Information Sciences 739 (2026) 123163

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Focal equilibrium: Bias reshaping for generalizable and robust visual understanding

Chao Wang[a,b,*], Weiwei Fu[a,b], Haoyang Li[c,e], Linqi Ye[a,b], Yang Zhou[c,d,**]

[a] School of Future Technology, Shanghai University, Shanghai, 200444, China
[b] Institute of Artificial Intelligence, Shanghai University, Shanghai, 200444, China
[c] School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, 200444, China
[d] Shanghai Artificial Intelligence Laboratory, Shanghai, 201114, China
[e] Australian Artificial Intelligence Institute, University of Technology Sydney, Sydney, NSW 2007, Australia

HIGHLIGHTS

- Unified debiasing framework balancing ID and OOD performance.
- Logits-based bias estimation without additional annotations.
- Consistent OOD gains with competitive ID results across backbones.

ARTICLE INFO

Keywords:
Bias mitigation
Language prior
Multimodal
Out-of-distribution
Visual question answering (VQA)

ABSTRACT

Visual Question Answering (VQA) models frequently rely on language priors while overlooking visual content. Current mainstream debiasing methods face limitations: data augmentation techniques demand high manual annotation costs and struggle to achieve balanced mitigation of biases, while ensemble-based approaches only capture language priors through a QA branch without fully identifying comprehensive bias. We propose `FAIR`, a bias reshaping method that utilizes pseudo-label functions to balance distribution bias and emphasizes learning weights for challenging samples. Moreover, we find that using model logit distributions as a substitute can achieve comparable effects to traditional data distribution annotations required by previous ensemble methods. Experimental results demonstrate that `FAIR` achieves the best balance among comparable methods, reaching 64.03% accuracy on VQA v2 and 60.96% on VQA-CP v2.

1. Introduction

Visual Question Answering (VQA) is a complex multimodal task that integrates natural language processing with computer vision [1]. It requires intelligent systems to generate accurate natural language responses to questions posed about given images. The field has attracted significant attention due to its applications in various domains, such as automated customer service, accessibility technologies and medical research [2]. Numerous studies have shown that VQA models are strongly influenced by inherent biases within datasets [3]. These biases often cause models to rely heavily on high-frequency answers within the dataset, thereby neglecting the actual content of the images and the contextual nuances of the questions. This phenomenon is illustrated in Fig. 1(a): for the

[*] Corresponding author at: School of Future Technology, Shanghai University, Shanghai, 200444, China.
[**] Corresponding author at: School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, 200444, China.
Email addresses: cwang@shu.edu.cn (C. Wang), fuweiwei2001@shu.edu.cn (W. Fu), haoyang-li@shu.edu.cn (H. Li), yelinqi@shu.edu.cn (L. Ye), saber_mio@shu.edu.cn (Y. Zhou).

https://doi.org/10.1016/j.ins.2026.123163
Received 20 September 2025; Received in revised form 26 January 2026; Accepted 26 January 2026
Available online 28 January 2026

question “How many umbrellas are seen?”, the model erroneously answers “2” instead of the correct answer “1”. The model’s prediction probability distribution closely mirrors the answer distribution in the dataset, indicating that the model has learned spurious correlations between questions and answers rather than understanding the content of the images.

Bias in VQA models is particularly problematic because it undermines the generalization ability of these models when applied to different datasets. When tested on datasets with varying answer distributions, the reliance on learned spurious correlations becomes a liability, resulting in poor performance and unreliable predictions. This challenge is exacerbated by the fact that VQA tasks inherently involve complex interactions between visual and textual information, making the models more susceptible to biases ingrained in either modality.

To address these biases, researchers have proposed various debiasing methods, which can broadly be categorized into two main approaches: *data augmentation methods* and *ensemble-based methods*. Data augmentation methods aim to mitigate data imbalance by creating new samples from the training set. For instance, annotation-based methods [4,5] leverage additional human annotations to ensure that model predictions are based on the correct regions of the images, thereby reducing bias. These methods help in aligning the model’s focus with the intended visual cues.

On the other hand, ensemble-based methods integrate an additional question-answering (QA) branch specifically designed to learn and mitigate bias. This approach enables the VQA model to discern and ignore bias during the prediction process, as illustrated in Fig. 1(b). By segregating the dataset into *In-Distribution (ID) datasets* and *Out-of-Distribution (OOD) datasets*, researchers can better diagnose the effectiveness of debiasing strategies. ID datasets have consistent distributions between the training and test sets, ensuring that models are tested on familiar patterns. Conversely, OOD datasets’ feature distributions are intentionally inconsistent or even opposite to those in the training sets, thereby challenging the models to generalize beyond learned biases.

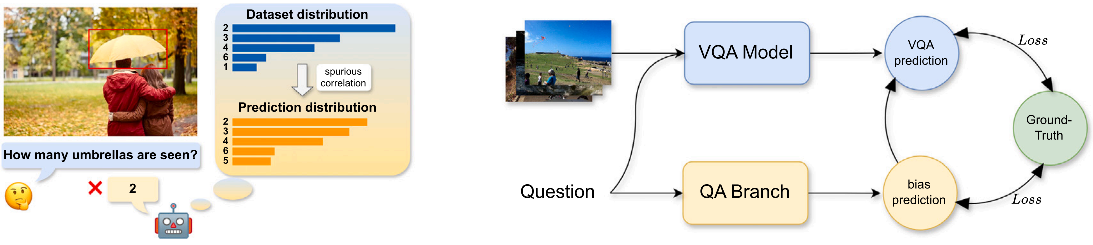
Despite substantial progress, existing VQA debiasing methods still face several fundamental contradictions. First, many approaches improve OOD generalization at the cost of degraded ID performance, revealing a persistent ID/OOD trade-off. Second, most approaches primarily focus on mitigating language priors, while their ability to enhance visual grounding remains limited, leaving visual shortcut biases largely underexplored. Third, data augmentation-based and some ensemble-based strategies often rely on counterfactual generation, region masking or question type, which not only introduce additional computational and annotation costs, but also conflict with the original philosophy of OOD datasets—namely, diagnosing whether models can learn unbiased knowledge directly from biased data.

Motivated by these limitations, we revisit ensemble-based and pseudo-label-driven debiasing frameworks from a unified perspective. Rather than treating pseudo-labels as heuristic re-weighting signals, we reinterpret them as a mechanism for reshaping distribution bias. Specifically, we propose to approximate distribution bias directly from model output logits, eliminating the need for annotated question-type statistics. This logits-based bias estimation is further integrated with Equalized Focal Loss (EFL) [6] to emphasize hard and under-represented samples, adversarial perturbations to suppress spurious textual patterns, and an explicit visual-answer (VA) branch to model visual shortcut biases.

We term the resulting framework **Focal Equilibrium** (`FAIR`), highlighting its objective of achieving a balanced equilibrium across multiple dimensions: robustness between ID and OOD settings, mitigation of language and visual biases, and stability between bias suppression and model performance. Through this design, `FAIR` provides a principled refinement of existing ensemble-based debiasing methods, enabling more robust and grounded visual question answering.

`FAIR` adopts a two-stage training process inspired by GGE. In the first stage, the model treats the training data distribution as a bias and generates pseudo-labels accordingly. Adversarial training is employed to inject textual perturbations into input questions, improving robustness against spurious linguistic patterns. The second stage extends this process by incorporating not only distribution bias and question shortcut bias but also visual shortcut bias. Specifically, visual shortcut bias is extracted through a VA branch to produce updated pseudo-labels, ensuring that the model explicitly identifies and mitigates visual biases during training.

Several prior VQA debiasing methods share a similar high-level goal but differ substantially in how bias is modeled and exploited during training. Re-Scaling [7] and GGE [8] primarily rely on dataset-level statistics conditioned on question types to estimate answer bias, which implicitly assumes access to question-type annotations or predefined linguistic groupings. GenB [9] further introduces



(a) Example of biased prediction

(b) Ensemble-based method

Fig. 1. **(a)** In the training set, the answer “2” has the highest frequency among question type “How many ...”, the predictions of VQA models tend to answer “2”, while the ground truth is “1”. **(b)** ensemble-based methods involve the learning of dataset biases with the QA branch, followed by training the VQA model using the bias prediction.

Table 1
Comparison with closely related VQA debiasing methods.

Method	Bias modeling	Pseudo-label	Visual bias	Adversarial
Re-Scaling [7]	statistics	×	×	×
GGE [8]	statistics + Q-only logits	✓	×	×
GenB [9]	statistics	✓	×	GAN-style
CSS [10]	No explicit bias model	×	✓	×
D-VQA [11]	statistics + Q-only logits + V-only logits	×	✓	×
FAIR (ours)	Model logits + Q-only logits + V-only logits	✓	✓	FGSM

a generative bias model with adversarial training, but still requires bias supervision derived from predefined bias-aligned data distributions. In contrast, CSS [10] and D-VQA [11] focus on mitigating bias by explicitly modeling visual bias through counterfactual data augmentation or unimodal branches, without directly characterizing the model's internal distribution bias. Unlike these approaches, FAIR models bias directly from the model's output logits, capturing implicit biases learned during training without relying on additional question-type annotations. By jointly leveraging model-level logits together with question-only and vision-only branches, FAIR offers a unified and annotation-free bias modeling strategy that is both adaptive to the model's behavior and compatible with adversarial training, distinguishing it from prior statistic-driven or architecture-specific debiasing methods. The detailed differences among these methods are summarized in Table 1.

Our benchmarks include widely used OOD datasets such as VQA-CP v1 and VQA-CP v2 [12], as well as the ID dataset VQA v2 [13]. Extensive experimental results demonstrate the effectiveness of FAIR in mitigating spurious correlations and improving performance across both ID and OOD datasets. **Our major contributions are as follows:**

- **A Unified Refinement of Ensemble-based Debiasing Frameworks.** We present FAIR, a systematic refinement of existing ensemble-based and pseudo-label-driven debiasing methods for VQA. Rather than introducing an entirely new pipeline, FAIR re-organizes and integrates pseudo-labeling, bias modeling, adversarial perturbation, and loss re-weighting into a coherent two-stage training framework, enabling a more stable trade-off between ID and OOD performance.
- **Logits-based distribution bias Approximation for Pseudo-labeling.** We reinterpret pseudo-labeling as a process of *distribution bias reshaping* and propose to approximate the distribution bias directly from model output logits, instead of relying on annotated question-type statistics. This design provides a lightweight and annotation-free alternative to prior pseudo-label formulations, while empirically achieving comparable effectiveness to methods that require explicit bias annotations.
- **Consistent Empirical Gains on Diverse Datasets and Backbones.** Across multiple commonly used VQA backbones, extensive experiments on VQA-CP v1/v2 and VQA v2 demonstrate that FAIR consistently improves OOD generalization while maintaining competitive ID performance. These results empirically indicate that reshaping distribution bias within existing ensemble-based frameworks can yield robust performance gains without introducing heavy architectural modifications.

The remainder of this paper is organized as follows: Section 2 reviews related work of this paper. Section 3 discusses the VQA bias problem and introduces how FAIR addresses these biases. Section 4 provides a detailed description of the experimental datasets, metrics, results, qualitative analysis, and a visual analysis of the impact of the pseudo-label function. And in Section 6, we present a brief summary of the paper.

2. Related work

2.1. VQA

VQA is a multidisciplinary field that bridges computer vision and natural language processing [14]. Its primary objective is to develop systems capable of answering questions about images by jointly understanding visual content and textual information [15]. This task involves complex reasoning and requires models to process and integrate data from both modalities effectively. The introduction of attention mechanisms marks a critical advancement in VQA. Anderson et al. [16] propose the Bottom-Up and Top-Down Attention (UpDn) model, which enables VQA systems to dynamically attend to relevant image regions, substantially improving performance. By facilitating more fine-grained visual grounding, attention-based approaches enhance the model's ability to generate accurate answers. Further developments include multi-modal fusion techniques, which seek to combine visual and textual features more effectively. Kim et al. [17] introduce the Bilinear Attention Network (BAN), leveraging bilinear pooling to capture complex interactions between image and text features.

Li et al. [18] extract parameter-shared multi-level concepts through multiple fusion modules, optimizing the trade-off between complexity and expressiveness while improving model capacity, and experimental results demonstrate that this approach effectively reduces language priors. BGML [19] introduces a bias model to guide a margin-based loss for explicitly separating biased answers, and integrates adversarial training, knowledge distillation, and contrastive learning to more comprehensively model bias. Mao et al. [20] propose an ensemble-based parameter-insensitive framework consisting of two representation learning branches and a joint learning block. Peng et al. [21] present a Dual Views Interaction Model (DVM) to address language priors in VQA by enhancing the contrast between correct answers and positive/negative predictions through a novel loss function.

Table 2
Key notations and descriptions used in this paper.

Notation	Description
$\mathcal{D} = \{(I_i, Q_i, y_i)\}_{i=1}^{N}$	The VQA dataset
g_q	The question-answering branch
g_v	The vision-answering branch
$f(;\Theta)$	The VQA model parameterized by learnable weights Θ
$g(;\theta_q)$	The question-answering branch parameterized by learnable weights θ_q
$\mathcal{F}$	The unimodal branch prediction
$\hat{\mathcal{F}}_q$	The question-answering branch prediction with adversarial perturbations
$\hat{Q}$	The question with adversarial perturbations
lr	The learning rate
p_{Dis}	The distribution bias
$c(\cdot)$	The classifier
$\mathcal{P}_l$	The pseudo-label function
$\mathcal{T}$	The set of question types set
$\mathcal{V}$	The model vocabulary

2.2. Biases

In recent years, several studies have highlighted the presence of systematic biases in many VQA models [12,16]. Biases primarily stem from severe imbalances in answer distributions within datasets, causing significant performance degradation in tasks such as reasoning and classification [22]. Post-training VQA models tend to disregard image content and instead rely on providing high-frequency answers prevalent in the training set, resulting in spurious correlations. In GGE [8], biases are categorized into shortcut bias and distribution bias. The former refers to the direct prediction of answers by a question-answering (QA) branch based on spurious correlations between questions and answers, while the latter denotes the model's inclination to provide high-frequency answers influenced by the answer distribution in the training set. More generally, for a vision-language model, if the model directly skips visual information and answers a question solely based on textual cues, this behavior is referred to as shortcut bias. D-VQA [11] identifies the presence of visual bias, which is referred to as visual shortcut bias in this work, and introduces a vision-answering (VA) branch to mitigate it. For instance, if the training set contains only images of yellow bananas, the model may incorrectly classify a green banana as yellow during inference.

2.3. Debias methods

Data augmentation. Dataset reconstruction is one of the most common data augmentation strategies. Anderson et al. [16] introduce bottom-up and top-down attention mechanisms that implicitly alleviate bias by encouraging more balanced visual attention. Some methods construct positive and negative samples by randomly masking or selectively masking image and text regions to increase data diversity. Chen et al. [10] propose CSS, which balances data distribution through counterfactual sample construction. These approaches effectively reduce the model's dependency on specific patterns, improving generalization. However, they may contradict the fundamental motivation of OOD benchmarks, as they alter the original data distribution and may inadvertently introduce new biases. Annotation-based methods rely on additional human annotations to ensure that model predictions are grounded in relevant image regions. Selvaraju et al. [4] leverage explanatory supervision to guide models toward correct visual regions, thereby reducing language bias. Similarly, Kv et al. [5] introduce a visually guided question encoder to suppress language bias and improve VQA performance.

Self-supervised learning and generative adversarial networks (GANs). Ramakrishnan et al. [23] employ adversarial regularization to mitigate language priors in VQA. Sheng et al. [24] introduce GANs to model global dataset distributions, thereby reducing label distribution bias. Li et al. [25] propose a multimodal attention mechanism that dynamically adapts to visual and textual inputs, enhancing model robustness. Ren et al. [26] combine adversarial training with multi-task learning to further improve performance.

Ensemble-based methods. Ensemble methods are widely used in VQA to mitigate biases and improve model performance. These methods typically combine multiple models or learning strategies to enhance robustness and generalization. One of the core approaches in ensemble methods is to combine predictions from multiple models. Pan et al. [27] propose a multimodal ensemble framework that integrates knowledge distillation and causal inference to capture diverse data characteristics. Ensemble methods often focus on combining visual and textual representations to mitigate bias. Kim et al. [17] introduce a bilinear attention network that implicitly ensembles multimodal interactions between visual and textual features. Hierarchical and modular ensemble methods further decompose the VQA task into smaller, more manageable components, enabling more structured bias modeling. In addition, some ensemble methods incorporate external knowledge to enhance VQA performance (Table 2).

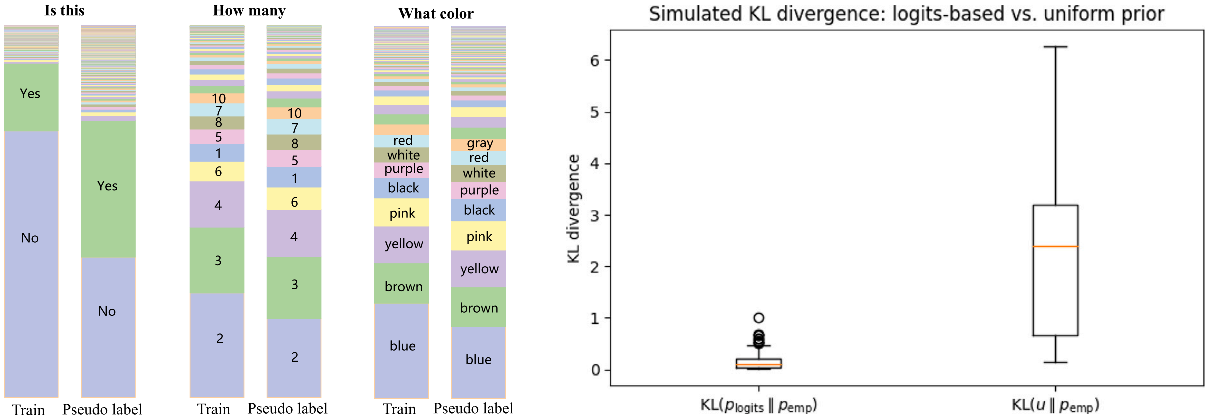


Fig. 2. (Left) The distribution of labels for different question types in the VQA-CP v2 training set and their corresponding pseudo-label distributions. (Right) KL divergence comparison between unimodal logits and empirical bias versus a uniform prior. The logits-based bias proxy exhibits substantially lower KL divergence, supporting the use of logits to estimate dataset-level bias.

3. Method

Section 3.2 delves into the pseudo-label proposed in GGE [8], which introduces a new function to better reshape distribution bias. Section 3.3 employs an improved focal loss as the loss function to encourage the model to focus more on challenging samples. Section 3.4 proposes adding adversarial noise to the text to enhance the model's grounding ability.

3.1. *Preliminaries*

Visual Question Answering (VQA) aims to provide accurate answers to questions based on a given image. Formally, given a VQA dataset $\mathcal{D} = \{(I_i, Q_i, y_i)\}_{i=1}^{N}$, where I_i represents an image, Q_i is a question, and y_i is the corresponding ground truth answer, the goal is to predict the correct answer $\hat{a}_i$ for each pair (I_i, Q_i). This can be expressed as:

$$\hat{a}_i = \arg\max_{y_i \in \mathcal{A}} p(y_i \mid Q_i, I_i; \Theta), \tag{1}$$

where $\mathcal{A}$ is the set of possible answers, Θ represents the model parameters, and $p(y_i \mid Q_i, I_i; \Theta)$ is the probability of the answer y_i given the image I_i and question Q_i. The probability is computed using a classifier $c(\cdot)$, which operates on the combined representation of the image and question features. Specifically, the image features are extracted using a function $g_v(\cdot)$, and the question features are extracted using $g_q(\cdot)$. These features are then fused and mapped to the answer space through a function $f(\cdot)$:

$$p(y_i \mid Q_i, I_i; \Theta) = c\left(f(g_v(I_i), g_q(Q_i))\right). \tag{2}$$

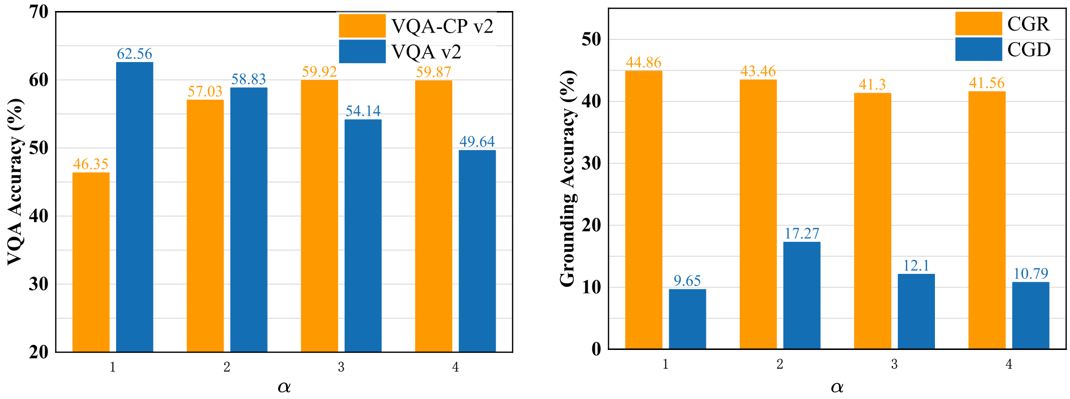
In ensemble-based methods, an additional question-answering (QA) branch is introduced. This branch operates solely on the question modality and directly predicts an answer without using image information. The QA branch consists of a question feature extractor $g_q(\cdot)$ and a classifier $c_q(\cdot)$, and its output can be expressed as:

$$\mathcal{F}_q(y_q \mid Q) = c_q(g_q(Q)). \tag{3}$$

3.2. *Reshaping bias with pseudo-label*

We posit that when a multimodal model receives unimodal input that is insufficient to determine the answer, an unbiased model should express high uncertainty in the output logits. A uniform distribution therefore serves as a conceptual reference for unbiased behavior. However, this does not universally hold; unimodal cues with strong semantic priors (e.g., "What color is the banana?") can legitimately induce non-uniform predictions. Consequently, we treat uniformity not as a strict requirement, but as a reference point that allows us to quantify deviations caused by unimodal bias.

To empirically validate this intuition, we compare unimodal logits against the empirical answer distribution of the training set. For each question type, we construct an empirical distribution bias p_{emp} from normalized answer frequencies, and a logits-based proxy p_{logits} by feeding unimodal input (e.g., text-only) into the model and applying softmax to the logits. We then measureKL($p_{\text{logits}} \parallel p_{\text{emp}}$) and KL($u \parallel p_{\text{emp}}$), where u denotes the uniform distribution. In our experiments with $K = 10$ and 300 question types, the average KL divergence is approximately 0.15 for KL($p_{\text{logits}} \parallel p_{\text{emp}}$) and 2.24 for KL($u \parallel p_{\text{emp}}$), indicating that unimodal logits form a much closer approximation to empirical bias than a uniform prior (see Fig. 2). This supports the use of logits as a proxy for dataset-level bias.



(a) VQA Accuracy (b) Grounding Accuracy

Fig. 3. The variation of α in the pseudo-label from 1 to 4 affects GGE's accuracy. As α increases, GGE's accuracy on VQA-CP v2 gradually improves, but its results on VQA v2, as well as its grounding ability, significantly deteriorate, highlighting a trade-off between dataset-specific accuracy and overall grounding effectiveness.

Motivated by this observation, we adopt a sigmoid-based pseudo-label to reshape biased predictions. Given the predicted distribution y and a distribution bias b, we define:

$$\mathcal{P}_l(y, b) = 2y\sigma(-\alpha yb), \tag{4}$$

where all operations are element-wise. This formulation was originally introduced in GGE with $\alpha = 2$ and later adopted in GenB. Although GGE interprets this form as a negative gradient of Sigmoid + BCE, its exact derivative corresponds to $y - b$, suggesting that the pseudo-label should be viewed as a heuristic suppressing term rather than a direct gradient.

To clarify its behavior, we decompose the pseudo-label into a confidence term y and a suppression factor $2\sigma(-\alpha yb)$. When y and b are simultaneously large, $2\sigma(-\alpha yb)$ decreases, meaning bias-aligned predictions are attenuated. Conversely, when y is large but $b \approx 0$, $2\sigma(-\alpha yb) \approx 1$, preserving confident predictions unsupported by bias. When y is small, $\mathcal{P}_l(y, b)$ remains close to y, avoiding excessive penalization. Thus, $\mathcal{P}_l$ performs bias-aware distribution reshaping by suppressing bias-aligned high-frequency classes while maintaining or relatively enhancing low-bias alternatives.

Fig. 3 illustrates that variations in the parameter α have a significant impact on the model's performance across different datasets. Specifically, as α increases, there is a notable improvement in performance on the VQA-CP v2 dataset, whereas a decline is observed on the VQA v2 dataset. Additionally, these changes in α are associated with variations in the Correct Grounding for Right prediction (CGR) values. This indicates that *arbitrarily changing the reshaping strength does not enhance the model's inherent debiasing ability*. The model does not answer based on the correct image regions but relies on spurious correlations to predict. Based on these observations, we posit that $\mathcal{P}_l$ fundamentally reshapes the distribution bias p_{Dis}, which refers to the answer distributions under the same question type $t \in \mathcal{T}$, where $\mathcal{T}$ is the set of question types. For each question type t, we calculate the occurrence count $C_{t,a}$ of each answer $y_j \in \mathcal{A}$, then compute the relative frequency of each answer as the distribution bias $p_{Dis}(t, y_i)$:

$$p_{Dis}(t, y_i) = \frac{C_{t,y_j}}{\sum_{y'_j \in \mathcal{A}} C_{t,a'_j}}, \tag{5}$$

where

$$C_{t,y_j} = \sum_{i=1}^{N} \mathbb{I}(Q_i \in t) \cdot \mathbb{I}(A_i = y_j). \tag{6}$$

This transformation reduces the proportion of high-frequency answers in the dataset and elevates the proportion of low-frequency answers. To validate this hypothesis, we use p_{Dis} and ground truth from VQA-CP v2 as inputs to the function and sort the obtained pseudo-labels. We observe the distribution of answers before and after the reshaping. As illustrated in Fig. 2, taking questions of the "Is this" type as an example, in the VQA-CP v2 training set, the proportion of "No" answers exceeds half, far surpassing the proportion of "Yes" answers. However, after transformation, the proportions of "No" and "Yes" in the obtained pseudo-labels become very similar. Additionally, the cumulative proportion of other types of answers significantly increases, aligning with our hypothesis.

However, the reliance on annotated data inherently limits model scalability. To address this issue, we propose an approximate estimation method for distribution bias that eliminates the need for labeled data. Specifically, we take the distribution of logits l of a VQA model as the distribution bias:

$$p_{Dis}(t, y_i) = \frac{e^{l_i}}{\sum_{j=1}^{\|\mathcal{V}\|} e^{l_j}}, \tag{7}$$

This approach captures the model's output distribution through training dynamics rather than direct dataset annotation.

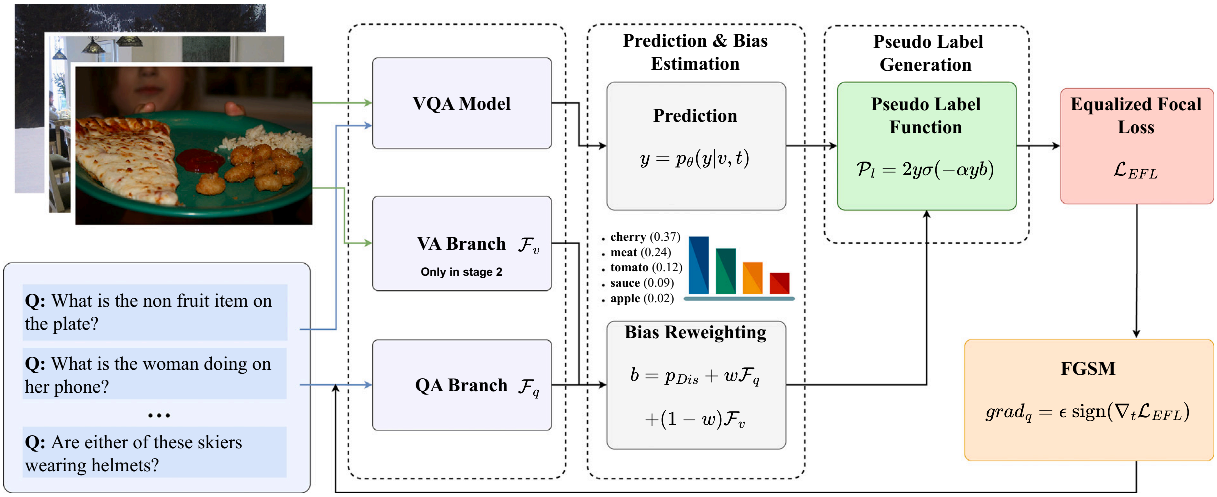

VQA Model
VA Branch $\mathcal{F}_v$
Only in stage 2
QA Branch $\mathcal{F}_q$
Q: What is the non fruit item on the plate?
Q: What is the woman doing on her phone?
...
Q: Are either of these skiers wearing helmets?
Prediction & Bias Estimation
Prediction
$y = p_\theta(y|v,t)$
cherry (0.37)
meat (0.24)
tomato (0.12)
sauce (0.09)
apple (0.02)
Bias Reweighting
$b = p_{Dis} + w\mathcal{F}_q$
$+(1-w)\mathcal{F}_v$
Pseudo Label Generation
Pseudo Label Function
$\mathcal{P}_l = 2y\sigma(-\alpha yb)$
Equalized Focal Loss
$\mathcal{L}_{EFL}$
FGSM
$grad_q = \epsilon\,\mathrm{sign}(\nabla_t \mathcal{L}_{EFL})$

Fig. 4. Overview of FAIR. $\mathcal{F}_q$ and $\mathcal{F}_v$ denote language shortcut and visual shortcut bias, respectively. $grad_q$ denotes the partial derivative of the loss function for the question q.

Our rationale stems from two observations: (1) dataset bias inevitably influences model output distributions, and (2) these output distributions partially reflect inherent dataset characteristics. In the experiments section, we compare the performance under two conditions: one using the true distribution bias and the other using the estimated distribution bias. Our experiments compare model performance using true vs. estimated distribution bias.

3.3. Equalized focal loss

Following the previous research, we consider the VQA task as a multiclass classification problem, utilizing binary cross-entropy (BCE) as the loss function:

$$\mathcal{L}_{BCE}(y,p) = \sum_{i=1}^{N} y_i \log(p_i) + (1-y_i)\log(1-p_i). \tag{8}$$

[7] finds that in the later stages of training, the contribution of hard samples to the loss in VQA is much smaller compared to easy samples, which is one of the reasons for bias. Focal loss [28] can effectively increase the contribution of hard samples to the loss value. It is a variant of binary cross-entropy loss, which can be combined with pseudo-labels. We use EFL [6] as our loss function, which is a more generalized form of focal loss:

$$\mathcal{L}_{EFL}(y,p) = -\sum_{i=1}^{N} \alpha \left(\frac{\gamma + \gamma_v^i}{\gamma}\right)(1-p_t)^{\gamma+\gamma_v^i}\log(p_t), \text{where } p_t = \begin{cases} p, & \text{if } y = 1 \\ 1-p, & \text{if } y = 0 \end{cases}, \tag{9}$$

where α and γ are hyperparameters controlling the class balancing and the focusing strength, respectively. p_t denotes the predicted probability of the ground-truth class. γ_v^i is a sample-dependent modulation term that adaptively adjusts the focusing strength for the i-th sample. In the original EFL formulation, γ_v^i is defined as $\gamma_v^i = s \cdot \gamma$, where s is a scaling factor that reflects the imbalance degree estimated from gradient statistics, enabling category-aware focusing for long-tailed data. EFL extends Focal Loss by dynamically reweighting samples according to their learning difficulty and category imbalance, and has been shown to be more effective in long-tailed scenarios. In our setting, instead of deriving γ_v^i from gradient-based imbalance estimation, we incorporate distribution bias explicitly by defining $\gamma_v^i = s \cdot p_{\text{Dis}}$, where p_{Dis} represents the estimated distribution bias of the sample. By injecting distribution bias into the focusing mechanism of EFL, the loss function becomes more sensitive to biased and underrepresented samples, encouraging the model to allocate more learning capacity to such samples and mitigating the influence of spurious correlations.

3.4. Fast gradient sign method

To enhance the grounding ability of the model and mitigate language prior bias, we incorporate FGSM-based adversarial perturbations into the textual input. The overall training process consists of two stages (Fig. 4), which differ in the modules being optimized and the incorporation of visual grounding.

In **Stage 1**, we focus on modeling distribution bias and question shortcut bias, as illustrated in Fig. 4. Given a question embedding t, we first compute the loss and obtain the gradient $\nabla_t \mathcal{L}$ through a forward and backward pass. Following FGSM [29], we generate adversarial perturbations $grad_q = \epsilon\,\mathrm{sign}(\nabla_t \mathcal{L}_{EFL})$ and inject them into the textual embedding (perturbation ratio $\epsilon = 1$). The perturbed input is fed forward again to produce the logits, which serve as a distribution bias estimate p_{Dis}. We then form pseudo-labels using

Algorithm 1: FGSM Stage 1.

```
Data: Questions $Q$, Distribution bias $p_{Dis}$, Labels $y$
Model: $g_q(Q;\theta_q)$;
Initialize $\mathcal{F}_q$ to 0;
for $batch \leftarrow 1$ to $B$ do
    $\mathcal{F}_q \leftarrow c_q(g_q(Q))$;
    $grad_q \leftarrow \partial\mathcal{L}_{EFL}(\mathcal{P}_l(y, \mathcal{F}_q + p_{Dis}), \mathcal{F}_q)/\partial Q$;
    $\hat{Q} \leftarrow Q + grad_q$;
    $\hat{\mathcal{F}}_q \leftarrow c_q(g_q(\hat{Q}))$;
    Update $\theta_q \leftarrow \theta_q - lr \cdot \nabla\mathcal{L}_{EFL}(\mathcal{P}_l(y, \hat{\mathcal{F}}_q + p_{Dis}), \hat{\mathcal{F}}_q)$
end
Return $y_q = g_q(Q;\theta_q)$;
```

Algorithm 2: FGSM Stage 2.

```
Data: Questions $Q$, Images $I$, Distribution bias $p_{Dis}$, Labels $y$
Model: $g_q(Q;\theta_q)$, $g_v(I;\theta_v)$, $f(I,Q;\Theta)$;
Initialize $\mathcal{F}_q$, $p_{bias}$ to 0;
for $batch \leftarrow 1$ to $B$ do
    $\mathcal{F}_q \leftarrow c_q(g_q(Q))$;
    $\mathcal{F}_v \leftarrow c_v(g_v(I))$;
    $\hat{a} = f(I,Q;\Theta)$;
    $grad_q \leftarrow \partial\mathcal{L}_{EFL}(\mathcal{P}_l(y, \mathcal{F}_q + p_{Dis} + \mathcal{F}_v), \hat{a})/\partial Q$;
    $\hat{Q} \leftarrow Q + grad_q$;
    $\hat{\mathcal{F}}_q \leftarrow c_q(g_q(\hat{Q}))$;
    $p_{bias} \leftarrow p_{Dis} + w \cdot \hat{\mathcal{F}}_q + (1-w) \cdot \mathcal{F}_v$;
    Update $\Theta \leftarrow \Theta - lr \cdot \nabla\mathcal{L}_{EFL}(\mathcal{P}_l(y, p_{bias}), \hat{a})$
end
Return $y_{debias} = f(I,Q;\Theta)$;
```

both p_{Dis} and ground truth and optimize model parameters via a standard backward pass. Stage 1 primarily updates the VQA model while explicitly exposing the model to semantic perturbations that force it to rely less on superficial textual cues.

In **Stage 2**, we introduce the VA branch to further enforce grounding, as shown in Algorithm 2. The VQA model, QA branch and the newly added VA branch are optimized jointly, while FGSM perturbation is still applied to the textual embeddings with the same perturbation ratio. In this stage, the pseudo-label mechanism incorporates both language-driven bias estimates and visual evidence, allowing visual bias modeling to complement the language-bias mitigation learned in Stage 1. The QA branch remains active throughout and its parameters are updated together with the VQA and VA branches. Only the VQA model is required for inference, with no need for any other branches.

We also clarify the role of the weighting terms: γ controls the contribution of pseudo-label reshaping to the overall loss, and we provide an ablation study on γ in Fig. 6(c).

4. Experiments

To demonstrate the effectiveness of our method, we conduct a comprehensive evaluation across three established benchmarks: VQA-CP v1&2 and VQA v2. We utilize CGR and Correct Grounding Difference (CGD) metrics to rigorously assess the grounding ability of our approach. Furthermore, we perform ablation experiments to validate the individual contributions of FAIR's components. Finally, we fine-tune hyperparameters to analyze the influence of reshaping bias on the debiasing process.

4.1. Datasets

We conduct experiments on VQA-CP v1&2 [12] datasets, two highly popular benchmarks for diagnosing biases in VQA. Additionally, we conduct experiments on VQA v2 [13] validation set to verify that FAIR maintains competitiveness in ID datasets. VQA v2 dataset contains a long-tailed distribution of answers with strong biases. The distribution of answers remains consistent across the training, validation, and test sets, and is widely used for evaluating a model's generalization capabilities. The VQA v2 dataset consists of 443 K training samples and 447 K test samples. The VQA-CP v1 and v2 datasets contain 2.5M and 4.4 M training samples, respectively, with 1.3 M and 2.2 M test samples.

VQA-CP v1&2 datasets are derived from VQA v2 but lack a validation set. The training and test sets exhibit incongruent answer distributions, making these datasets suitable for diagnosing whether a model relies on spurious correlations within the dataset to answer.

Table 3
Experimental results (%) on the VQA-CP v2 test set and VQA v2 validation set for SOTA methods. **Bold** font denotes the best result within the column, and numbers <u>underlined</u> indicate the second-best results. FAIR's performance surpasses all other non-data augmentation methods on VQA-CP v2, and even shows a slight improvement in accuracy compared to the base model UpDn on VQA v2. † indicates that the distribution bias is estimated from the output logits. We conducted the experiment ten times and recorded the mean and variance.

Method	Base Model	VQA-CP v2					VQA v2				Avg.
		Overall	Y/N	Num.	Others	CGD	Overall	Y/N	Num.	Others	
GVQA [12]	–	31.30	57.99	13.68	22.14	–	48.24	72.03	31.17	34.65	39.77
UpDn [16]	–	39.89	43.01	12.07	45.82	3.91	63.79	80.94	42.51	55.78	51.84
S-MRL [30]	–	38.46	42.85	12.81	43.20	–	63.10	–	–	–	50.78
Modifying language module											
VGQE [5]	UpDn	48.75	–	–	–	–	<u>64.04</u>	–	–	–	56.40
Weakening language prior											
Re-Scaling [7]	UpDn	47.09	68.42	21.71	42.88	–	55.50	64.22	39.61	53.09	51.30
AReg [23]	UpDn	41.17	65.49	15.48	35.48	–	62.75	79.84	42.35	55.16	51.96
CF-VQA(SUM) [31]	UpDn	53.55	**91.15**	13.03	44.97	–	63.54	<u>82.51</u>	<u>43.96</u>	54.30	58.55
CF-VQA(SUM) [31]	S-MRL	55.05	<u>90.61</u>	21.50	45.61	–	60.94	81.13	43.86	50.11	58.00
PW-VQA [32]	UpDn	59.06	88.26	<u>52.89</u>	45.45	–	62.63	81.80	43.90	53.01	60.85
Strengthening visual attention											
HINT [4]	UpDn	47.50	67.21	10.67	46.80	10.34	63.38	81.18	42.14	55.66	55.44
RUBi [30]	UpDn	45.42	63.03	11.91	44.33	6.27	58.19	63.04	41.00	54.43	51.81
LM [33]	UpDn	48.78	70.37	14.24	46.42	11.33	63.26	81.16	42.22	55.22	56.02
LMH [33]	UpDn	52.73	72.95	31.90	47.79	10.60	56.35	65.06	37.63	54.69	54.54
SCR [34]	UpDn	49.45	72.36	10.93	48.02	–	62.2	78.8	41.6	54.4	55.83
Ensemble-based methods											
MDDC [25]	UpDn	54.70	83.58	19.93	49.10	–	63.33	81.64	42.56	54.88	59.00
GGE [8]	UpDn	57.32	87.04	27.75	49.59	15.27	59.11	73.27	39.99	54.39	58.22
GGD [35]	UpDn	59.37	88.23	38.11	49.82	13.31	62.15	79.25	42.43	54.66	60.76
GenB [9]	UpDn	59.15	88.03	40.05	49.25	–	–	–	–	–	–
CMQEF [18]	UpDn	40.88	41.40	12.73	47.81	–	**65.51**	**82.89**	**47.71**	<u>56.64</u>	53.20
BGML [19]	UpDn	<u>62.28</u>	89.82	51.31	**53.87**	–	60.84	77.81	37.81	54.05	61.56
CL [20]	UpDn	57.45	82.45	39.76	49.21	–	63.94	81.23	40.03	**57.13**	60.70
DCJ [21]	UpDn	61.24	90.02	51.05	48.95	–	60.98	77.67	37.51	54.51	61.11
BILI [36]	UpDn	59.76	88.20	41.76	49.79	–	–	–	–	–	–
MSB-VQA [37]	UpDn	**62.40**	89.01	**55.12**	<u>50.46</u>	–	60.89	77.58	37.24	54.28	61.65
FAIR (Ours)	UpDn	61.14 ± 0.08	87.59 ± 0.09	50.76 ± 0.06	50.20 ± 0.10	**16.37** ± 0.11	63.81 ± 0.17	80.52 ± 0.15	43.72 ± 0.13	56.49 ± 0.16	<u>62.48</u> ± 0.11
FAIR† (Ours)	UpDn	60.96 ± 0.16	87.89 ± 0.19	50.35 ± 0.14	50.01 ± 0.10	<u>15.66</u> ± 0.13	64.03 ± 0.12	80.82 ± 0.15	43.78 ± 0.12	56.61 ± 0.08	**62.50** ± 0.15

4.2. Metrics

We adopt the standard VQA evaluation metric for estimating the model's capacity. The accuracy is computed as:

$$Acc = \min\left(\frac{\#humananswers}{3},\ 1\right), \tag{10}$$

where $\#human\ answers$ is the number of times each answer is annotated by humans for the question. Specifically, if the predicted answer matches three or more human-provided answers, the score is 1. If it matches two or one human answers, the score is 2/3 or 1/3, respectively. Otherwise, the score is 0. Solely evaluating the model's accuracy is insufficient. Instead, we aim for the model to predict correct answers based on the accurate image regions, thereby demonstrating grounding ability. To evaluate the model's grounding ability, we adopt the metrics CGR, and CGD proposed by GGE:

$$\%CGR = \frac{N_{rg,rp}}{N_{rp}} \times 100\%,\ \%CGD = \left(\frac{N_{rg,rp}}{N_{rp}} - \frac{N_{rg,wp}}{N_{wp}}\right) \times 100\%, \tag{11}$$

where N_{rp} denotes the number of samples predicted correctly, $N_{rg,rp}$ denotes the number of samples for which the model made correct predictions based on the correct image regions, N_{wp} is the number of samples predicted incorrectly, and $N_{rg,wp}$ is the number of samples for which the model made incorrect predictions based on the correct image regions.

4.3. Implementation details

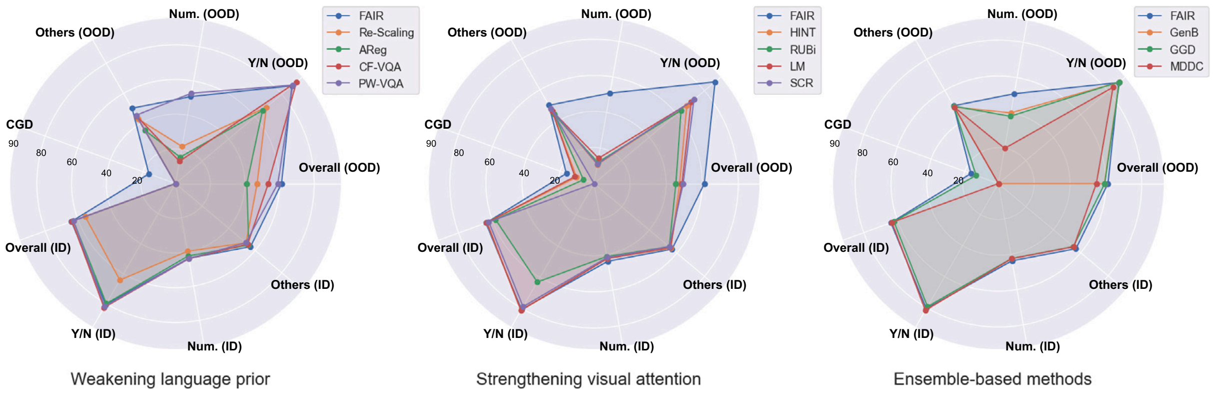
We adopt UpDn as the base model, which is highly sensitive to biases. Table 4 showcases the hyperparameter settings used during training on the VQA-CP v2 and VQA v2 datasets. Unless otherwise specified, our default parameter settings remain unchanged.

4.4. Baselines

We compare FAIR with the state-of-the-art (SOTA) methods (total of 26), including:

Table 4
Hyperparameter settings on VQA-CP v1&2 and VQA v2.

Parameter	VQA-CP v1&2	VQA v2
learning rate	0.001	0.001
batch size	512	512
epoch	25	25
α	2	1
s	0.3	0.1
γ	2	1
ϵ	1	1
w	0.7	0.7



Weakening language prior Strengthening visual attention Ensemble-based methods

Fig. 5. An intuitive comparison of FAIR with three types of methods: weakening language prior, strengthening visual attention, and ensemble-based methods. This figure summarizes the results presented in Table 3.

- **Base models:** GVQA [12], UpDn [16], S-MRL [30], LXMERT [38], LLaVA1.5-7b [39]. Both UpDn and GVQA are models that employ Faster R-CNN combined with an attention mechanism. LXMERT is a BERT-based model that incorporates cross-attention to introduce visual capabilities. LLaVA connects CLIP with a pre-trained large language model using a simple linear layer and achieves remarkable performance after fine-tuning.
- **Methods based on modifying language module: VGQE** [5].
- **Methods based on weakening language prior: Re-Scaling** [7], **AReg** [23], **CF-VQA** [31], **PW-VQA** [32].
- **Methods based on strengthening visual attention: HINT** [4], **RUBi** [30], **LMH** [33], **LM** [33], **SCR** [34].
- **Ensemble-based methods: GGE** [8], **GenB** [9], **CMQEF** [18], **BGML** [19], **CL** [20], **DCJ** [21], **MDDC** [25], **GGD** [35], **BILI** [36], MSB-VQA [37].
- **Data augmentation methods: CSS** [10], **D-VQA** [11], **RandImg** [40], **Mutant** [41], **CVL** [42], **KDDAug** [43].

4.5. Main results

Accuracy on VQA-CP v2. Fig. 5 summarizes the main results of FAIR. Table 3 showcases that FAIR achieves an impressive accuracy of 60.96% on the VQA-CP v2 dataset *without any visual annotations*, marking a remarkable improvement of 21.07% in accuracy and 11.75% in CGD over the base model UpDn. For the "Other" category of questions, FAIR attains an accuracy of 50.01%, surpassing all other non-augmentation methods. In the particularly challenging "Num." category, our accuracy reaches 50.35%, outperforming the previous best method, GenB, by a significant margin of 10.30%. We also estimate the distribution bias using the model's output logits distribution, completely eliminating the need for additional textual annotations. The results demonstrate that, compared to using annotated question types, this approach achieves highly comparable performance on both ID and OOD datasets. This further indicates the feasibility of estimating distribution bias during training through the logits distribution.

Robustness & Grounding. Compared to GGE, FAIR demonstrates superior performance, with a 3.64% higher accuracy and a 0.39% higher CGD. This underscores FAIR's enhanced capability to mitigate dataset biases and improve the model's grounding ability. Additionally, while most debiasing methods experience a significant drop in accuracy on ID datasets after improving performance on OOD datasets, FAIR maintains higher accuracy on the VQA v2 dataset compared to other ensemble-based methods. Specifically, FAIR shows a 0.24% and 4.92% improvement in accuracy over UpDn and GGE, respectively. Under the "Num." and "Other" question types, FAIR achieves 43.78% and 56.61% accuracy, respectively. These experimental results clearly demonstrate that FAIR significantly enhances the model's robustness and grounding ability.

FAIR vs. Data Augmentation. In Table 5, we compare FAIR with various data augmentation techniques. *FAIR is also highly competitive when compared to augmentation methods. For instance, when compared to the best data augmentation method, D-VQA, our*

Table 5
Comparison with data augmentation Methods on VQA-CP v2 and VQA v2. FAIR is highly competitive, lagging only 0.74% behind the best-performing method, D-VQA, which combines ensemble-based techniques and data augmentation.

Method	Base Model	VQA-CP v2				VQA v2			
		Overall	Y/N	Num.	Others	Overall	Y/N	Num.	Others
FAIR (Ours)	UpDn	61.17	87.66	50.89	50.11	63.89	80.57	43.75	56.53
CVL [42]	UpDn	42.12	45.72	12.45	48.34	–	–	–	–
CSS [10]	LMH	58.95	84.37	49.42	48.21	59.91	73.25	39.77	55.11
RandImg[40]	UpDn	55.37	83.89	41.60	44.20	57.24	76.53	33.87	48.57
D-VQA [11]	UpDn	**61.91**	**88.93**	52.32	50.39	**64.96**	**82.18**	**44.05**	**57.54**
Mutant [41]	UpDn	61.72	88.90	49.68	**50.78**	62.56	82.07	42.52	53.28
KDDAug [43]	UpDn	60.24	86.13	**55.08**	48.08	62.86	80.55	41.05	55.18

Table 6
Experimental results (%) on VQA-CP v1. FAIR significantly improves the robustness of UpDn, surpassing all known methods (to our best knowledge) reported on this dataset.

Method	Base model	VQA-CP v1				Δ Gap
		Overall	Y/N	Num.	Other	
UpDn	–	36.38	42.72	42.14	40.35	–
FAIR (Ours)	UpDn	**64.39**	**86.55**	**48.99**	48.54	+**28.01**
AReg [23]	UpDn	62.75	79.84	42.35	55.19	+26.37
RUBi [30]	UpDn	58.19	63.04	41.00	54.43	+21.81
CF-VQA(SUM) [31]	UpDn	63.65	82.63	44.01	54.38	+27.27
GenB [9]	UpDn	62.74	86.18	43.85	47.03	+26.36
CSS [10]	UpDn	60.95	85.60	40.57	**57.03**	+24.57

Table 7
Ablation results (%) on VQA-CP v2 test set. In the case where the loss function is sigmoid + BCE, we perform ablation studies on the pseudo-label equation, EFL, and FGSM. When the loss function is softmax + CE, we ablate the pseudo-label equation and FGSM.

Loss function	Module	Accuracy	CGR	CGD
Sigmoid + BCE	UpDn	39.89	–	–
	+ Pseudo	57.03	43.46	**17.27**
	+ Pseudo + EFL	61.00	78.71	15.98
	+ Pseudo + EFL + FGSM	**61.17**	**81.36**	16.35
Softmax + CE	+ Pseudo	47.32	56.36	3.17
	+ Pseudo + FGSM	46.87	56.05	− 0.07

approach is only slightly behind by 0.95% on VQA-CP v2. It is noteworthy that D-VQA combines data augmentation with an ensemble-based method, whereas FAIR achieves similar performance without any data augmentation techniques. It is important to note that data augmentation often contradicts the original intent behind constructing OOD datasets, as it manipulates these datasets directly rather than enabling the model to learn unbiased knowledge from biased data.

Accuracy on VQA-CP v1. Table 6 evaluates the performance of FAIR on the VQA-CP v1 dataset, comparing it with selected SOTA methods. The ΔGap indicates the improvement in *Overall Accuracy* achieved by FAIR over the base model. FAIR significantly enhances the accuracy of the base model, surpassing all known methods tested on this benchmark, thus demonstrating the robustness and effectiveness of our approach across different datasets.

Our extensive evaluations show that FAIR is effective in mitigating bias and improving grounding in VQA models. The method reshapes bias distributions and incorporates adversarial training to strengthen robustness. As a result, FAIR achieves notable gains on standard benchmarks and maintains consistent improvements across challenging categories. Moreover, it does not require additional data augmentation or manual annotations, underscoring the value of a debiasing strategy that remains lightweight and practical. Overall, FAIR offers a versatile and powerful approach for advancing VQA systems.

4.6. Ablations

We conduct ablation experiments on VQA-CP v2 to validate the contribution of each module to debiasing and grounding. As shown in Table 7, we evaluate two settings: Sigmoid + BCE and Softmax + Cross-Entropy (CE). The Sigmoid + BCE configuration yields significantly stronger debiasing performance. This is likely due to the multi-label nature of many VQA answers. For instance, questions such as “What objects are in this scene?” may correspond to multiple valid responses (e.g., “cat” and “table”). Sigmoid activation with BCE loss allows independent probability estimation for each label, making it well-suited for such cases. In contrast, Softmax assumes a single-label output and therefore restricts performance when multiple answers are correct. Within the Softmax + CE

Table 8
In ablation studies using different base models, FAIR consistently improves accuracy on OOD datasets without modifying the underlying structure of the VQA models, the improvement is most pronounced for LXMERT. Compared with supervised fine-tuning (SFT), FAIR can bring a 4.96% improvement.

Method	VQA-CP v2				Δ Gap
	Overall	Y/N	Num.	Other	
UpDn [16]	39.89	43.01	12.07	45.82	+21.28
UpDn+FAIR	61.17	87.66	50.89	50.11	
S-MRL [30]	38.46	42.85	12.81	43.20	+22.22
S-MRL+FAIR	60.68	87.05	50.53	49.64	
BAN[17]	37.35	41.96	12.08	41.71	+22.07
BAN+FAIR	59.42	88.53	50.99	46.47	
LXMERT [38]	46.23	42.84	18.91	55.51	+**23.66**
LXMERT+FAIR	69.89	87.81	63.46	60.10	
LLaVA [39]	74.33	91.71	72.11	68.48	+4.96
LLaVA+FAIR	**79.29**	**92.41**	**73.26**	**72.13**	

Table 9
In debiasing experiments on LXMERT, FAIR slightly outperforms data augmentation method D-VQA and ranks second only to GenB.

Method	Base Model	VQA-CP v2			
		Overall	Y/N	Num.	Others
MUTANT [41]	LXMERT	69.52	**93.15**	**67.17**	57.78
D-VQA [11]		69.75	80.43	58.57	**67.23**
SAR [34]		62.12	85.14	41.63	55.68
GenB [9]		**71.16**	92.24	64.71	61.89
MDDC [25]		69.77	87.88	52.80	64.93
Re-Scaling [7]		69.47	86.82	45.03	64.47
FAIR (Ours)		69.89	87.81	63.46	60.10

Table 10
Using S-MRL as the base model. FAIR significantly enhances the OOD performance of S-MRL, achieving the best OOD accuracy compared to other methods.

Method	Base Model	VQA-CP v2			
		Overall	Y/N	Num.	Others
VGQE [5]	S-MRL	50.11	66.35	27.08	46.77
RUBi [30]		47.11	68.65	20.28	43.18
CF-VQA(SUM) [31]		55.05	**90.61**	21.50	45.61
PW-VQA [32]		60.26	88.09	**59.13**	45.99
GGE [8]		54.03	79.66	20.77	46.72
FAIR (Ours)		**60.68**	87.05	50.53	**49.64**

setting, introducing the pseudo-label yields an accuracy improvement of 18.14%. Incorporating EFL leads to a further gain of 3.97%, demonstrating the complementary effect of these modules.

Module Ablation. Notably, performance on CGR shows a significant improvement of 35.25%, while CGD is 1.29% worse than GGE. This indicates that EFL significantly enhances the accuracy of the model's answers based on correct image regions, yet also improves the accuracy of answers based on incorrect image regions, suggesting that the model still learns spurious correlations. Additionally, when adding FGSM, there is a marginal increase of 0.17% in accuracy on VQA-CP v2. Both CGR and CGD increase by 2.65% and 0.37%, respectively. The experimental results suggest that FGSM can partially suppress spurious correlations and enhance debiasing capabilities as well as the model's robustness.

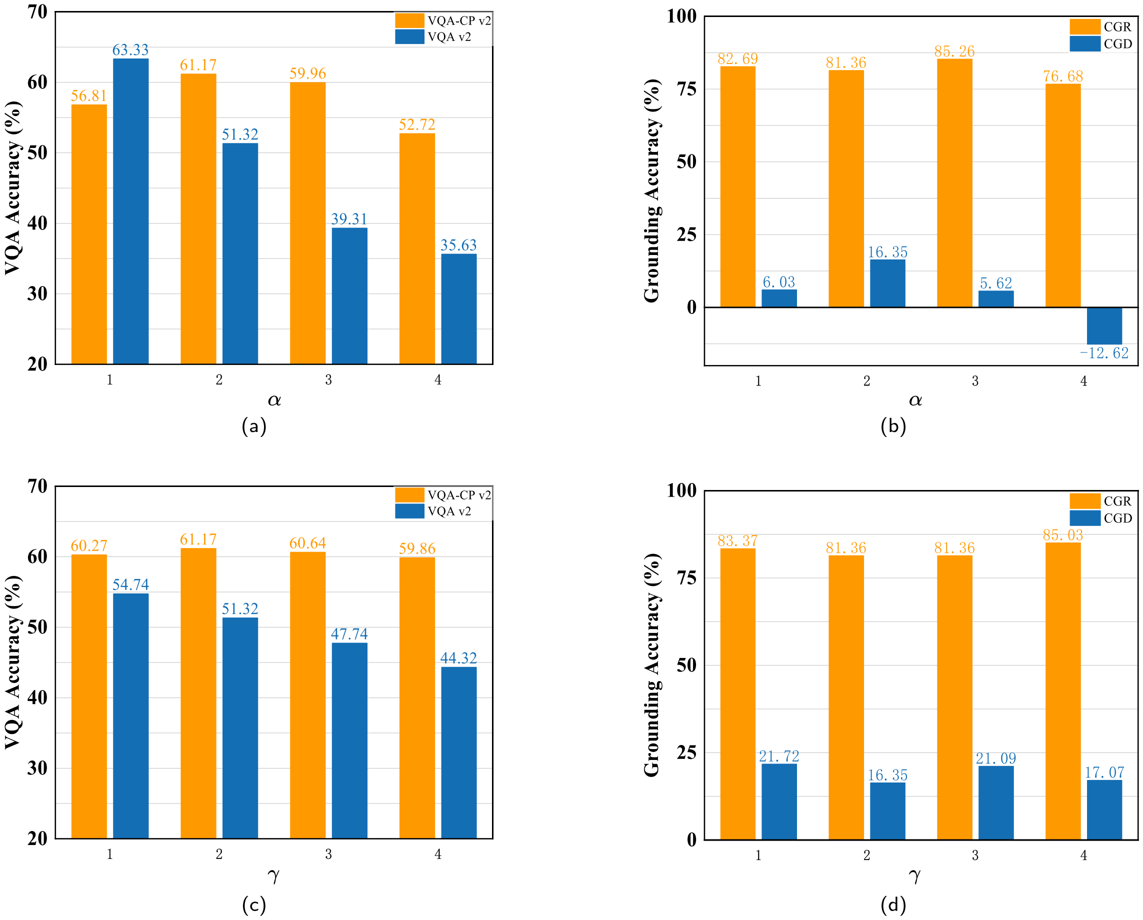
Base Model Ablation. Our approach can be conveniently applied to other base models, as shown in Table 8. In addition to UpDn, we also explore the enhancements FAIR brings to various base models. FAIR significantly enhances the performance of UpDn, S-MRL, BAN, and LXMERT on the OOD dataset. Notably, LXMERT [38], a multimodal variant of BERT [44], is fine-tuned on VQA-CP v2 for 4 epochs. More details are available in Section 4.7. LLaVA achieves outstanding performance due to its training on extensive vision-language data and its large parameter size of 7B. By comparing LoRA [45] fine-tuning with FAIR training, we observe that our approach further enhances performance on the VQA task.

4.7. Base model comparison

In this study, we undertake a detailed comparative analysis of various methods applied across multiple foundational models within the context of the VQA-CP v2 dataset. This comprehensive evaluation is meticulously documented in Tables 9–11, which present the

Table 11
Using BAN as the base model. We compare FAIR with existing approaches that use BAN as the base model and find that it significantly outperforms both GenB and GGE by a substantial margin.

Method	Base Model	VQA-CP v2			
		Overall	Y/N	Num.	Others
GenB [9]	BAN	57.37	**89.11**	29.52	48.37
GGE [8]		50.75	74.56	20.59	46.54
FAIR (Ours)		**59.42**	88.53	**46.47**	**50.99**



(a) (b) (c) (d)

Fig. 6. The first and second rows of images respectively illustrate the impact of the hyperparameters α and γ on model performance. From left to right, they show the changes in VQA accuracy and grounding accuracy as the hyperparameters increase. The second row shows the impact of the hyperparameter γ on model performance. Both α and γ significantly affect all metrics we evaluate, with the model being more sensitive to the value of α. This sensitivity is because α directly participates in reshaping bias.

performance metrics of the three models under consideration: LXMERT, S-MRL, and BAN. With UpDn, S-MRL, and BAN as base models and default parameter settings, FAIR requires less than 12GB of GPU memory. When using LXMERT, the model requires no more than 20GB of GPU memory and the training time does not exceed 8 h on a single 3090 GPU. The LoRA training for LLaVA is conducted using two RTX 4090 GPUs. To reduce additional computational overhead, we collect gradients for one-fifth of the samples during the FGSM process.

FAIR distinguishes itself by delivering superior performance on the S-MRL and BAN models, as highlighted by the comparative data. Specifically, our approach not only outperforms competing methods on these models but also establishes new benchmarks for accuracy and efficiency in processing visual and textual cues within the VQA framework. On the LXMERT model, however, FAIR secures the second position. It closely trails behind GenB, which currently sets the performance standard for this model.

This pattern of results underscores the efficacy of FAIR, particularly in enhancing model responsiveness and accuracy across different architectures within the challenging VQA-CP v2 dataset. Furthermore, it suggests areas for potential refinement in our approach, especially in terms of optimizing performance on the LXMERT model to possibly surpass GenB in future evaluations. The

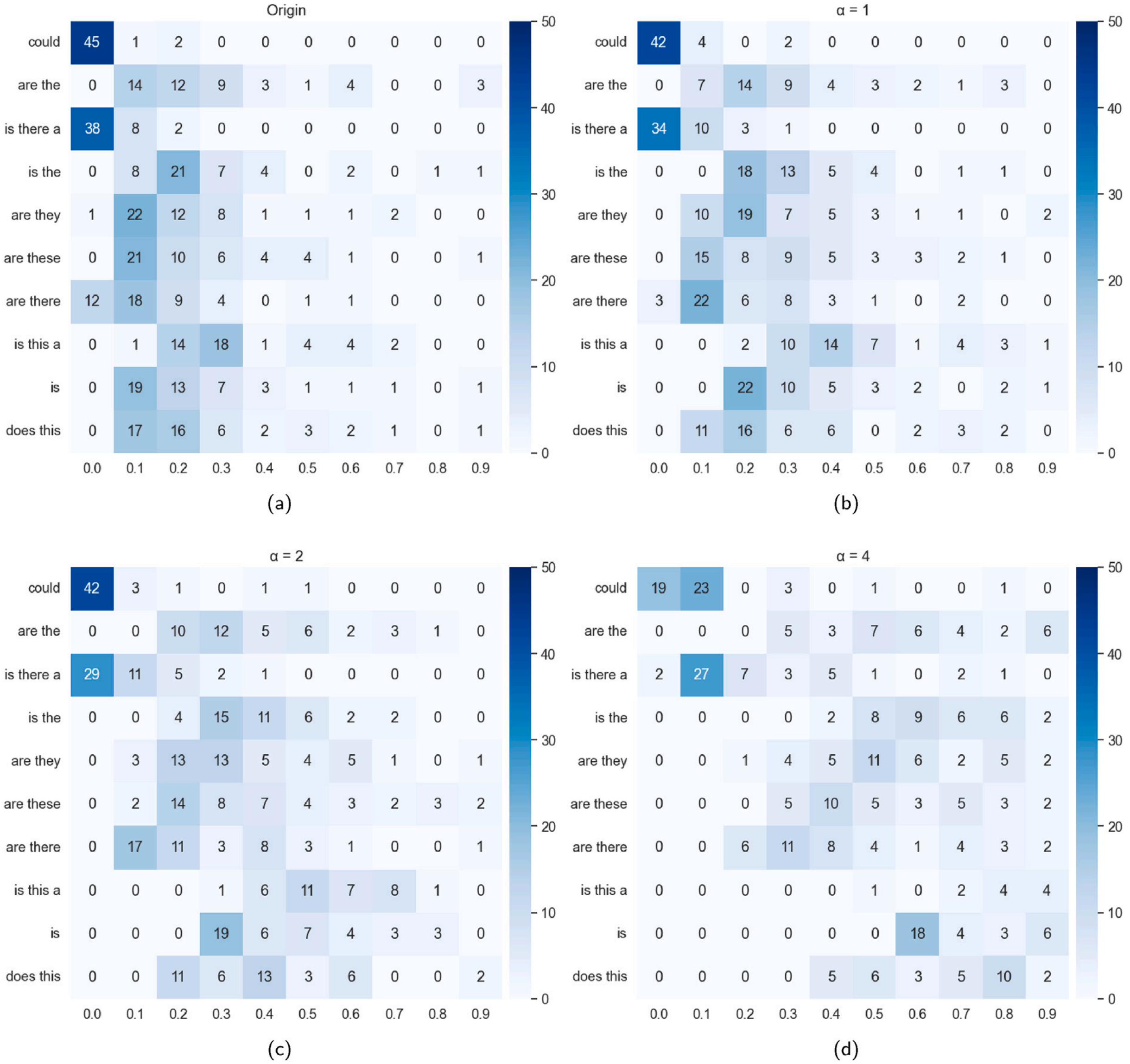


Fig. 7. Heatmaps illustrating the distribution of answer labels for 10 selected question types in the VQA-CP v2. The horizontal axis represents the probability values of the model's answers to different question types, and the vertical axis denotes the question types in the dataset. The probability range of 0 to $2e-3$ is divided into 10 sub-intervals, with labels closer to the left side of the heatmaps being rarer. The impact of varying α values (1, 2, 4) on the pseudo-label distribution is shown, highlighting the reduction in rare labels and the smoothing of the overall label distribution as α increases.

insights gained from this comparative study are pivotal for guiding future research and development efforts aimed at advancing the SOTA in VQA systems.

4.8. Impact of hyperparameters

In this section, we conduct experiments to explore the effects of two crucial hyperparameters, α and γ, on the performance of our model. The hyperparameter α is instrumental in adjusting the distribution bias within the model, while γ determines the relative weighting of losses attributed to hard versus easy samples in the training dataset.

Our findings, as illustrated in Fig. 6(a) and (b), demonstrate that the model's performance is notably sensitive to variations in α. When other hyperparameters are maintained at their default values, increasing α leads to a marked degradation in performance on the VQA v2 dataset. Optimal results on the VQA-CP v2 dataset are achieved when α is set to 2. Moreover, the influence of α extends to the model's capability to generalize across different distributions, as evidenced by its performance on the CGD metric. Specifically, setting α to 4 resulted in a CGR of 76.68%, but this configuration also produced a negative CGD value, underscoring the emergence of significant spurious correlations within the model.

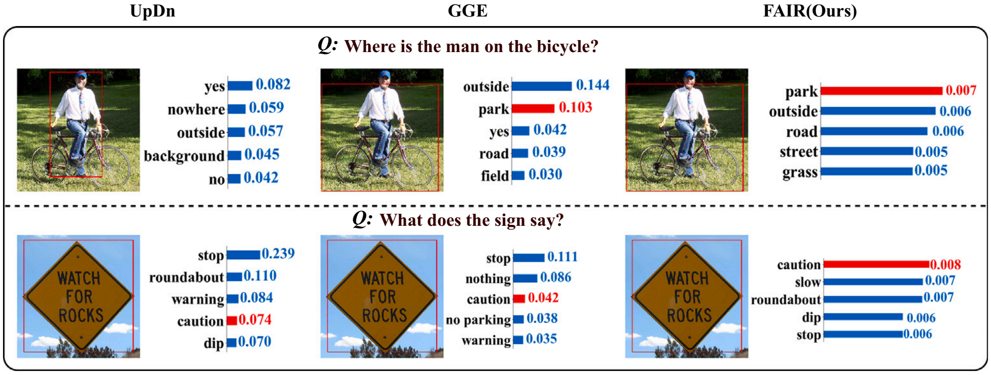


Fig. 8. Qualitative analysis for FAIR. We compare FAIR, UpDn, and GGE models by visualizing the regions of interest during our model's inference.

Further analysis provided in Fig. 6(c) and (d) reveals that γ plays a less pronounced role in the performance on the VQA-CP v2 dataset, yet it substantially affects the model's grounding capabilities. With γ set to 4, the model not only achieved an impressive CGD of 86.26% but also recorded a CGR of 17.07%. These results highlight the delicate balance required in tuning γ to enhance the model's ability to discriminate between more and less challenging samples, thereby improving its overall grounding accuracy. The scaling term s shows minimal sensitivity in our experiments (varying s from 0.1 to 0.6 yields less than 0.22 difference on both ID and OOD results), thus we primarily focus on the more influential γ and α.

4.9. Visualization of the effect of the pseudo-label function

This section visualizes the pseudo-label function's effect on label distributions to enhance methodological interpretability. Using the VQA-CP v2 training set, we perform detailed statistical profiling of answer distributions across question types. For 10 representative question types, we extract the top 50 most probable labels and partition their probability range (0 to 2×10^{-3}) into 10 equidistant intervals. Fig. 7 displays the resulting heatmaps, where leftward-positioned labels indicate lower-frequency responses within the dataset's long-tail distribution.

The distribution bias values serve as input to our pseudo-label function for generating transformed distributions under varying α parameters (1, 2, 4). The corresponding heatmaps demonstrate α's smoothing effect: as α increases, we observe (1) reduced density in rare-label intervals, and (2) probability mass redistribution toward the distribution tail. This adaptive reshaping produces more uniform label distributions while preserving relative frequency relationships.

From the heatmaps, we can observe that higher values of α lead to a more balanced distribution of labels. This adjustment ensures that the rare labels, which are often underrepresented, receive a higher probability, thereby smoothing out the distribution. This smoothing effect is crucial for improving the model's ability to generalize across different types of questions and answer distributions, as it reduces the impact of spurious correlations and biases that typically arise from an imbalanced dataset. The process of adjusting α and observing its impact on the distribution through heatmaps provides valuable insights into how the distribution bias can be reshaped to achieve a more equitable representation of all labels.

4.10. Qualitative analysis

As illustrated in Fig. 8, we show two cases for qualitative analysis of the model's grounding ability (More cases can be found in Section 4.10). The red bounding box indicates the most relevant image region for the model's predicted answer. We compare FAIR with UpDn and GGE. The red bounding box regions represent the image regions upon which the model's predictions are based. For the second question "What does the song say?", although UpDn, GGE, and FAIR attend to the correct image region, only FAIR can answer precisely.

In Fig. 9, we provide a more detailed case study analysis featuring the UpDn, GGE, and FAIR models, particularly examining their performance on specific questions. For the question "Where do these animals live?", UpDn correctly identifies the habitat of the animals. However, it fails to accurately localize the relevant image region, highlighting a discrepancy in its attention mechanism. In contrast, GGE, while accurately focusing on the correct image region, derives an incorrect answer, suggesting a misalignment in its interpretative processing. On the other hand, the FAIR model excels by not only concentrating on the correct image region but also providing the correct response, showcasing its effective integration of visual and contextual cues. In another scenario involving the question "What color is the man's hat?", UpDn continues to exhibit difficulties in focusing on the pertinent image area. GGE, although focusing correctly, erroneously identifies a woman's hat as the object in question, demonstrating a misdirection in its contextual understanding. Conversely, FAIR demonstrates its superiority and robustness by accurately pinpointing and responding to the query about the man's hat, reaffirming its superior analytical capabilities.

These instances illustrate the varied capabilities and challenges of each model in handling complex visual and textual interplays, providing valuable insights into their operational strengths and areas for improvement. This analysis not only informs potential enhancements in model design but also underscores the importance of robust and accurate visual attention in the field of VQA.

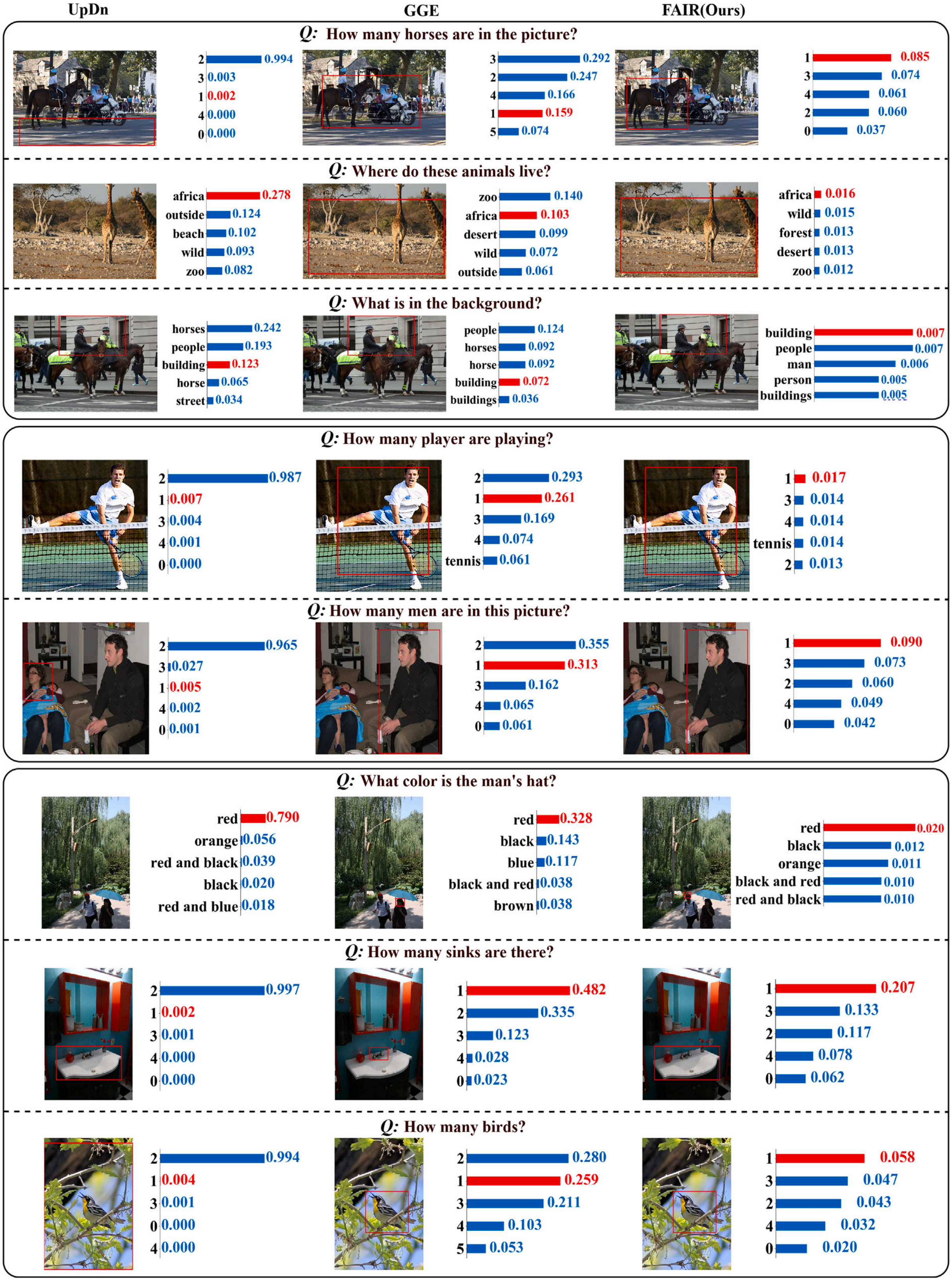


Fig. 9. More qualitative analysis.

5. Discussion

While FAIR demonstrates strong effectiveness in VQA tasks, several limitations warrant discussion. First, the sensitivity of the pseudo-label parameter α introduces dependency on dataset-specific bias characteristics; when bias severity is unknown, iterative tuning may be required and improper calibration could induce excessive distributional shift. A potential mitigation strategy is preliminary sampling-based estimation of dataset bias, which can facilitate selecting reasonable α values prior to full-scale training. Second, similar to most baselines, our evaluation primarily relies on mainstream VQA-CP v1&2 datasets, and broader validation on more diverse bias settings remains a direction for future work.

Beyond these limitations, FAIR is compatible with larger transformer-based architectures and LVLMs because it does not alter network structures or introduce inference-time modules. We validate this on LLaVA-1.5-7B, where FAIR yields consistent OOD improvements. The additional cost arises mainly from FGSM-based perturbation during training and is moderate, while inference overhead remains negligible. These results suggest that FAIR can scale to modern LVLMs with manageable training overhead and maintain efficiency at inference.

6. Conclusion

This paper presents a novel debiasing framework FAIR that introduces a pseudo-label function capable of transforming ground truth distributions by filtering out bias. Unlike existing approaches that rely on annotated question-type distributions in VQA v2 and VQA-CP v2 datasets—a requirement that severely limits LLM adaptation—we pioneer the use of model logits distributions as effective bias proxies during training. Experimental validation with LLaVA models confirms the efficacy of this approach. Furthermore, we enhance visual grounding capabilities through question-specific adversarial noise injection. The framework demonstrates strong transferability across architectures, with promising future applications in addressing challenges such as LLM hallucination.

CRediT authorship contribution statement

Chao Wang: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Weiwei Fu:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Haoyang Li:** Writing – original draft, Methodology, Data curation. **Linqi Ye:** Writing – review & editing, Writing – original draft, Methodology. **Yang Zhou:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank the anonymous reviewers for their valuable comments. This work was supported by the National Key Research and Development Program of China (Grant No. 2022ZD0160603), and the Natural Science Foundation of Shanghai (No. 23ZR1422800).

Data availability

Data will be available upon request.

References

[1] X. Liang, D. Wang, B. Jing, Z. Jiao, R. Li, R. Liu, Q. Miao, Q. Wang, Fine-grained knowledge fusion for retrieval-augmented medical visual question answering, Inf. Fusion 120 (2025) 103059.

[2] C. Wang, J. Yang, Q. He, Event extraction for visual: De-biasing with causality-guided attention mechanism, Neurocomputing 649 (2025) 130783.

[3] C. Chen, D. Han, Z. Guo, C.-C. Chang, Towards bias-aware visual question answering: rectifying and mitigating comprehension biases, Expert Syst. Appl. 264 (2025) 125817.

[4] R.R. Selvaraju, S. Lee, Y. Shen, H. Jin, S. Ghosh, L. Heck, D. Batra, D. Parikh, Taking a hint: leveraging explanations to make vision and language models more grounded, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2591–2600.

[5] G. Kv, A. Mittal, Reducing language biases in visual question answering with visually-grounded question encoder, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16, Springer, 2020, pp. 18–34.

[6] B. Li, Y. Yao, J. Tan, G. Zhang, F. Yu, J. Lu, Y. Luo, Equalized focal loss for dense long-tailed object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 6990–6999.

[7] Y. Guo, L. Nie, Z. Cheng, Q. Tian, M. Zhang, Loss re-scaling VQA: revisiting the language prior problem from a class-imbalance view, IEEE Trans. Image Process. 31 (2021) 227–238.

[8] X. Han, S. Wang, C. Su, Q. Huang, Q. Tian, Greedy gradient ensemble for robust visual question answering, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1584–1593.

[9] J.W. Cho, D.-J. Kim, H. Ryu, I.S. Kweon, Generative bias for robust visual question answering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 11681–11690.

[10] L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, Y. Zhuang, Counterfactual samples synthesizing for robust visual question answering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10800–10809.

[11] Z. Wen, G. Xu, M. Tan, Q. Wu, Q. Wu, Debiased visual question answering from feature and sample perspectives, Adv. Neural Inf. Process. Syst. 34 (2021) 3784–3796.

[12] A. Agrawal, D. Batra, D. Parikh, 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, 2018, pp. 4971–4980.

[13] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, 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, 2017, pp. 6904–6913.

[14] C. Wang, J. Yang, Y. Zhou, X. Yue, CooKie: commonsense knowledge-guided mixture-of-experts framework for fine-grained visual question answering, Inf. Sci. 695 (2025) 121742.

[15] C. Wang, L. Zhang, Y. Zhou, Toward profundity and precision: reinventing knowledge retrieval capabilities guided by human cognition, Knowl.-Based Syst. 322 (2025) 113711.

[16] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, 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, 2018, pp. 6077–6086.

[17] J.-H. Kim, J. Jun, B.-T. Zhang, Bilinear attention networks, Adv. Neural Inf. Process. Syst. 31 (2018).

[18] S. Li, C. Gong, Y. Zhu, C. Luo, Y. Hong, X. Lv, Context-aware multi-level question embedding fusion for visual question answering, Inf. Fusion. 102 (2024) 102000.

[19] Y. Sun, J. Qi, Z. Zhu, K. Li, L. Zhao, L. Lv, Bias-guided margin loss for robust visual question answering, Inf. Process. Manag. 62 (2) (2025) 103988.

[20] A. Mao, F. Chen, Z. Ma, K. Lin, Overcoming language priors in visual question answering with cumulative learning strategy, Neurocomputing 608 (2024) 128419.

[21] D. Peng, Z. Li, Robust visual question answering via polarity enhancement and contrast, Neural Networks 179 (2024) 106560.

[22] P. Soltanzadeh, M.R. Feizi-Derakhshi, M. Hashemzadeh, Addressing the class-imbalance and class-overlap problems by a metaheuristic-based under-sampling approach, Pattern Recognit. 143 (2023) 109721.

[23] S. Ramakrishnan, A. Agrawal, S. Lee, Overcoming language priors in visual question answering with adversarial regularization, Adv. Neural Inf. Process. Syst. 31 (2018).

[24] T. Sheng, C. Shen, Y. Liu, Y. Ou, Z. Qu, Y. Liang, J. Wang, Modeling global distribution for federated learning with label distribution skew, Pattern Recognit. 143 (2023) 109724.

[25] Y. Li, B. Hu, F. Zhang, Y. Yu, J. Liu, Y. Chen, J. Xu, A multi-modal debiasing model with dynamical constraint for robust visual question answering, in: Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 5032–5045.

[26] Q. Ren, X. Cheng, S. Su, Multi-task learning with generative adversarial training for multi-passage machine reading comprehension, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 8705–8712.

[27] Y. Pan, Z. Li, L. Zhang, J. Tang, Distilling knowledge in causal inference for unbiased visual question answering, in: Proceedings of the 2nd ACM International Conference on Multimedia in Asia, 2021, pp. 1–7.

[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[29] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, arXiv preprint arXiv:1412.6572, 2014.

[30] R. Cadene, C. Dancette, M. Cord, D. Parikh, et al., RUBi: reducing unimodal biases for visual question answering, Adv. Neural Inf. Process. Syst. 32 (2019).

[31] Y. Niu, K. Tang, H. Zhang, Z. Lu, X.-S. Hua, J.-R. Wen, Counterfactual VQA: a cause-effect look at language bias, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12700–12710.

[32] A. Vosoughi, S. Deng, S. Zhang, Y. Tian, C. Xu, J. Luo, Cross modality bias in visual question answering: a causal view with possible worlds VQA, IEEE Trans. Multimedia 26 (2024) 8609–8624.

[33] C. Clark, M. Yatskar, L. Zettlemoyer, Don't take the easy way out: ensemble based methods for avoiding known dataset biases, in: EMNLP/IJCNLP (1), Association for Computational Linguistics, 2019, pp. 4067–4080.

[34] J. Wu, R. Mooney, Self-critical reasoning for robust visual question answering, Adv. Neural Inf. Process. Syst. 32 (2019).

[35] X. Han, S. Wang, C. Su, Q. Huang, Q. Tian, General greedy de-bias learning, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2023) 9789–9805.

[36] L. Zhao, K. Li, J. Qi, Y. Sun, Z. Zhu, Robust visual question answering utilizing bias instances and label imbalance, Knowl.-Based Syst. 305 (2024) 112629.

[37] J. Gu, X. Zhuang, Z. Li, MSB-VQA: overcoming multiple source biases for robust visual question answering, Neural Netw. 192 (2025) 107908.

[38] H. Tan, M. Bansal, LXMERT: learning cross-modality encoder representations from transformers, arXiv preprint arXiv:1908.07490, 2019.

[39] H. Liu, C. Li, Q. Wu, Y.J. Lee, Visual instruction tuning, Adv. Neural Inf. Process. Syst. 36 (2023) 34892–34916.

[40] D. Teney, E. Abbasnejad, K. Kafle, R. Shrestha, C. Kanan, A. Van Den Hengel, On the value of out-of-distribution testing: an example of goodhart's law, Adv. Neural Inf. Process. Syst. 33 (2020) 407–417.

[41] T. Gokhale, P. Banerjee, C. Baral, Y. Yang, MUTANT: a training paradigm for out-of-distribution generalization in visual question answering, in: EMNLP (1), Association for Computational Linguistics, 2020, pp. 878–892.

[42] E. Abbasnejad, D. Teney, A. Parvaneh, J. Shi, A. Hengel van den, Counterfactual vision and language learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10044–10054.

[43] L. Chen, Y. Zheng, J. Xiao, Rethinking data augmentation for robust visual question answering, in: European Conference on Computer Vision, Springer, 2022, pp. 95–112.

[44] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805, 2018.

[45] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., LoRA: low-rank adaptation of large language models, ICLR 1 (2) (2022) 3.