Debiasing Visual Question Answering via Ensemble Gradient Detection and Iterative Attention Forgetting

Qiuying Han[1], Shaohui Zhang[2,3,*], Peng Wang[2,3], Boyuan Li[3] and Qiwen Lu[4]
[1]School of Computer Science and Technology, Zhoukou Normal University, Zhoukou 466001, China
[2]School of Artificial Intelligence, Zhoukou Normal University, Zhoukou 466001, China
[3]School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou 450001, China
[4]School of Computer and Information Engineering, Henan University, Zhengzhou 450046, China
E-mail: {hanqy, zhangsh, wangp}@zknu.edu.cn, ieboyuan@163.com, easonglll@henu.edu.cn
*Corresponding Author

Keywords: Visual question answering, language bias, multi-modal reasoning, attention mechanism, forgotten attention algorithm, ensemble models, gradient debiasing approach

Received: April 25, 2025

In Visual Question Answering (VQA), the model's ability to understand and reason across different modalities—language, visual, and multimodal—is crucial for accurate predictions. However, recent studies have identified a significant challenge of language bias in VQA models, where the model's reasoning often hinges on incorrect linguistic associations rather than genuine multimodal understanding. To address this issue, We propose a novel Ensemble Bias Gradient Debiasing Approach (EBGDA) that combines bias detection with a dynamic forgetting mechanism. Our method utilizes a bias detector to identify and score biases across linguistic, visual, and multimodal data, enabling the model to focus on unbiased information for more accurate predictions. Additionally, inspired by human reasoning, we introduce the Forgotten Attention Algorithm (FAA), which iteratively "forgets" irrelevant visual content, progressively concentrating attention on the image regions most relevant to the question. This combination of bias mitigation and attention focusing enhances the model's ability to make multimodal inferences, reducing bias and improving overall performance. Extensive experiments on the VQA-CP v2, VQA v2, VQA-VS, GQA-OOD, and VQA-CE datasets demonstrate the effectiveness of our approach, showing state-of-the-art performance in mitigating biases and excelling in complex multimodal scenarios. Our approach achieves a 21.32% improvement over UpDn on VQA-CP v2, establishing a new state-of-the-art among methods without data augmentation.

Povzetek: Članek predlaga metodo za VQA, ki z detekcijo pristranosti in dinamičnim pozabljanjem odstranjuje jezikovno pristranskost ter preusmeri pozornost na relevantne vizualne regije, da izboljša večmodalno sklepanje.

1 Introduction

In recent years, Visual Question Answering (VQA) has become one of the prominent tasks in the field of deep learning [1], achieving significant accomplishments in various applications, such as intelligent service systems [2, 3]. However, recent research has found that many existing VQA methods tend to rely on false associations between questions and answers, without sufficiently extracting accurate visual information from images to answer questions. For example, when answering questions "What color?", some VQA models are inclined to use the most common answers from training data of that type, like "yellow", rather than extracting genuine color information from images. Additionally, some studies [4, 5] have indicated deficiencies in the existing methods' understanding of images, resulting in answers generated by the model relying on image regions with low relevance to the questions. In other words, specific methods often provide correct answers based on incorrect image regions, which does not genuinely reflect the model's performance in the question-answering task. Consequently, the factors affecting the robustness of VQA models can be summarized into two primary aspects: inherent biases in the language distribution of training and testing datasets, and the improper shortcut biases caused by the inadequate utilization of visual information [6]. However, biases in VQA are multifaceted, affecting not only language but also the improper use of visual information, resulting in models that often focus on irrelevant regions of the image. As shown in Fig. 1, when the object elements in the image are nearly the same, the model during training will generate the answer "Standing" based on the object and keyword ("What" and "doing"). However, during testing, for the same question, the model will generate the incorrect answer based on similar objects and keywords (the correct answer is "Jumping", but the model gives the wrong answer

Q:What are the animals doing in the picture?

A: Standing

Q:What are the animals doing in the picture?

A: Jumping

Figure 1: An example of multimodal bias: the presence of visual features related to the dog and horse, along with the keyword "what" and "doing" in the question, misleads the model into predicting an incorrect answer. While the correct action for the second image is "Jumping", the model incorrectly predicts "Standing" based on similar objects and keywords.

"Standing").

The state-of-the-art and noteworthy methods primarily revolve around data augmentation techniques and attention-based approaches. Data augmentation methods [7] aim to enhance a model's understanding of critical features within the data by expanding the dataset with samples, such as counterfactual instances and additional annotations [8, 9], which help eliminate biases and enhance robustness [10, 11] by obtaining more critical sample features and supplementary information. However, it is still of great interest and challenge to remove the language biases in VQA model without resorting to data augmentation [12]. Regarding attention-based methods [13], the majority currently integrate these into pre-trained models for efficient feature fusion [13–17], with limited emphasis on fully utilizing visual information.

We introduce an Ensemble Bias Gradient Debiasing Approach (EBGDA), which incorporates ensemble learning to dynamically identify and score different types of biases in the data. The method applies overfitting techniques to certain biases to ensure that the main model focuses on unbiased and relevant information during prediction. Our bias detector, tailored to the multimodal nature of VQA, assesses the degree of bias in both linguistic and visual domains, helping the model to adapt its learning process accordingly.

Additionally, inspired by the idea of iterative human reasoning, where we progressively focus on relevant areas of an image based on the question, we propose a novel approach called the Forgotten Attention Algorithm (FAA). This algorithm progressively "forgets" irrelevant image regions after each iteration, allowing the model to concentrate on the most relevant visual information for answering the question. In Fig. 2, it is evident that prominent objects (*i.e.*, the bench) often dominate the model's attention, causing it to overlook the finer image area that is relevant to the question (*i.e.*, the people). This observation poses a new challenge: how to focus on the right image area that is the most relevant to the question. To address this problem, we are inspired by the process of answering questions of human beings, where people always gradually reduce the focus area in the image with the aid of question information until the final related area is retained. This iterative forgetting process helps the model mitigate the impact of biased associations by focusing on the crucial regions of the image that relate to the question, improving its reasoning capacity.

To validate the effectiveness of our approach, we conducted extensive experiments on the VQA-CP v2, VQA v2, VQA-VS, GQA and VQA-VE. The results show that our method, combining FAA and EBGDA, significantly outperforms existing models, without relying on additional annotations or data augmentation. Our approach effectively mitigates language and visual biases, enhancing the robustness of VQA models across various challenging scenarios. The main contributions of this study are summarized as follows:

(1) Ensemble Bias Gradient Debiasing Approach: We propose a unified multi-bias debiasing approach which leverages ensemble learning to dynamically identify and evaluate the impact of linguistic, visual, and multimodal biases. The method selectively overfits to harmful biases, enabling the main model to focus more effectively on unbiased and relevant information during prediction.

(2) Cognitive-Inspired Attention Mechanism: We propose a novel Forgotten Attention Algorithm (FAA) inspired by human cognitive processes, which progressively refines visual focus by iteratively "forgetting" irrelevant image regions, thereby enhancing alignment between question-guided attention and critical visual content.

(3) Bias Mitigation via Deliberate Forgetting: The FAA addresses language bias in VQA by systematically suppressing spurious question-answer correlations through multi-round attention refinement, forcing the model to rely on genuine multi-modal reasoning rather than linguistic shortcuts.

(4) Iterative Attention Refinement for Precision: Unlike conventional static attention mechanisms, our FAA dynamically narrows the model's focus across successive iterations, ensuring precise localization of question-relevant visual features while discarding distracting information.

(5) Extensive Experimental Validation: On VQA-CP v2, our enhancements in leveraging visual information led to optimal performance. Notably, without additional annotations, our approach attained a 21.32% improvement compared to the UpDn baseline model.

To address these challenges, our study focuses on the following research questions: RQ1: Can ensemble gradient modeling effectively identify and suppress linguistic, visual, and multimodal biases without relying on data augmentation? RQ2: Can an iterative forgetting mechanism,

What are the people doing?

Wrong Answer: Sitting

Right Answer: Cycling

Figure 2: Due to the presence of biases, the influence of the size of prominent objects in the image on model reasoning leads to incorrect answers, while the image regions relevant to the answers often occupy a small portion. FAA achieves this by masking irrelevant regions in the image, allowing the model to focus on image details for inference.

inspired by human cognition, progressively refine attention maps to enhance multimodal alignment and reasoning accuracy?

2 Related works

2.1 Visual question answering

The Visual Question Answering (VQA) task requires models to provide accurate answers to natural language questions based on visual content. Since its introduction, significant progress has been made in dataset development and multimodal fusion techniques. Benchmark datasets such as VQA v2 [18], GQA [19], and CLEVR [20] have driven advancements by providing large-scale annotated samples with balanced linguistic distributions. For more complex reasoning, OK-VQA [21] focuses on outside-knowledge-based questions, while VideoQA [22] extends the task to dynamic video understanding.

Current state-of-the-art approaches primarily adopt single-stream or dual-stream architectures [23–27]. Single-stream models (*e.g.*, ViLBERT [24], LXMERT [25]) employ unified transformers to jointly encode image-text pairs, enabling deep cross-modal interaction. In contrast, dual-stream methods (*e.g.*, MCAN [26], UNITER [27]) process visual and linguistic features separately before fusion, often achieving higher precision through modular design. These models benefit from large-scale pretraining on datasets like Conceptual Captions [28] or Visual Genome [29], which enhances their ability to align visual and textual representations.

However, despite their strong performance, these methods still face two critical challenges:

(1) Language Bias: Models tend to exploit statistical priors in training data (*e.g.*, favoring frequent answers like "yellow" for color questions) rather than grounding responses in visual evidence [30].

(2) Weak Visual Grounding: Attention mechanisms often focus on salient but irrelevant objects (*e.g.*, a "bench" dominating attention when the question concerns "people"), leading to incorrect reasoning paths [11].

Recent work attempts to address these issues through debias strategies (e.g., adversarial learning[31], counterfactual samples [7]) and enhanced atten tion mechanisms (e.g., compositional attention [15]).Yet, most methods either rely on costly data augment tation or fail to fully leverage visual cues. Such as HINT[32], SCR[33], and GGE-DQ[34] address VQA bias from different angles. HINT aligns model attention with human annotations, but relies on external supervision. SCR introduces causal regularization to suppress language priors, while GGE-DQ uses gradient-based ensemble training with dual questions, increasing computational cost. Our work bridges this gap by introducing a human-inspired iter ative attention refinement approach, eliminating language biases without external data.

2.2 Mitigate language bias

In recent research, researchers have proposed a range of debiasing methods to address language bias concerning existing defined bias issues. These methods include adversarial-based techniques [38], regularization approaches [4, 12, 39–41], and data augmentation strategies. Our approach focuses on addressing bias issues from the perspective of the visual modality.

The pervasive issue of language bias in VQA systems has been extensively documented in recent literature [4, 12]. These biases manifest when models exploit statistical regularities in question-answer pairings while neglecting genuine visual evidence, leading to compromised generalization capabilities, particularly on out-of-distribution datasets like VQA-CP [7]. Existing methodologies for mitigating language bias can be categorized into three primary paradigms:

(1) Adversarial-based Techniques: Methods such as [38] employ gradient reversal layers or auxiliary adversarial networks to explicitly disentangle language priors from visual reasoning pathways. While effective in

Table 1: Comparison of methods on VQA-CP v2 dataset

Method	Type	External Data/ Augmentation	VQA-CP v2 Acc(%)	Notes
RUBi [35]	Adversarial	No	44.23	Learns bias via gradient reversal
LMH [36]	Ensemble + Heur.	No	52.01	Uses language prior estimation
CSS [7]	Augmentation	Yes	58.95	Generates counterfactual samples
CF-VQA [37]	Regularization	No	53.55	Uses bias-weighted contrastive loss
GGE [30]	Gradient-based	No	57.32	Trains bias expert model
D-VQA [11]	Feature-level Aug	Yes	61.91	Debiasing in feature and sample space
Ours	Gradient + FAA	**No**	**62.78**	No extra annotation, no augmentation

theory, these approaches often face optimization challenges, as noted in [42], where the adversarial training dynamics can inadvertently suppress semantically relevant linguistic patterns along with harmful biases.

(2) Regularization Approaches: Recent advances in regularization-based debiasing employ specialized loss functions to mitigate language prior overfitting while preserving model generalizability. Representative approaches include: a question-only branch with entropy maximization to discourage shortcut learning [4], attention divergence constraints to ensure visual grounding [40], and a contrastive learning framework to distinguish between vision-relevant and bias-driven features [41]. These approaches demonstrate that structured regularization can effectively reduce linguistic bias without compromising model capacity, offering computationally efficient alternatives to adversarial training or data augmentation.

(3) Data Augmentation Strategies: Techniques like [7, 11, 43] generate counterfactual samples or synthetic QA pairs to break spurious correlations. The VQA-CP dataset [7] represents a notable example of this approach through systematic answer distribution inversion. However, such methods may introduce new biases during synthetic data generation while incurring significant annotation costs[43].

2.3 Attention mechanism

In the context of Visual Question Answering (VQA), attention mechanisms are employed to integrate information from different modalities [13], allowing models to focus on the most relevant regions between images and texts. Presently, attention-based methodologies include linear attention [44], co-attention [16], detection attention [13], and relational attention [45]. Consequently, in our approach, we explore the integration of attention mechanisms into debiasing methods in VQA, strengthening the model's retrieval capabilities between images and questions. Leveraging attention mechanisms enhances the role of visual information, ultimately aiding in debiasing strategies.

2.4 Summarize

To provide a clearer comparison, we summarize representative debiasing methods in Table 1, categorized by their core technique, reliance on external data, and performance on VQA-CP v2. This table highlights the novelty and advantage of our method in balancing performance and data efficiency. To provide a clearer comparison, we summarize representative debiasing methods in Table 1, categorized by their core technique, reliance on external data, and performance on VQA-CP v2. This table highlights the novelty and advantage of our method in balancing performance and data efficiency.

3 Method

We describe the architecture and algorithmic flow of FAA. As shown in Fig. 3, the left side illustrates the primary structure of the UpDn baseline model [13], responsible for extracting visual-language features. On the right side, there are stacked $Attention\_Layers$ that iteratively mask irrelevant features and make answer predictions.

3.1 Visual information combination

On the left side of Fig. 3, we utilize the UpDn encoding layer to extract features. For a given text, the UpDn leverages a standard GRU to encode each question, generating a question vector. Regarding the provided image, UpDn uses the detected visual features as input. The visual feature set is represented as $F = \{f_1, ..f_i.., f_n\}$, where f_i denotes the feature of the i-th object in the image. In our method, we also incorporate factors such as spatial position. We re-encode all the outputs from Faster-RCNN [46] into new visual features. The visual input V is represented as Eq. (1):

$$V = Visual\_Encoder(F, S, Cls, Ari) \tag{1}$$

where $Visual\_Encoder$ represents the visual encoder responsible for re-encoding the four types of features into visual input. These four types of features are represented as visual feature vectors F, spatial features S, classification scores Cls, and attribute information Ari. During the initialization phase, this re-encoded visual data V is introduced as the visual input for the VQA process.

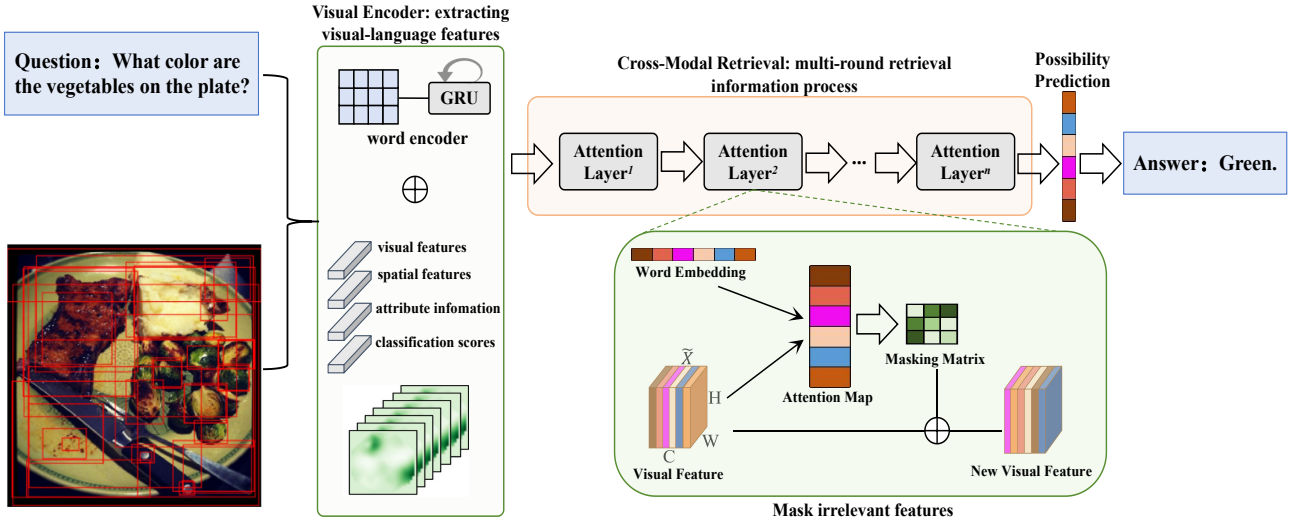

Figure 3: Our proposed FAA follows the architecture of the UpDn baseline model, comprising the feature extraction stage and the attention layers. The attention layers aim to retrieve information from the encoded question and image features, facilitating a multi-round retrieval process. In each round of retrieval, an image mask matrix is constructed to mask out the information deemed irrelevant by the model during this round, retaining crucial information for subsequent reasoning.

3.2 Bias definition

For the VQA, traditional methods typically treat it as a multi-class classification problem. The model generally takes a given triplet dataset $D = \{v_i, q_i, a_i\}_{i=1}^{N}$ as input, where the i-th image is $v_i \in \mathcal{V}$, the question is $q_i \in \mathcal{Q}$, and the answer is $a_i \in \mathcal{A}$. The main goal is to train a mapping that generates an accurate answer distribution over the answer set $\mathcal{A}$ based on the input data. In VQA-VS, the authors define nine different types of biases across three modalities, namely, linguistic, visual, and multimodal biases. In practice, we focus on the VS dataset as the primary research dataset, and for the biases corresponding to the three modalities, we follow the previous work by modeling each modality individually to capture biases within each modality. Furthermore, we treat bias in VQA as the absence of modality-specific reasoning during the inference process. Specifically, for the input sample pair (V, Q), we establish three bias models to perform bias feature modeling on the sample pair.

$$B_q = c_q(\sigma(q_i)) \tag{2}$$
$$B_v = c_v(\sigma(v_i)) \tag{3}$$
$$B_m = c_m(\sigma(m_i)) \tag{4}$$

where q_i, v_i, and m_i represent the question, image, and multimodal representation, respectively. In our implementation, the components c_q, c_v, and c_m are each implemented as two-layer MLPs with ReLU activation and Layer Normalization. These modules encode the question-only, vision-only, and joint features respectively to generate bias predictions B_q, B_v, and B_m for ensemble gradient analysis.

3.3 Bias detection

Firstly, we believe that in practical prediction tasks, a model must be able to identify different types of biases, which enables it to maintain strong generalization performance when handling diverse types of data. Additionally, in VQA, not all biases necessarily have a purely negative impact on the model. For instance, bias based on keywords or key objects may require the model to incorporate some additional real-world knowledge, which can act as supplementary information for making accurate predictions, rather than misleading the model. Therefore, in this paper, we propose a bias detector that not only assists the model in identifying biases but also estimates the weights of the detected biases to assess their influence on the model's performance.

Specifically, based on the bias feature results and definitions obtained in the previous step, we calculate the loss of the corresponding bias features to determine the type of bias the model is currently encountering.

$$Loss_q = \mathcal{L}(B_q, A) \tag{5}$$
$$Loss_v = \mathcal{L}(B_v, A) \tag{6}$$
$$Loss_m = \mathcal{L}(B_m, A) \tag{7}$$

where A represents the ground truth answers in the dataset, and $Loss_*$ denotes the loss calculated for the corresponding bias feature results. The reason for calculating the loss between the bias feature results and the ground truth labels in our method is that if the loss between the bias feature and the actual result is too large, it indicates that the predictions generated by the bias model are poor, meaning that the model has a weak ability to capture the bias. In the ensemble model approach, we aim to sufficiently train the bias model to achieve overfitting, so that the overall model develops a strong capacity for bias recognition.

In summary, in our method, we treat the modality with the largest $Loss_*$ as the one where the model exhibits bias, and we mark it accordingly.

$$F = B_* \tag{8}$$

Where F represents the bias flag. B_* represents the corresponding modality bias in $\max(Loss_q, Loss_v, Loss_m)$.

3.4 Bias score

After identifying the biases present in the sample pair, we use the bias detector to score the bias features, with the scores reflecting the degree of harm the bias imposes on the model (*i.e.*, whether the bias is beneficial or harmful). If the bias in the current sample aids the model's performance, there is no need to overfit to it during training. However, if the bias harms the model, it is necessary to ensure the model thoroughly learns the features corresponding to that bias during training. Accordingly, beneficial biases receive higher scores, while harmful biases are assigned lower scores.

$$score = 1 - softmax(m_c(F, A)) \tag{9}$$

where, F represents the bias flag obtained in the previous step, and m_c denotes the linear layer, which linearly fits the bias features corresponding to the flag with the labels. Finally, the bias score is computed through the softmax layer. In our implementation, the bias scoring function mc is a two-layer MLP with ReLU activation. Its input consists of the fused feature vector concatenated with a one-hot encoded bias type flag (length 3, indicating linguistic, visual, or multimodal bias). The output is a scalar score representing the confidence of bias activation. The MLP uses a hidden dimension of 512 with LayerNorm for stability. The bias score represents the softmax probability that the bias-specific features F support the ground truth label A. If the score is high (close to 1), it implies the bias aligns with the correct answer and is potentially helpful. If the score is low (close to 0), it indicates that the bias contradicts the correct label and is more likely to be harmful. We leverage this score to modulate the influence of gradient contributions: harmful biases (low score) are down-weighted, while benign ones (high score) are preserved. This design allows our model to adaptively suppress misleading bias without discarding all bias signals.

3.5 Fitting gradients

Based on the obtained scores, we follow the Boosting concept in our method by employing different classifiers to overfit the training data. For weak features associated with bias, we apply an overfitting approach to learn the bias distribution. By leveraging a biased model's gradient to train the base model, we aim to eliminate language bias in VQA tasks.

In training, based on the ensemble model concept, the main goal of the method is to minimize the gap between the predictions of the main model, bias models, and the true labels. Therefore, the three bias models in this method are defined as $M = \{M_v, M_q, M_m\}$, representing the visual, question, and multimodal bias models, respectively. The ultimate objective is to fit the prediction set of these three models, along with the main model, to the ground truth labels Y.

$$\min\mathcal{L}(\sigma(\sum_{i=v}^{v,q,m} M_i + f(v,q)), Y) \tag{10}$$

where, M represents the biased model, and f denotes the neural network. Our goal is to ensure that the biased parts of the dataset are fit exclusively through the biased model. In our method, we use the opposite direction of the gradient as pseudo-labels to guide the model in overfitting the biased training data. We differentiate between the effects of different types of biases on the model, and based on the bias scores obtained in the previous step, we scale the magnitude of the gradient during the gradient descent process. Simply put, higher scores correspond to beneficial biases in the training data, resulting in less decrease during backpropagation, as the model should leverage such biases to acquire knowledge. Conversely, for harmful biases, the gradient descent is amplified during backpropagation to reduce the model's reliance on such biases.

$$Y_{new} = -\nabla\mathcal{L}(f(v,q)) \tag{11}$$

$$\min\mathcal{L}(\sigma(\sum_{i=v}^{v,q,m} M_i + f(v,q)), s * Y_{new}) \tag{12}$$

where s represents the bias score. During the testing phase, we only use the main model for predictions, allowing the main model to handle the prediction of unbiased data under normal conditions. To clarify, Eq. 10 is the main objective applied to all samples, while Eq. 11 is an auxiliary loss used only for biased samples. These two losses are optimized jointly. Eq. 11 provides a targeted correction signal for harmful bias features by leveraging pseudo-gradients, scaled by bias scores. The combined optimization ensures both general supervision and bias suppression.

3.6 Attention layers

Each attention iteration in FAA corresponds to one instance of the Attention Layer stack, as implemented in Algorithm 1. Thus, the N-layer configuration of FAA represents N iterative refinement steps in attention focus. In the right side of Fig. 3, we have stacked N layers of $Attention\_Layer$ to achieve visual information masking and retrieval. The number of Attention Layers (N) in FAA equals the number of iterative forgetting steps, where each layer corresponds to one round of attention refinement (masking irrelevant regions and updating focus)". Specifically, the $Attention\_Layer$ module consists of three main components:

(1) **Initial Impression**. After obtaining visual and text features, the next step in our process is to employ

the $Self\_Attention$ mechanism. This mechanism helps identify the most critical components within each modality, similar to how humans instinctively react when first encountering an image or text. We establish the model's initial assessment of the pivotal image regions and word vectors within the provided features. As formally delineated in Algorithm 1, this process can be mathematically defined as follows:

$$\begin{aligned} att_v &= Self(V) \\ att_q &= Self(Q) \\ V^1 &= att_v * V \\ Q^1 &= att_q * Q \end{aligned} \tag{13}$$

where att_v and att_q represent the initial attention. V^1 and Q^1 represent the features obtained after fusing the initial attention with the original data.

Algorithm 1: Forgetting Attention Algorithm

Input : Representation of Object Detection Outputs:$\mathcal{F}, \mathcal{S}, Cls, Ari$; Text coded representation: Q; Number of layers of attention stack: N; Attention threshold: α.
Output: Predicted answer probability:$\mathcal{A}$.

Initialize:$V \leftarrow [\mathcal{F}, \mathcal{S}, Cls, Ari]$, $k \leftarrow 3$.
Function *FAA(V, Q)*:
  while $n \leq N$ **do**
    $att_v, att_q \leftarrow SelfAttention(V, Q)$
    $V^1, Q^1 \leftarrow att_v \odot V, att_q \odot Q$
    $V^2, Q^2 \leftarrow CrossAttention(V^1, Q^1)$
    $Att \leftarrow V^2 \odot Q^2$
    if $Att \geq \alpha$ **then**
      $V_{mask} \leftarrow 1$;
    else
      $V_{mask} \leftarrow 0$;
    $V^3 \leftarrow V_{mask} \oplus V^2$
    $V, Q \leftarrow V^3, Q^2$
  $\mathcal{A} \leftarrow V^3, Q^2$
return $\mathcal{A}$

(2) **Cross-Modal Retrieval**. With the obtained features V^1 and Q^1, we consider using the Cross Attention mechanism [14] to explore information across modalities. This step is analogous to how humans associate objects with words. We perform cross-modal information retrieval separately in the image and text domains. This is defined as Eq. (14):

$$\begin{aligned} V^2 &= CrossAtt_{v \to q}(Q^1, V^1) \\ Q^2 &= CrossAtt_{q \to v}(V^1, Q^1) \end{aligned} \tag{14}$$

where V^2 and Q^2 represent the feature outputs after conducting cross-modal retrieval for the image and text, respectively. $CrossAtt$ respectively represents the cross-modal information retrieval layer, with "image" and "question" as the primary modalities.

(3) **Masking Matrix**. After Cross-Modal Retrieval, we calculate the masking matrix for V^2 and Q^2. Initially, we employ the Top-Down attention mechanism [13] to obtain an attention weight matrix: Att, which is then compared to a predefined threshold α to determine the masking matrix. As depicted in Algorithm 1, this is defined as Eq. (15):

$$\begin{aligned} Att &= V^2 \odot Q^2 \\ V_{mask} &= Mask(Att \geq \alpha) \\ V^3 &= V_{mask} \oplus V^2 \end{aligned} \tag{15}$$

where V_{mask} represents the masking matrix, and $Mask()$ denotes the process in which Att is compared to α in Algorithm 1. The value of α is determined by the mean of attention. $\oplus$ denotes linear fusion.

In the FAA, the threshold α is a critical parameter for controlling the "forgotten regions" in the attention distribution, used to mask image regions with attention scores below α in each iteration to focus on visual features relevant to the question. The definition of α supports two strategies: (1) a fixed threshold strategy, where an empirical value is determined by analyzing the distribution of attention scores (ranging from 0 to 1); for instance, we found that α = 0.6 performs optimally in balancing masking strength and preserving key features by examining sample attention scores; (2) a dynamic threshold strategy, where α is set as a hyperparameter and adaptively adjusted based on the dispersion of attention responses in each iteration to prevent over-forgetting or under-forgetting. This adaptive adjustment dynamically updates α by calculating the standard deviation of attention scores, ensuring the model effectively focuses on relevant regions across different scenarios.

V^3 represents the features obtained by merging the masking matrix with visual features. $\oplus$ denotes the linear fusion of two types of features.

Specifically, in each attention layer, we establish a masking matrix based on the magnitude of attention weights, which masks regions in the image that contribute less to the answer. Through N such attention layers, we allow the model to progressively identify precise regions with high relevance to the given question.

3.7 Training phase

Our model is designed to effectively identify and mitigate various biases that arise in VQA tasks. The process begins with the integration of multimodal features, where visual information is extracted using object detection techniques such as Faster R-CNN. These visual features, including object location, classification scores, and spatial attributes, are then combined with textual features derived from the question to form a comprehensive representation for each input pair. A key component of this phase is the Bias Detection

mechanism, as shown in Fig. 4. Here, the model evaluates and scores biases in both the linguistic and visual domains. The linguistic biases, such as word associations, are detected through a specialized language bias model, while visual biases are identified by analyzing the model's tendency to focus on irrelevant or dominant visual features. By dynamically assessing these biases, the model adjusts its learning process to ensure that it prioritizes unbiased information, allowing it to better reason across both modalities.

Furthermore, the FAA is employed to progressively filter out irrelevant image regions during training. This iterative attention mechanism helps the model to focus on the most relevant parts of the image that contribute to answering the question, simulating the human-like reasoning of gradually reducing focus to the most pertinent visual areas. The FAA works by masking less important image regions in each iteration, ensuring that the model refines its attention and concentrates on the critical visual details.

Through this training process, which integrates bias detection, bias scoring, and the FAA, the model becomes more robust and capable of handling complex, real-world VQA scenarios without relying on biased shortcuts.

4 Experimental analysis and discussion

In this section, we will present the quantitative and qualitative analysis results from the experiments.

4.1 Experimental setup

We adopt the following unified training setup: all models are trained for 30 epochs using the Adam optimizer with a learning rate of $1e^{-4}$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The batch size is set to 256, and the cross-entropy loss is used as the objective function. All experiments are conducted on a single NVIDIA A100 32GB GPU, with an average training time of approximately 6 hours.

Dataset and evaluation metric. We validated our method on multiple VQA datasets: VQA v2, VQA-CP v2, VQA-VS, GQA, and VQA-CE. The VQA-CP v2 is employed to evaluate bias in VQA models across various scenarios, while GQA-OOD assesses the model's robustness on both ID and OOD data. VQA-CE and VQA-VS concentrate on multimodal biases, analyzing shortcut learning patterns to provide a more detailed evaluation of OOD performance. Standard evaluation on VQA v2 gauges the approach's generalizability. All test results are based on standard VQA evaluation metrics.

Baseline. We mainly used the UpDn[13] model as the baseline model.

4.2 Comparisons with SOTAs

We compared our method with other bias reduction methods, as shown in Table 2. First, we introduce the method names and their corresponding baseline models, followed by the experimental results on VQA-CP v2 and VQA v2, respectively. In Table 2, our method achieves state-of-the-art performance on VQA-CP v2 (62.78%). Moreover, compared to other methods using the same baseline models, our approach demonstrates superior performance across all three categories, particularly achieving the best results in the Num. and Other categories. On VQA v2, we also attain state-of-the-art performance (65.47%), ranking in the top three across all specific categories.

We further compared our method to approaches that utilize supplementary data. As seen, our method remains competitive against methods like CSS and D-VQA, which are trained on balanced datasets, even though such comparisons are considered unfair in [37].

As shown in Table 3, all reported results are derived from multiple independent replicate experiments to ensure reliability and reproducibility. Mean accuracy and standard deviation are computed over 5 random seeds. 95% confidence intervals are calculated using Student's t-distribution with 4 degrees of freedom. Two-sided t-tests are used to evaluate statistical significance against the UpDn baseline. $p < 0.05$ (*) indicates statistically significant improvement; $p < 0.01$ (**) indicates strong significance.

4.3 Multimodal debiasing experiments

Recent studies emphasize that VQA-CP v2 primarily addresses limited biases, particularly shortcuts between question types and answers. In contrast, VQA-VS reorganizes VQA v2 by incorporating three different forms of bias to tackle various incorrect correlations. Therefore, we conducted debiasing experiments on the VQA-VS benchmark, with the results presented in Table 4. Table 4 illustrates the performance of the EBGDA+FAA model on the VQA-VS benchmark, where it demonstrates superior performance in addressing linguistic, visual, and multimodal biases, outperforming most baseline models. Compared to the pretrained LXMERT model, EBGDA+FAA exhibits considerable robustness, effectively mitigating various biases in the VQA task.

4.4 Analysis of other metrics

In our approach, we aim to increase the role of visual content in reasoning. To assess its effectiveness, we use additional metrics. Table 5 reports the performance of FAA compared with reverse attention and linear decay. FAA achieves higher accuracy on all benchmarks, and we further introduce three auxiliary metrics to quantify its effect on bias suppression at the gradient level: CGR (Counterfactual Gradient Reduction): the average decrease in gradient magnitude for biased predictions after FAA masking. CGW

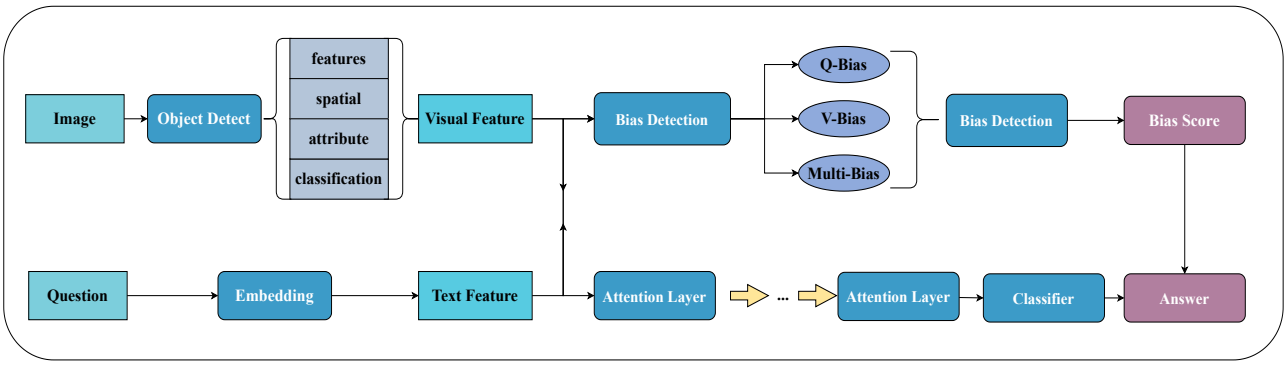

Figure 4: Training phase diagram

(Counterfactual Gradient Weighting): the normalized contribution of biased gradients in the overall weight update. CGD = CGR × CGW, reflecting the overall suppression effect on gradient-level bias flow. Compared to GGE [30], our approach performs better in terms of CGD, indicating improved utilization of visual information for answer prediction.

4.5 Results on GQA-OOD and VQA-CE

GQA-OOD and VQA-CE are new datasets and evaluation methods for VQA debiasing. Table 6 presents the experimental results of EBGDA+FAA. On GQA-OOD, EBGDA+FAA outperforms methods such as CSS, LMH, and RUBi. In VQA-CE, EBGDA+FAA achieves the highest accuracy in the Counter, validating its debiasing capability. These results strongly demonstrate the debiasing effectiveness of EBGDA+FAA.

4.6 Ablation experiments

The ablation experiment section verifies the performance of the EBGDA and FAA components in the method. EBGDA suppresses bias gradients at the feature level, while FAA refines visual focus via iterative masking; their combination enhances debiasing by addressing both feature-level and spatial-level biases.

The effectiveness of EBGDA. In this section, we conduct experiments using only the EBGDA component. The experimental results are shown in Table 7. By comparing EBGDA with methods that use the same baseline model, we can see that EBGDA also achieves performance improvements.

The effectiveness of FAA. In this section, we conduct experiments using only the FAA component. The experimental results are shown in Table 8. When comparing FAA with methods using the same baseline model, we can observe that FAA also achieves performance improvements.

As shown in Table 9, we compared the experimental results of FAA with those of other methods using attention mechanisms.

As shown in Table 10, these values confirm FAA introduces moderate overhead while significantly improving performance and focus alignment.

Forgotten Sequence. The concept of forgetting attention in this paper is based on the process of human answering relevant questions. The ablation experiment considers the order of forgetting in the algorithm to verify the validity of the concept. Table 11 shows the effect of three different sequences of attention on the performance of the model:

(1) In the method of this article, we follow the normal attention process, pay attention to the image areas that are most relevant to the problem, and forget the results obtained from the irrelevant areas.

(2) In contrast to the normal attention flow, the model first notices the areas of the image that are less relevant to the problem, assigns higher weights to them, and finally forgets the areas of the image with lower weights.

(3) The feature areas in the image are arranged in the original linear order, and the model forgets the relevant features in the same order.

By comparing the model performance of different forgetting sequences, we were able to observe that the FAA achieved excellent performance by forgetting irrelevant regions, while the other forgetting sequences resulted in decreased performance. This suggests that the human attentional forgetting mechanism on which the FAA is based works.

Attention Threshold Selection. In this method, we set thresholds to allow the model to filter image regions to determine which are forgotten and which are retained. Different threshold sizes make the model achieve different performance when forgetting the image region. If the threshold is too high, the model will forget most of the image region, resulting in the model being unable to obtain useful information from the image, while if the threshold is too low, the model will retain useless information, thus degrading the algorithm to a common attention algorithm. Therefore, this section sets up a comparative analysis of different threshold sizes to determine the appropriate parameter as the forgetting threshold. As shown in Table 12, the corresponding

Table 2: The results of VQA-CP v2 test set and VQA v2 validation set are presented in the following table. Each column illustrates the **Best** performances of each method, excluding data augmentation techniques. Our method has been compared with state-of-the-art methods on both datasets

Method \ Dataset	Base	VQA-CP v2 test				VQA v2 val			
		All	Y/N	Num.	Other	All	Y/N	Num.	Other
UpDn[13]	-	39.96	43.01	12.07	45.82	63.48	81.18	42.14	55.66
AdvReg[47]	UpDn	41.17	65.49	15.48	35.48	62.75	79.84	42.35	55.16
RUBi[35]	UpDn	44.23	67.05	17.48	39.61	-	-	-	-
LMH[36]	UpDn	52.01	72.58	31.12	46.97	56.35	65.06	37.63	54.69
AdaVQA[48]	UpDn	54.67	72.47	53.81	45.58	-	-	-	-
CF-VQA[37]	UpDn	53.55	**91.15**	13.03	44.97	63.54	82.51	43.96	54.30
CIKD[49]	UpDn	54.05	90.01	15.10	45.88	61.29	76.34	40.20	55.43
GGE[30]	UpDn	57.32	87.04	27.75	49.59	59.11	73.27	39.99	54.39
D-VQA[11]	UpDn	61.91	88.93	52.32	50.39	64.96	82.18	44.05	57.54
GGD[34]	UpDn	59.37	88.23	38.11	49.82	62.15	79.25	42.43	54.66
GenB[40]	UpDn	59.15	88.03	40.05	49.25	62.74	**86.18**	43.85	47.03
Template-based[50]	UpDn	39.75	43.03	14.98	44.83	63.83	81.61	41.98	56.10
ASL[51]	UpDn	46.00	58.24	29.49	44.33	-	-	-	-
CVL[39]	UpDn	42.12	45.72	12.45	48.34	-	-	-	-
GradSup[52]	UpDn	46.80	64.50	15.30	45.90	-	-	-	-
RangImg[53]	UpDn	55.37	83.89	41.60	44.20	57.24	76.53	33.87	48.57
CSS[7]	UpDn	58.95	84.37	49.42	48.24	59.91	73.25	39.77	55.11
Mutant[9]	UpDn	61.72	88.90	49.68	50.78	62.56	82.07	42.52	53.28
Unshuffling[54]	UpDn	43.47	47.82	14.35	49.18	43.47	81.99	43.07	55.21
SimpleAug[55]	UpDn	52.65	66.40	43.43	47.98	64.34	81.97	43.91	56.35
ECD[56]	LMH	59.92	83.23	52.29	49.71	57.38	69.06	35.74	54.25
KDDAug[57]	UpDn	60.24	86.13	**55.08**	48.08	62.86	80.55	41.05	55.18
AttReg[58]	UpDn	46.75	66.23	11.94	46.09	64.13	81.72	43.77	56.13
SSL[59]	UpDn	57.59	86.53	29.87	50.03	-	-	-	-
CSS+CL[60]	UpDn	40.49	42.90	12.44	46.93	-	-	-	-
MMBS[61]	UpDn	48.19	65.00	14.05	48.75	63.84	79.61	44.23	57.05
LSP[6]	UpDn	61.95	89.50	52.44	50.12	65.26	82.38	44.77	57.67
DM[62]	UpDn	61.13	88.13	45.98	51.13	63.53	81.09	39.61	56.52
GVQA[63]	UpDn	31.30	57.99	13.68	22.14	48.24	72.03	31.17	34.65
AttAlign[32]	UpDn	39.37	43.02	11.89	45.00	63.24	80.99	42.55	55.22
HINT[32]	UpDn	46.73	67.27	10.61	45.88	63.38	81.18	42.99	55.56
SCR[33]	UpDn	49.45	72.36	10.93	48.02	62.2	78.8	41.6	54.5
DLR[31]	UpDn	48.87	70.99	18.72	45.57	57.96	76.82	39.33	48.54
LPF[64]	UpDn	51.57	87.33	12.25	43.61	62.63	79.51	42.90	55.02
CCB[65]	UpDn	57.99	86.41	45.63	48.76	60.73	78.37	36.88	53.17
Loss-Rescaling[66]	UpDn	47.09	68.42	21.71	42.88	55.50	64.22	39.61	53.09
RMLVQA[67]	UpDn	60.41	89.98	45.96	48.74	59.99	76.68	37.54	53.2
LRLGC[68]	UpDn	60.91	89.95	45.13	50.03	60.81	77.65	39.25	53.71
Ours	UpDn	**62.78**	87.90	51.74	**52.79**	**65.47**	82.51	**57.84**	**58.64**

experimental results of the three threshold sizes selected by us are reported.

Considering the three different choices of attention threshold, in the original attention scheme of UpDn, the contribution degree of different image regions to the answer is realized by the assigned weight, whose value is between 0 and 1. Therefore, we choose the sizes of 0.5, 0.6 and 0.7 for experimental comparison. The final experimental results show that when the threshold size is 0.6, the experimental effect reaches the optimal performance, and the model achieves a balanced trade-off in attention precision. Additionally, we also present the approach using a dynamic threshold, where the threshold is treated as a hyperparameter and optimized during training.

To further justify the choice of dynamic threshold $\alpha = 0.6$, we present two analytical curves in Fig. 5. The left subplot shows that as α increases, the recall of relevant visual regions steadily declines, indicating that overly aggressive masking may suppress informative content. Conversely, the right subplot shows that model accuracy improves with moderate filtering, peaking around $\alpha = 0.6$, and then drops as too much information is retained. These trends suggest

Table 3: Comparison of methods across different datasets with accuracy, 95% confidence intervals, *p*-value, and significance levels

Method	Dataset	Accuracy (%)	95% CI (%)	*p*-value (vs UpDn)	Significance
EBGDA	VQA-CP v2	60.10 ± 0.26	[59.81, 60.39]	0.0004	**
	VQA v2	64.01 ± 0.23	[63.75, 64.27]	0.043	*
	VQA-VS	56.44 ± 0.21	[56.21, 56.67]	0.006	**
	GQA-OOD	52.68 ± 0.22	[52.44, 52.92]	0.012	*
	VQA-CE	58.42 ± 0.21	[58.19, 58.65]	0.0012	**
FAA	VQA-CP v2	60.18 ± 0.31	[59.83, 60.53]	0.0003	**
	VQA v2	63.87 ± 0.22	[63.63, 64.11]	0.048	*
	VQA-VS	56.52 ± 0.24	[56.26, 56.78]	0.004	**
	GQA-OOD	52.60 ± 0.20	[52.38, 52.82]	0.017	*
	VQA-CE	58.17 ± 0.19	[57.95, 58.39]	0.0024	**
EBGDA+FAA	VQA-CP v2	62.78 ± 0.28	[62.49, 63.07]	0.0001	**
	VQA v2	64.39 ± 0.25	[64.09, 64.69]	0.026	*
	VQA-VS	57.90 ± 0.20	[57.68, 58.12]	0.002	**
	GQA-OOD	53.74 ± 0.23	[53.47, 54.01]	0.006	**
	VQA-CE	59.26 ± 0.18	[59.06, 59.46]	0.0009	**

Table 4: Regarding the experimental outcomes of EBGDA+FAA on the VQA-VS dataset, we have presented the relevant experimental performance reports associated with this dataset. Each column displays the performance results of the corresponding best and second-performing models

Model	Base	VQA-VS OOD Test Sets									
		Language-based				Visual-based		multi-modality			mean
		QT	KW	KWP	QT+KW	KO	KOP	QT+KO	KW+KO	QT+KW+KO	
UpDn		32.43	45.10	56.06	55.29	33.39	41.31	46.45	54.29	56.92	46.80
+LMH	UpDn	33.36	43.97	54.76	53.23	33.72	41.39	46.15	51.14	54.97	45.85
BAN	UpDn	33.75	46.64	58.36	57.11	34.56	42.45	47.92	56.26	59.77	48.53
LXMERT	-	**36.46**	**51.95**	**64.17**	**64.22**	**37.69**	**46.40**	**53.54**	**62.46**	**67.44**	**53.70**
EBGDA+FAA	UpDn	32.45	47.83	59.27	60.59	34.75	43.98	44.47	56.69	55.6	51.33

Table 5: Experiment on the evaluation metric CGD using the FAA method on the VQA-CP v2 dataset. **Best** results are displayed in each column

Method	CGR	CGW	CGD
UpDn	44.27	40.63	3.91
RUBi	39.60	33.33	6.27
CSS	**46.70**	37.89	8.87
GGE-DQ-iter	44.35	27.91	16.44
GGE-DQ-tog	42.74	**27.47**	15.27
EBGDA+FAA	45.09	27.54	**17.56**

that $\alpha = 0.6$ achieves a well-balanced trade-off between visual focus (selectivity) and reasoning completeness (coverage), supporting the empirical results reported in Table 12.

Visual Information. All experiments in Table 13 are conducted using the UpDn baseline model with modified visual input features. The purpose is to isolate and evaluate the contribution of diverse visual inputs independent of debiasing components (FAA and EBGDA). As shown in Table 13, While object-only features offer a reasonable baseline, adding semantic attributes and classification logits consistently improves performance across all question types, especially in the "Other" category where visual grounding is more complex. The best configuration (Obj + Attr + Cls) achieves a +2.62% gain in All accuracy and +3.51% in Other compared to object-only features.

Layers of Attention. As shown in Table 14, we set up different levels of attention in the method to perform validation experiments. Specifically, when humans answer questions by focusing on different areas in the image, they may go through multiple target shifts to determine the final area, while in the simulation of a computer, this operation can be achieved by setting the number of layers of attention. In this experiment, we set up a total of five different layers, as you can see the model performs the best with two attention layers. This configuration significantly improves performance during inference compared to fewer layers. However, increasing the number of layers yields a slightly worse overall accuracy as well as increasing inference time by nearly 200 seconds. Regarding accuracy degradation, we believe that this phenomenon arises due to the fact that the model rec-

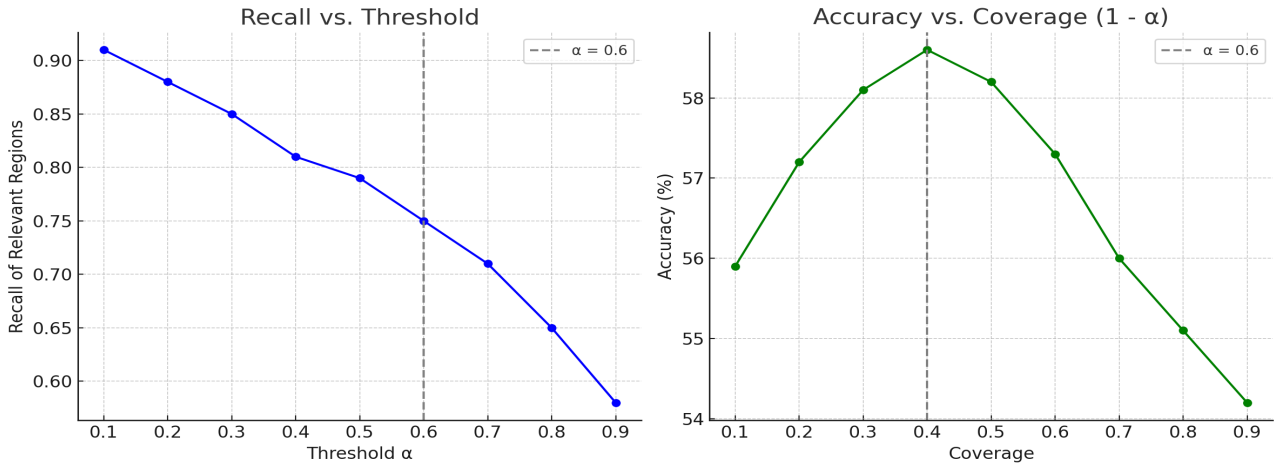

Figure 5: (Left) Recall vs. Threshold α: As the threshold α increases, the model becomes more selective, leading to a gradual decrease in the recall of question-relevant visual regions. (Right) Accuracy vs. Coverage: Coverage denotes the proportion of retained visual tokens. Model accuracy peaks around $\alpha = 0.6$, indicating a favorable balance between retaining useful information and filtering out noise.

Table 6: Experimental results on GQA-OOD and VQA-CE. Compared to the Best methods, EBGDA+FAA achieves the best performance on GQA-OOD. On VQA-CE, EBGDA+FAA also demonstrates the best performance in the Counter

Model	GQA-OOD Test			VQA-CE	
	All	Tail	Head	Counter	Easy
UpDn	46.87	42.13	49.16	33.91	**76.69**
RUBi	45.85	43.37	47.37	32.25	75.03
LMH	43.96	40.73	45.93	34.26	73.13
CSS	44.24	41.20	46.11	34.36	62.08
RandImg	-	-	-	34.41	76.21
Genb	49.43	45.63	51.76	34.80	68.15
EBGDA+FAA	**51.71**	**46.54**	**54.29**	**35.93**	63.98

ognizes incorrect visual information and masks relevant regions, thus hindering accurate answer retrieval. With three layers, the model may excessively mask relevant regions due to cumulative attention filtering, leading to information loss and accuracy degradation

Q-CSS. In our approach, we opted for a single-word replacement strategy, combined with FAA. The experimental results in Table 15 encompass a partial replication of the Q-CSS strategy from the CSS method and the QV-CSS strategy, incorporating FAA into both Q-CSS and CSS. Notably, our approach exhibits approximately 2% improvement in accuracy over Q-CSS and CSS.

Table 7: Ablation experiments with EBGDA

Method	All	Yes/No	Num.	Other	Δ
UpDn	$39.96_{\pm 0.18}$	43.01	12.07	45.82	-
SAN	$38.65_{\pm 0.20}$	40.09	12.98	44.67	-1.31
BAN	$35.94_{\pm 0.22}$	40.39	12.24	40.51	-4.02
S-MRL	$38.46_{\pm 0.25}$	42.85	12.81	43.20	-1.50
EBGDA	$60.10_{\pm 0.26}$	86.99	36.72	52.42	+20.14

Table 8: Ablation experiments with FAA

Method	All	Yes/No	Num.	Other	Δ
UpDn	$39.96_{\pm 0.18}$	43.01	12.07	45.82	-
SAN	$38.65_{\pm 0.20}$	40.09	12.98	44.67	-1.31
BAN	$35.94_{\pm 0.22}$	40.39	12.24	40.51	-4.02
S-MRL	$38.46_{\pm 0.25}$	42.85	12.81	43.20	-1.50
FAA	$60.18_{\pm 0.31}$	83.27	37.82	54.21	+20.22

Table 9: Comparison of methods on VQA-CP v2 dataset

Method	VQA-CP v2 (%)
UpDn(baseline)	39.96
SCR	49.45
HINT	46.73
FAA(Ours)	60.18

Table 10: Placeholder caption for the performance metrics comparison

Metric	UpDn	FAA (N=3)	Overhead
Training Time/ epoch	23.1 min	26.6 min	+15.2%
Inference Time/ sample	58.0 ms	65.4 ms	+12.7%
Peak GPU memory usage	14.6 GB	16.5 GB	+13.0%

Figure 6: The results of qualitative analysis show the flow of our model when making predictions by masking different image regions so that the model focuses on the effective ones

Table 11: Impact of forgetten order on performance

		VQA-CP v2 test			
Order	Base	All	Y/N	Num.	Other
FAA	UpDn	60.74	83.99	41.45	53.85
Reverse	UpDn	37.86	78.00	15.34	23.00
Linear	UpDn	27.13	71.99	6.18	9.37

Table 12: Performance corresponding to different attention thresholds

		VQA-CP v2 test			
Thresholds	Base	All	Y/N	Num.	Other
0.5	UpDn	56.84	83.29	43.87	46.54
0.6	UpDn	60.18	83.99	41.45	53.85
0.7	UpDn	59.75	82.85	35.21	54.38
dynamic	UpDn	61.65	85.37	40.68	54.98

4.7 Qualitative results

As depicted in Fig. 6, the original image, after two rounds of attentional operations, masks out irrelevant areas based on attentional weights in the Fig. 6 (1), ultimately identifying the target region relevant to the answer.

In Fig. 6, more examples are given to analyze the effect of forgotten attention on changes in image areas. For example, in the example of the Fig. 6 (2), the image of the animal is the area where the zebra is located, and there is overlap between some areas that are unrelated to the problem and the zebra, which is covered by the FAA to some extent, but most of the zebra area is still captured by the model. Similarly, in the Fig. 6 (3) and Fig. 6 (4), the areas of the sign are somewhat obscured, but the model retains key semantic features in the remaining areas, enabling accurate semantic inference.

5 Conclusion

In this paper, we proposed a novel debiasing approach for Visual Question Answering by leveraging an ensemble model framework to address and mitigate various types of biases, including linguistic, visual, and multimodal biases. Through the integration of a bias detection mechanism, we dynamically identified and evaluated the impact of biases on the model's performance. By selectively applying overfitting strategies to biased components, the proposed method allows the main model to concentrate on unbiased and informative data, thereby enhancing its generalization ability across diverse scenarios. Our experiments on VQA-CP v2, VQA v2, and VQA-VS benchmarks demonstrated state-of-the-art performance, especially in challenging scenarios involving complex multimodal correlations. This approach provides an effective solution for reducing bias in VQA tasks, leading to more robust and accurate predictions.

6 Discussion

This section provides an in-depth analysis and discussion of our proposed FAA+EBGDA approach. First, compared with the LMH method, our approach does not rely on heuristic strategies based on language prior scores. Instead, we achieve more adaptive bias modeling by introducing multimodal bias detection and gradient modulation. Unlike methods such as CSS and Mutant, which depend on external adversarial samples, our approach operates without any data augmentation or external annotations, offering stronger generalizability and deployment flexibility. Second, the Forgotten Attention Algorithm (FAA) effectively suppresses visually salient but question-irrelevant regions in the image through a multi-round attention filtering mechanism. This helps avoid "saliency traps" that may mislead the model's decision-making, thereby enhancing both vi-

Table 13: Comparison of visual features across different VQA question categories (Yes/No, number, other)

Visual Features	All (%)	Yes/No	Number	Other
Object (Obj)	57.12 ± 0.21	82.14	35.47	50.91
Attribute (Attr)	58.03 ± 0.19	82.31	35.81	52.14
Classification (Cls)	58.65 ± 0.23	83.02	36.04	52.88
Obj + Attr	58.83 ± 0.22	83.15	36.19	53.01
Obj + Cls	59.02 ± 0.25	83.28	36.41	53.38
Attr + Cls	59.16 ± 0.20	83.36	36.52	53.63
Obj + Attr + Cls	**59.74 ± 0.24**	**83.85**	**37.04**	**54.42**

Table 14: The performance under different number of attention layers

Layers	All	Yes/No	Num.	Other	Time/Epoch
1	56.95	83.77	42.99	46.72	1048s
2	60.74	83.99	41.45	53.85	1303s
3	57.67	82.66	46.44	47.66	1380s
4	58.56	80.81	50.68	47.26	1532s
5	56.21	84.48	35.66	47.03	1540s

Table 15: Ablation experiments on the combination of FAA with Q-CSS and CSS methods

Method	All	Yes/No	Num.	Other
Q-CSS	56.19	80.83	40.33	47.63
CSS	58.17	84.57	46.99	47.40
FAA+Q-CSS	58.31	80.83	48.90	49.10
FAA+CSS	60.09	88.55	53.16	47.09

sual focus and multimodal alignment. In terms of computational overhead, our method introduces approximately a 15% increase in training cost compared to the standard UpDn model. During inference, with a two-layer attention configuration, the overhead remains manageable. Detailed analysis is provided in Table 10. Finally, while our method achieves competitive performance without relying on large-scale pretraining, it still has limitations when handling questions requiring external knowledge or complex reasoning. In the future work, we can explore integrating our approach with pretrained language models to further enhance its capabilities.

Data availability statement

Data will be made available on request.

Conflicts of 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.

Author contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Qiuying Han, Peng Wang, Boyuan Li and Qiwen Lu. The first draft of the manuscript was written by Qiuying Han and Shaohui Zhang. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (62172457), Science and Technology Research Project in Henan Province of China (242102210104, 252102210032), Key Scientific Research Project in Henan University of China (25B520021).

Abbreviations

The following abbreviations are used in this paper:

VQA	Visual Question Answering
FAA	Forgotten Attention Algorithm
EBGDA	Ensemble Bias Gradient Debiasing Approach
GRU	Gated Recurrent Unit
UpDn	Bottom-Up and Top-Down Attention Model

References

[1] Drew Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. *Advances in Neural Information Processing Systems*, 32, 2019.

[2] Haonan Luo, Guosheng Lin, Yazhou Yao, Fayao Liu, Zichuan Liu, and Zhenmin Tang. Depth and video segmentation based visual attention for embodied question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6807–6819, Jun 2023.

[3] Jingyao Wang, Manas Ranjan Pradhan, and Nallappan Gunasekaran. Machine learning-based human-robot interaction in its. *Information Processing and Management*, 59(1):102750, Jan 2022.

[4] Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. Greedy gradient ensemble for robust visual question answering. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1564–1573, Los Alamitos, CA, USA, 2021. IEEE Computer Society.

[5] Yibing Liu, Yangyang Guo, Jianhua Yin, Xuemeng Song, Weifeng Liu, Liqiang Nie, and Min Zhang. Answer questions with right image regions: A visual attention regularization approach. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 18(4):1–18, 2022.

[6] Jin Liu, ChongFeng Fan, Fengyu Zhou, and Huijuan Xu. Be flexible! learn to debias by sampling and prompting for robust visual question answering. *Information Processing and Management*, 60(3):103296, 2023.

[7] Long Chen, Yuhang Zheng, Yulei Niu, Hanwang Zhang, and Jun Xiao. Counterfactual samples synthesizing and training for robust visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13218–13234, 2023.

[8] Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. Learning to contrast the counterfactual samples for robust visual question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3285–3292, Online, November 2020. Association for Computational Linguistics.

[9] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. MUTANT: A training paradigm for out-of-distribution generalization in visual question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 878–892. Association for Computational Linguistics, 2020.

[10] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9687–9695, Los Alamitos, CA, USA, 2020. IEEE Computer Society.

[11] Zhiquan Wen, Guanghui Xu, Mingkui Tan, Qingyao Wu, and Qi Wu. Debiased visual question answering from feature and sample perspectives. In *35th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 3784–3796, Red Hook, NY, USA, 2021. Curran Associates Inc.

[12] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12695–12705, 2021.

[13] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, Los Alamitos, CA, USA, 2018. IEEE Computer Society.

[14] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[15] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6274–6283, 2019.

[16] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *30th International Conference on Neural Information Processing Systems(NeurIPS)*, pages 289–297, Barcelona, Spain, 2016.

[17] Pan Lu, Hongsheng Li, Wei Zhang, Jianyong Wang, and Xiaogang Wang. Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering. In *the AAAI Conference on Artificial Intelligence*, 2022.

[18] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433, Chile, 2015. IEEE Computer Society.

[19] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702, 2019.

[20] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997, 2017.

[21] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3195–3204, 2019.

[22] Kewei Tu, Meng Meng, Mun Wai Lee, Tae Eun Choe, and Song Chun Zhu. Joint video and text parsing for

understanding events and answering queries. *IEEE MultiMedia*, 21(02):42–70, 2014.

[23] Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. End-to-end open-domain question answering with bertserini. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 72–77, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[24] Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. Multi-passage bert: A globally normalized bert model for open-domain question answering. In *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5878–5882, Hong Kong, China, 2019. Association for Computational Linguistics.

[25] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 874–880. Association for Computational Linguistics, 2021.

[26] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *56th Annual Meeting of the Association for Computational Linguistics*, pages 784–789, Melbourne, Australia, 2018. Association for Computational Linguistics.

[27] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *16th European Conference on Computer Vision*, pages 104–120, 2020.

[28] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Jan 2018.

[29] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, page 32–73, May 2017.

[30] Xinzhe Han et al. Greedy gradient ensemble for robust visual question answering. In *ICCV*, pages 1584–1593, 2021.

[31] Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. Overcoming language priors in vqa via decomposed linguistic representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11181–11188, 2020.

[32] R. R. Selvaraju et al. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2591–2600, 2019.

[33] Jialin Wu et al. Self-critical reasoning for robust visual question answering. *Advances in Neural Information Processing Systems*, 32, 2019.

[34] Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. General greedy de-bias learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[35] Remi Cadene et al. Rubi: Reducing unimodal biases for visual question answering. In *NeurIPS*, Red Hook, NY, USA, 2019. Curran Associates Inc.

[36] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082, Hong Kong, China, November 2019. Association for Computational Linguistics.

[37] Yulei Niu et al. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710, 2021.

[38] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 1548–1558, Red Hook, NY, USA, 2018. Curran Associates Inc.

[39] Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Antonvanden Hengel. Counterfactual vision and language learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10041–10051, 2020.

[40] JaeWon Cho, Dong-Jin Kim, Hyeonggon Ryu, and InSo Kweon. Generative bias for robust visual question answering. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11681–11690, 2023.

[41] Abhipsa Basu, Sravanti Addepalli, and R.Venkatesh Babu. Rmlvqa: A margin loss approach for visual question answering with language biases. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11671–11680, 2023.

[42] Gabriel Grand and Yonatan Belinkov. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. *arXiv preprint arXiv:1906.08430*, 2019.

[43] Tianyu Huai, Shuwen Yang, Junhang Zhang, Jiabao Zhao, and Liang He. Debiased visual question answering via the perspective of question types. *Pattern Recognition Letters*, 178:181–187, 2024.

[44] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21–29, Los Alamitos, CA, USA, 2016. IEEE Computer Society.

[45] Chenfei Wu, Jinlai Liu, Xiaojie Wang, and Xuan Dong. Object-difference attention: A simple relational attention for visual question answering. In *26th ACM international conference on Multimedia*, pages 519–527, 2018.

[46] Shaoqing Ren et al. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1137–1149, Jun 2017.

[47] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. *Neural Information Processing Systems*, 2018.

[48] Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Feng Ji, Ji Zhang, and Alberto Del Bimbo. Adavqa: Overcoming language priors with adapted margin cosine loss. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 708–714, Montreal, 2021. ijcai.org.

[49] Yonghua Pan, Zechao Li, Liyan Zhang, and Jinhui Tang. Distilling knowledge in causal inference for unbiased visual question answering. In *Proceedings of the 2nd ACM International Conference on Multimedia in Asia*, page 1–7, Mar 2021.

[50] Kushal Kafle, Mohammed Yousefhussien, and Christopher Kanan. Data augmentation for visual question answering. In *Proceedings of the 10th International Conference on Natural Language Generation*, Jan 2017.

[51] Damien Teney and Antonvanden Hengel. Actively seeking and learning from live data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1940–1949, Jun 2019.

[52] Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. Learning what makes a difference from counterfactual examples and gradient supervision. In *Computer Vision – ECCV 2020,Lecture Notes in Computer Science*, page 580–599, Jan 2020.

[53] Damien Teney, Kushal Kafle, Robik Shrestha, Ehsan Abbasnejad, Christopher Kanan, and Antonvanden Hengel. On the value of out-of-distribution testing: An example of goodhart's law. *arXiv: Computer Vision and Pattern Recognition*, May 2020.

[54] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. Unshuffling data for improved generalization in visual question answering. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021.

[55] Jihyung Kil, Cheng Zhang, Dong Xuan, and Wei-Lun Chao. Discovering the unknown knowns: Turning implicit knowledge in the dataset into explicit training examples for visual question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6346–6361. Association for Computational Linguistics, 2021.

[56] Camila Kolling, Martin More, Nathan Gavenski, Eduardo Pooch, Otavio Parraga, and RodrigoC. Barros. Efficient counterfactual debiasing for visual question answering. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan 2022.

[57] Long Chen, Yuhang Zheng, and Jun Xiao. Rethinking data augmentation for robust visual question answering. In *ECCV*, pages 95–112, Berlin, Heidelberg, 2022. Springer, Springer-Verlag.

[58] Yibing Liu, Yangyang Guo, Jianhua Yin, Xuemeng Song, Weifeng Liu, Liqiang Nie, and Min Zhang. Answer questions with right image regions: A visual attention regularization approach. *ACM Transactions on Multimedia Computing, Communications, and Applications*, page 1–18, Nov 2022.

[59] Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Bin Wang, and Yongdong Zhang. Overcoming language priors with self-supervised learning for visual question answering. *arXiv preprint arXiv:2012.11528*, 2020.

[60] Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. Learning to contrast the counterfactual samples

for robust visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Jan 2020.

[61] Qingyi Si, Yuanxin Liu, Fandong Meng, Zheng Lin, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. Towards robust visual question answering: Making the most of biased samples via contrastive learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Jan 2022.

[62] Yike Wu, Yu Zhao, Shiwan Zhao, Ying Zhang, Xiaojie Yuan, Guoqing Zhao, and Ning Jiang. Overcoming language priors in visual question answering via distinguishing superficially similar instances. *arXiv preprint arXiv:2209.08529*, 2022.

[63] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4971–4980, USA, 2018. Computer Vision Foundation / IEEE Computer Society.

[64] Zujie Liang, Haifeng Hu, and Jiaying Zhu. Lpf: A language-prior feedback objective function for debiased visual question answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul 2021.

[65] Chao Yang, Su Feng, Dongsheng Li, Huawei Shen, Guoqing Wang, and Bin Jiang. Learning content and context with language bias for visual question answering. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, Jul 2021.

[66] Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Qi Tian, and Min Zhang. Loss-rescaling vqa: Revisiting language prior problem from a class-imbalance view. *IEEE Transactions on Image Processing*, page 227–238, Jan 2022.

[67] Abhipsa Basu, Sravanti Addepalli, and R. Venkatesh Babu. RMLVQA: A margin loss approach for visual question answering with language biases. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 11671–11680, Canada, 2023. IEEE.

[68] Runlin Cao and Zhixin Li. Overcoming language priors for visual question answering via loss rebalancing label and global context. In Robin J. Evans and Ilya Shpitser, editors, *Uncertainty in Artificial Intelligence, UAI 2023, July 31 - 4 August 2023, Pittsburgh, PA, USA*, volume 216 of *Proceedings of Machine Learning Research*, pages 249–259, USA, 2023. PMLR.