018.
- [2] P. Anderson, X. He, C. Buehler, D. Teney, and M. Johnson. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, 2018.
- [3] A. Basu, S. Addepalli, and R. V. Babu. Rmlvqa: A margin loss approach for visual question answering with language biases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11671–11680, 2023.
- [4] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 41–48, 2009.
- [5] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1578–1587, 2022.

- [6] R. Cadene, C. Dancette, H. Ben-younes, M. Cord, and D. Parikh. Rubi: Reducing unimodal biases in visual question answering. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.
- [7] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.
- [8] L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, and Y. Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10800–10809, 2020.
- [9] J. W. Cho, D.-J. Kim, H. Ryu, and I. S. Kweon. Generative bias for robust visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11681–11690, 2023.
- [10] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [11] C. Dancette, R. Cadene, D. Teney, and M. Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1574–1583, 2021.
- [12] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019.
- [13] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913, 2017.
- [14] Y. Guo, L. Nie, Z. Cheng, F. Ji, J. Zhang, and A. Del Bimbo. Adavqa: Overcoming language priors with adapted margin cosine loss. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 708–714, 2021.
- [15] X. Han, S. Wang, and C. Su. Greedy gradient ensemble for robust visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1584–1593, 2021.
- [16] X. Han, S. Wang, C. Su, Q. Huang, and Q. Tian. General greedy de-bias learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45:1–17, 2023.
- [17] A. Jha, B. Patro, L. Van Gool, and T. Tuytelaars. Barlow constrained optimization for visual question answering. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1084–1093, 2023.
- [18] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. Hauptmann. Self-paced curriculum learning. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*, volume 29, 2015.
- [19] C. Kervadec, G. Antipov, M. Baccouche, and C. Wolf. Roses are red, violets are blue... but should vqa expect them to? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2776–2785, 2021.
- [20] S. Khan, M. Hayat, S. W. Zamir, J. Shen, and L. Shao. Striking the right balance with uncertainty. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 103–112, 2019.
- [21] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 18661–18673, 2020.

- [22] Y. Koishekenov, S. Vadgama, R. Valperga, and E. J. Bekkers. Geometric contrastive learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 206–215, 2023.
- [23] M. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 23, 2010.
- [24] G. Kv and A. Mittal. Reducing language biases in visual question answering with visually-grounded question encoder. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 18–34. Springer, 2020.
- [25] T. Li, P. Cao, Y. Yuan, L. Fan, Y. Yang, R. S. Feris, P. Indyk, and D. Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6918–6928, 2022.
- [26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.
- [27] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 212–220, 2017.
- [28] Y. Niu, K. Tang, H. Zhang, Z. Lu, X.-S. Hua, and J.-R. Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12700–12710, 2021.
- [29] Y. Pan, J. Liu, L. Jin, and Z. Li. Unbiased visual question answering by leveraging instrumental variable. *IEEE Transactions on Multimedia (TMM)*, 26:6648–6662, 2024.
- [30] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.
- [31] S. Ramakrishnan, A. Agrawal, and S. Lee. Overcoming language priors in visual question answering with adversarial regularization. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.
- [32] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, 2015.
- [33] Q. Si, F. Meng, M. Zheng, Z. Lin, Y. Liu, P. Fu, Y. Cao, W. Wang, and J. Zhou. Language prior is not the only shortcut: A benchmark for shortcut learning in vqa. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3698–3712, 2022.
- [34] H. Tan and M. Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.
- [35] A. Vosoughi, S. Deng, S. Zhang, Y. Tian, C. Xu, and J. Luo. Cross modality bias in visual question answering: A causal view with possible worlds vqa. *IEEE Transactions on Multimedia (TMM)*, 2024.
- [36] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5265–5274, 2018.
- [37] T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 9929–9939. PMLR, 2020.

- [38] Z. Wen, G. Xu, M. Tan, Q. Wu, and Q. Wu. Debiased visual question answering from feature and sample perspectives. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 3784–3796, 2021.
- [39] S. Yu, J. Guo, R. Zhang, Y. Fan, Z. Wang, and X. Cheng. A re-balancing strategy for class-imbalanced classification based on instance difficulty. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 70–79, 2022.
- [40] Z. Zhou, J. Yao, F. Hong, Y. Zhang, B. Han, and Y. Wang. Combating representation learning disparity with geometric harmonization. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 20394–20408, 2023.
- [41] J. Zhu, Y. Liu, H. Zhu, H. Lin, Y. Jiang, Z. Zhang, and B. Chen. Combating visual question answering hallucinations via robust multi-space co-debias learning. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 955–964, 2024.
- [42] X. Zhu, Z. Mao, C. Liu, P. Zhang, B. Wang, and Y. Zhang. Overcoming language priors with self-supervised learning for visual question answering. *arXiv preprint arXiv:2012.11528*, 2020.