mathematics

MDPI

Article

HQD-EM: Robust VQA Through Hierarchical Question Decomposition Bias Module and Ensemble Adaptive Angular Margin Loss

SeongHyeon Noh [1] and Jae Won Cho [2,*]

[1] Department of Computer Engineering, Sejong University, Gwangjin-gu, Seoul 05006, Republic of Korea; watercar67@sju.ac.kr
[2] Department of Artificial Intelligence & Data Science, Sejong University, Gwangjin-gu, Seoul 05006, Republic of Korea
* Correspondence: chojw@sejong.ac.kr

Abstract

Recent studies in Visual Question Answering (VQA) have revealed that models often rely heavily on language priors rather than vision–language understanding, leading to poor generalization under distribution shifts. To address this challenge, we propose HQD-EM, a unified debiasing framework that combines the Hierarchical Question Decomposition (HQD) module with an Ensemble adaptive angular Margin (EM) loss. HQD systematically decomposes questions into multi-granular representations to capture layered language biases, while EM leverages bias confidence to modulate per-sample decision margins dynamically. Our method integrates an ensemble-based method with adaptive margin learning in an end-to-end trainable architecture. Experiments on VQA benchmarks demonstrate that HQD-EM significantly outperforms prior works on VQA-CP2 and VQA-CP1.

Keywords: Visual Question Answering; vision and language; bias issues; robustness

MSC: 68T07



Academic Editor: Jianping Gou

Received: 14 October 2025
Revised: 10 November 2025
Accepted: 11 November 2025
Published: 14 November 2025

Citation: Noh, S.; Cho, J.W. HQD-EM: Robust VQA Through Hierarchical Question Decomposition Bias Module and Ensemble Adaptive Angular Margin Loss. *Mathematics* **2025**, *13*, 3656. https://doi.org/10.3390/math13223656

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Visual Question Answering (VQA) [1,2] is a multimodal task that accepts natural language questions and images as inputs, integrating complementary information from both modalities to derive correct answers. This task requires models to deeply understand the intrinsic meanings embedded in both text and images, as well as to capture the complex relationships between them. VQA models have been actively researched since their inception and have consistently improved in performance [3–8]. However, VQA methods have been shown to rely heavily on language priors and learning shortcuts [9–12], leading to poor generalizability on the out-of-distribution (OOD) dataset and the development of robust or debiasing in VQA models.

Existing debiasing approaches for VQA fall into four main categories. Grounding-based methods [13,14] use additional attention maps or human annotations to align language with visual regions, but their reliance on external supervision limits scalability, and analyses show they often underutilize true visual content and default to learned language priors. Counterfactual-based techniques [15–18] augment data via synthetic perturbations, such as question rephrasing or image inpainting, to break spurious correlations; yet, these perturbations frequently violate the original dataset's generation philosophy and

incur heavy annotation or heuristic design costs. Adaptive margin loss-based methods (e.g., RMLVQA [19], AdaVQA [20]) dynamically adjust per-sample decision margins to improve feature separation, but they typically operate under single-modality assumptions and use only basic cross-entropy for bias learning. Ensemble-based models such as RUBi, LMH, GGE, and GenB [11,12,21,22] train auxiliary bias predictors to reweight or regularize the main model; yet, they lack systematic analysis and utilization of the complex, overlapping language biases inherent in natural questions.

Despite these advances, two key limitations remain. First, most methods address language bias at a single granularity, namely, whole questions or fixed heuristic splits, without multi-level, hierarchical analysis of nested syntactic and semantic bias patterns. Moreover, these approaches tend to suffer performance degradation as the number of tokens in the question increases. Second, ensemble- and margin-based techniques have been studied largely in isolation or connected sequentially, without an integrated framework that fuses ensemble-based methods and adaptive margin learning.

To overcome these issues, we propose HQD-EM (Hierarchical Question Decomposition with Ensemble adaptive angular Margin), as shown in Figure 1. This illustrates a concrete example of how a standard VQA model falls prey to language priors, where it memorizes the most common answer and outputs an erroneous response. Whereas we show an example where our HQD-EM framework is able to overcome that bias and infer the correct answer from the image.

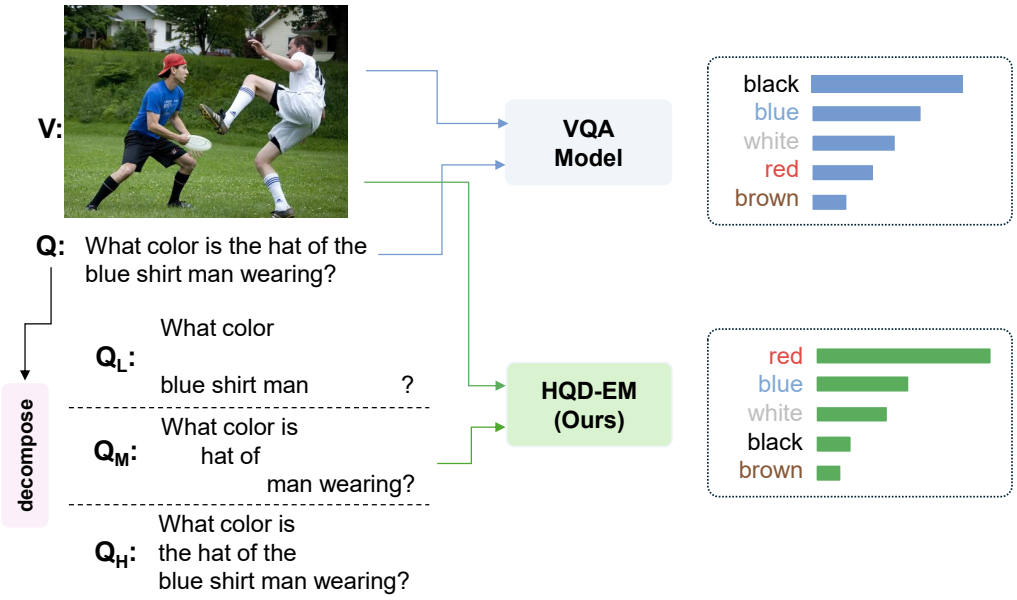


Figure 1. An example of how traditional VQA models suffer from language biases. An example where our HQD-EM framework mitigates this issue.

In detail, we present HQD-EM, a unified debiasing framework that proposes two key components: (1) Hierarchical Question Decomposition (HQD), which decomposes each question into low-, middle-, and high-level variants and ensembles their predictions to learn language priors more reliably; and (2) the Ensemble adaptive angular Margin (EM) loss, where "ensemble" explicitly refers to using the ensemble of HQD outputs as the bias signal whose instance-level confidence dynamically modulates class-specific margins in a confidence-aware manner.

Our contributions are listed as follows:

- A novel architecture combining Hierarchical Question Decomposition Bias Module (HQD) and Ensemble adaptive angular Margin (EM) loss for unified debiasing.

- A multi-granular token-based decomposition that captures and mitigates complex, overlapping language biases.
- State-of-the-art results on VQA-CP2 and VQA-CP1, demonstrating superior generalization under challenging bias conditions while maintaining competitive performance on the standard VQA v2 dataset with minimal degradation in accuracy.

2. Related Work

Visual Question Answering (VQA) [1,2,23,24] has been an active field of research in recent years, with powerful models being developed on standard benchmarks [3–8,25]. However, researchers have shown that models can be reliant on dataset biases [9,26–29], particularly language priors, which can cause models to neglect visual information and rely on spurious correlations.

To address these challenges, diverse debiasing approaches for VQA can be broadly categorized into four families: grounding-based, counterfactual-based, adaptive margin loss-based, and ensemble-based methods. Grounding-based techniques seek to reinforce visual reasoning by leveraging additional attention maps or human annotations [13,14]; however, their reliance on external supervision limits scalability, and recent analyses have shown they often fail to truly leverage visual content, defaulting instead to regularization. Counterfactual-based methods augment training data with synthetic perturbations, such as question rephrasing or image inpainting, to reduce spurious correlations [16–18]; yet, they frequently diverge from the dataset's original generation philosophy and require extensive annotation or heuristic design. Adaptive margin loss approaches (e.g., RMLVQA [19], AdaVQA [20]) assign per-sample margins based on answer frequency or difficulty, improving feature separation but typically operating within a single modality and under simplified bias assumptions. Ensemble-based models such as RUBi, LMH, GGE, and GenB [11,12,21,22] incorporate auxiliary bias predictors, typically question-only or image-only models, to reweight or regularize the main model. However, they fail to systematically analyze and address the complex, overlapping language biases embedded in natural questions. Another line of work explores adapter-based debiasing [30] to explore in- and out-of-distribution performance through adapters.

Despite these advances, two key limitations persist in prior works. First, most methods treat language bias at a single granularity, such as entire questions or fixed heuristic splits, without performing systematic, multi-level analysis of the rich, overlapping biases embedded in natural language. As a result, complex biases arising from nested syntactic or semantic patterns remain underexplored. Second, while ensemble and margin-based techniques both mitigate bias, they are typically studied in isolation or connected only sequentially, lacking an organic framework that fuses ensemble-based methods with adaptive margin learning in a unified manner.

To address these gaps, we propose HQD-EM (Hierarchical Question Decomposition with Ensemble adaptive angular Margin), the first framework to (i) hierarchically decompose each question into multi-granular variants for fine-grained bias detection and (ii) organically integrate an ensemble-based method with confidence-aware adaptive angular margin loss to achieve robust debiasing under challenging distribution shifts.

3. Proposed Method

Our HQD-EM framework is composed of two synergistic parts: the HQD module, which employs ensemble-based bias modeling to capture complex, layered language biases via hierarchical question decomposition, and the EM loss, which leverages these learned biases to adaptively adjust decision margins and perform effective debiasing (see Figure 2 for an overview). In Section 3.1, we first describe our VQA baseline. Section 3.2 presents the

HQD question decomposition strategy. Section 3.3, illustrated in Figure 2a,b, details how the decomposed questions are used to train the bias model and how to apply debiasing, and finally, Section 3.4, depicted in Figure 2c, explains how the EM loss uses the trained bias models to modulate margins for robust generalization.

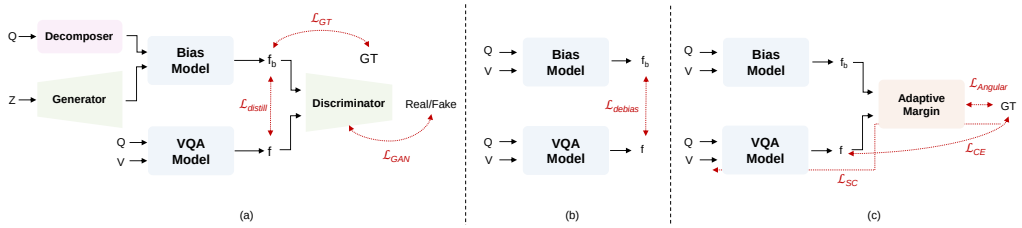


Figure 2. Overview of HQD-EM. (**a**) Bias model training using hierarchical questions and pseudo-visual features. (**b**) Debiasing with pseudo-labels. (**c**) Ensemble-aware adaptive margin loss.

3.1. VQA Baseline

We adopt the Bottom-Up Top-Down (UpDn) architecture [31] as our VQA baseline. Given an input image v, a pre-trained Faster R-CNN [32] extracts n object-level feature vectors $e_v(v) \in \mathbb{R}^{n \times d_v}$. Concurrently, the question q is encoded by GloVe [33] embeddings followed by a GRU [34] to produce $e_q(q) \in \mathbb{R}^{d_q}$. An attention-based fusion module m_f, then computes weights over the visual features conditioned on the question embedding, yielding a joint multimodal representation of x where

$$x = m_f\big(e_v(v), e_q(q)\big). \tag{1}$$

This feature x is then passed through an MLP classifier $c(\cdot)$ to produce logits y where

$$y = c(x) \in \mathbb{R}^{|\mathcal{A}|}, \tag{2}$$

which are transformed into answer probabilities $\hat{p} = \sigma(y)$. We denote this combined network of $c(m_f(\cdot,\cdot))$ as $P(\cdot,\cdot)$ for ease of notation, where necessary. We optimize this baseline using the standard cross-entropy loss against the one-hot ground-truth labels.

3.2. Hierarchical Question Decomposition Strategy

In this work, we refer to our multi-scale question decomposing strategy as Hierarchical Question Decomposition (HQD). HQD systematically splits each question into three levels—low, middle, and high—by first fixing the question token that encodes the question type and then randomly sampling the remaining tokens at different ratios. In low-level decomposition, we randomly select one-third of the question tokens (without replacement) and then form Q_L by keeping the selected tokens in their original left-to-right order (i.e., no reordering or shuffling). In middle-level decomposition, we similarly retain roughly half of the tokens in order. High-level decomposition makes no changes and retains the full token sequence. For example, the question *"How many donuts have light purple sprinkles on them?"* is transformed into a low-level variant *"How many donuts purple?"* (denoted Q_L), a middle-level variant *"How many have purple sprinkles on?"* (denoted Q_M), and a high-level variant *"How many donuts have light purple sprinkles on them?"* (denoted Q_H). We choose the minimum ratio, one-third, for Q_L based on the token statistics of VQA-CP: the maximum question length used in the model is 14 tokens, with most questions being much shorter. Further decomposition could leave too few tokens. In addition, as we deterministically keep the question-type token, this could leave only one to two additional content tokens, creating a sentence that is too sparse to preserve meaningful context. We find that using a decomposition rate of one-third typically preserves 4–5 tokens while still enforcing substantial compression. We also present a variant that fixes both the question-type token

and noun tokens (sampling over the remaining tokens to align with meaningful units), but empirically find that fixing only the question-type token yielded stronger performance and more robust debiasing; therefore, we adopt this setting (see Section 4.7). By exposing the model to these compressed and expanded variants, HQD enables learning of subtle distributed language biases at multiple granularities without sacrificing overall context.

3.3. Hierarchical Question Decomposition Training and Debiasing

As we also utilize an ensemble-based method, we create 2 models: a *target model*, which is the main model that is used for inference and testing, and a *bias model*, which is a similar model that we utilize to capture the biases and use for debiasing the target model. In our case, to capture both the dataset priors and the intrinsic biases of the target model, we adopt the same architectural structure as the target model. As previously mentioned, the target model produces logits: $y = c(x)$, where the bias model would produce the logits: $y_b = c(x)$, given a multimodal intermediary feature x. However, for the bias model, following GenB [12], we feed a Gaussian noise tensor z through a simple generator to create generated image features so the bias model learns bias representations that remain architecturally aligned with the backbone while stochastically capturing biases from the question.

$$z \sim \mathcal{N}(0, I) \ \in \ \mathbb{R}^{n \times 128} \tag{3}$$

z is then used to synthesize a pseudo-visual feature $\hat{v}$,

$$\hat{v} = G(z) \ \in \ \mathbb{R}^{n \times d_v}, \tag{4}$$

where $G : \mathbb{R}^{n \times 128} \to \mathbb{R}^{n \times d_v}$ is a small generator network, and n, d_v denote the number of region features and their dimensionality.

Our HQD bias model P_b directly receives the decomposed questions Q_L, Q_M, and Q_H, along with the pseudo-visual feature $\hat{v}$.

$$y_b \ = \ P_b\big(\hat{v}, (\ Q_L,\ Q_M,\ Q_H)\big), \tag{5}$$

Inside P_b, each question variant is first encoded by its own subnetwork:

$$\begin{aligned} h_{\text{low}} &= e_q^{\text{low}}(Q_L), \\ h_{\text{mid}} &= e_q^{\text{mid}}(Q_M), \\ h_{\text{high}} &= e_q^{\text{high}}(Q_H). \end{aligned} \tag{6}$$

We empirically found that element-wise summation of the three subnetwork outputs yields the best performance, resulting in the fused question feature h_q.

$$h_q \ = \ h_{\text{low}} \ + \ h_{\text{mid}} \ + \ h_{\text{high}}, \tag{7}$$

The fused question feature h_q and the pseudo-visual feature $\hat{v}$ are then passed through the attention-based fusion module and an MLP classifier to generate the final logits y_b.

$$x \ = \ m_f\big(e_v(\hat{v}), h_q\big). \tag{8}$$

$$y_b \ = \ \mathrm{c}(x), \tag{9}$$

Binary Cross-Entropy Loss. This loss forces P_b to overfit to dataset priors by matching its sigmoid outputs to the ground-truth answer distribution y_{gt}:

$$L_{\text{GT}} = \text{BCE}\big(\sigma(y_b),\, y_{\text{gt}}\big). \tag{10}$$

Adversarial Loss. A discriminator D [35] is trained to distinguish true VQA logits y, the output of the VQA model given image features v and question embedding q, from HQD bias logits y_b, which capture spurious correlations learned by the bias model. Meanwhile, the bias model P_b and generator G are optimized to fool D by producing logits that resemble the true VQA outputs. The resulting minimax game is:

$$\begin{aligned} \min_{P_b, G} \max_{D} \; & \mathbb{E}_{(v,q)\sim\mathcal{D}}\big[\log D\big(y\big)\big] \\ & + \mathbb{E}_{(v,q)\sim\mathcal{D}}\big[\log\big(1 - D(y_b)\big)\big]. \end{aligned} \tag{11}$$

This encourages P_b to mimic the target model's output distribution and capture its biases.

Knowledge Distillation Loss. To align P_b more closely with the target VQA model, we minimize the Kullback–Leibler divergence [36] between the target logits and bias logits:

$$L_{\text{distill}} = \; \mathbb{E}_{v,q,z}\big[D_{\text{KL}}\big(y \,\|\, y_b\big)\big]. \tag{12}$$

This distillation transfers fine-grained prediction patterns into the bias model.

Overall Loss of Training Bias. The overall objective for bias model training combines these terms:

$$L_{\text{bias}} = L_{\text{GAN}} \; + \; \lambda_1 \, L_{\text{distill}} \; + \; \lambda_2 \, L_{\text{GT}} \tag{13}$$

where λ_1 and λ_2 are hyperparameters that balance each component's contribution.

Debiasing the Target Model. After training P_b, we generate bias logits y_b by feeding the real image features v and each decomposed question into the bias model, following the design choices in GenB [12]. Let the target model's logits be y and its predicted probabilities $\hat{p} = \sigma\big(y\big)$. We then construct pseudo-labels y^{DL} that reduce overly biased predictions while preserving bias intensity:

$$y_i^{DL} \; = \; \min\Big(1,\; 2\, y_i^{gt}\; \sigma\big(-2\, y_i^{gt}\; y_{b,i}\big)\Big) \tag{14}$$

where y^{gt} is the one-hot ground truth and $y_{b,i}$ the i-th bias logit. Finally, the target model is optimized with a cross-entropy loss against these pseudo-labels:

$$L_{\text{debias}} = \; \text{BCE}\big(\hat{p},\; y^{DL}\big). \tag{15}$$

This debiasing loss leverages the bias model's confidence to steer the target model away from spurious correlations without discarding useful bias strength.

3.4. Ensemble Adaptive Angular Margin Loss

To further enhance robustness, inspired by previous work RMLVQA [19], we calibrate the loss using dataset-level statistics together with a bias signal learned via standard cross-entropy. In contrast, we tackle the limitation of such a CE (Cross Entropy) trained model, which is a largely static bias estimate by introducing an *Ensemble adaptive angular Margin (EM)* that dynamically modulates class-specific margins using both dataset priors and instance-level confidence of our HQD bias model, yielding a more effective debiasing model.

First, inspired by RMLVQA [19], for each question type q^k, we compute *frequency-based margins* to penalize models less on common answers and more on rare ones:

$$\begin{aligned} \bar{m}^k_{\text{freq}}[i] &= \frac{n^k_i + \epsilon}{\sum_{j=1}^{|A|}(n^k_j + \epsilon)}, \\ m^k_{\text{freq}}[i] &= 1 - \bar{m}^k_{\text{freq}}[i]. \end{aligned} \tag{16}$$

where n^k_i is the count of answer a_i under q^k and ϵ avoids division by zero. Here, $\bar{m}^k_{\text{freq}}[i]$ reflects how frequent an answer is, and $m^k_{\text{freq}}[i]$ becomes a larger margin for rarer answers.

Next, we inject *controlled randomness* into these margins to prevent over-correction:

$$\begin{aligned} \bar{m}^k_{\text{rand}}[i] &\sim \mathcal{N}\big(m^k_{\text{freq}}[i], \sigma\big), \\ m_{\text{rand}}[i] &= 1 - \bar{m}^k_{\text{rand}}[i]. \end{aligned} \tag{17}$$

Sampling $\bar{m}^k_{\text{rand}}$ from a Gaussian centered on m^k_{freq} with standard deviation σ smooths margin extremes and improves generalization. In parallel, we derive an *instance-level margin* from the dedicated bias classifier's output logits. We do this by training a separate classifier $y_{bm} = c_{bm}(x)$ with standard cross-entropy loss to capture language biases. Then, we normalize the cross-entropy logits y_{bm} of the stand-alone bias model.

$$\bar{m}_b = \text{softmax}\big(y_{bm}/\tau\big), \tag{18}$$

where τ is a temperature hyperparameter. Unlike previous works, here we then compute the per-instance margin as

$$m'_{\text{ins}}[i] = 1 - \phi\,\big(\bar{m}_b[i]\;\sigma(y_{b,i})\big). \tag{19}$$

Here, $(y_{b,i})$ denotes the output of HQD, and ϕ is a hyperparameter. By injecting HQD's output into the instance-specific margin $m'_{\text{ins}}[i]$, we organically couple HQD and EM. The well-trained HQD bias model provides instance-level confidence to modulate the base $\bar{m}_b[i]$, complementing the standard CE-based bias estimation and capturing residual priors that the model might miss. We propose that this synergy yields a more context-aware margin and, ultimately, better debiasing.

We then merge the *randomized* and *instance-level* margins for the ground-truth class gt:

$$\begin{aligned} m_{\text{comb}}[gt] = \beta\, m_{\text{rand}}[gt] \;+\; (1-\beta)\, m'_{\text{ins}}[gt], \\ m_{\text{comb}}[l] = m_{\text{rand}}[l], \end{aligned} \tag{20}$$

where β is annealed from 1 to 0 over training, gradually shifting emphasis from global frequency priors to per-sample bias signals.

Finally, we impose these combined margins in an *angular margin loss*, which maximizes the angular distance between feature x and classifier weight W_i:

$$L_{\text{Angular}} = -\sum_{i=1}^{|A|} a_i \log \frac{\exp\big(s\cos(\theta_i + m_{\text{comb}}[i])\big)}{\sum_{j=1}^{|A|}\exp\big(s\cos(\theta_j + m_{\text{comb}}[j])\big)}, \tag{21}$$

where s scales features and θ_i is the angle to W_i.

To train the bias-capturing classifier $c_{bm}(\cdot)$ (which outputs logits y_{bm}), we include its cross-entropy loss:

$$L_{\text{CE}} = -\sum_{i=1}^{|A|} a_i \log p_i. \tag{22}$$

To further enhance feature discrimination of our model, we add an additional Supervised Contrastive (SC) loss, L_{SC}, as contrastive losses [37,38] have been shown to aid models in learning more representative features. Specifically, given a mini-batch of size B with feature vectors $\{x_j\}_{j\in B}$ and corresponding labels $\{a_j\}_{j\in B}$, we define for each anchor j:

$$P_j = \{\, i \in B \mid a_i = a_j \},\quad N_j = \{\, i \in B \mid a_i \neq a_j \}.$$

The SC loss [37] is then:

$$L_{SC} = \sum_{j\in B} -\frac{1}{|P_j|} \sum_{p\in P_j} \log \frac{\exp\left(x_j^\top x_p/\tau\right)}{\sum_{n\in N_j} \exp\left(x_j^\top x_n/\tau\right)}, \tag{23}$$

where we set the temperature $\tau = 1$ for all experiments.

The full training objective is

$$L_{EM} \;=\; L_{CE} \;+\; L_{SC} \;+\; L_{\text{Angular.}} \tag{24}$$

This design leverages HQD's bias signals to tailor margins at both global and instance levels, yielding fine-grained debiasing and robust generalization under distribution shifts.

4. Experiments

4.1. Setup

Datasets and evaluation protocol. We evaluate our method on two widely used benchmarks for assessing VQA model robustness to bias: VQA-CP2, VQA-CP1 [10]. VQA-CP2 and VQA-CP1 [10] are specifically designed to test overreliance on language priors by altering answer distributions between train and test splits. We additionally also show our experiments on VQA v2 [1] for reference. VQA v2 serves as an in-distribution benchmark to evaluate overall performance and performance degradation when debiasing is applied. We use the standard VQA accuracy metric for all evaluations.

Experimental Settings. All experiments were conducted on a single NVIDIA (NVIDIA, Santa Clara, USA) RTX 5000 Ada GPU. The end-to-end training time was approximately 10 h, and the peak GPU memory usage at roughly 8 GB. For reference, our baseline configuration was trained for about 8.5 h and with roughly 6 GB of memory usage. At inference time, as we only utilize the target model, evaluating 2048 samples took 0.98 s for the baseline and 1.07 s for our model. At training time, the parameter count for UpDn is 24.4 M, GenB is 55.5 M, and Ours is 67.0 M. At test time, the parameter count is for UpDn and GenB is 24.4 M, while Ours is 27.6 M. Thus, with only a modest overhead time, memory, and parameters relative to the baseline, our full model yields a notable improvement in accuracy.

4.2. Results on VQA-CP2 and VQA-CP1

As shown in Table 1, we evaluate HQD-EM by integrating it with the UpDn backbone. For fair comparison, all methods, excluding the base models, use UpDn as their backbone. When applied on VQA-CP2, HQD-EM establishes a new state-of-the-art in overall accuracy (62.53%) among all prior UpDn-based methods, including those employing data augmentation or balancing, and outperforms them by a notable margin. In particular, our model exhibits strong performance on both of the challenging tasks, Num (55.67%) and Other (50.44%), demonstrating its effectiveness.

Table 1. Comparison of VQA-CP2 and VQA-CP1 performance across different debiasing methods. Our model, HQD-EM, achieves state-of-the-art overall accuracy on both VQA-CP2 and VQA-CP1. **Bold and underline** indicates best scores, **bold** indicates second best scores.

Method	VQA-CP2 Test				VQA-CP1 Test			
	All	Yes/No	Num	Other	All	Yes/No	Num	Other
SAN [39]	24.96	38.35	11.14	21.74	32.50	36.86	12.47	36.22
GVQA [10]	31.30	57.99	13.68	22.14	39.23	64.72	11.87	24.86
S-MRL [21]	38.46	42.85	12.81	43.20	36.38	42.72	12.59	40.35
UpDn [31]	39.94	42.46	11.93	45.09	36.38	42.72	42.14	40.35
Methods that include additional annotation								
HINT [13]	46.73	67.27	10.61	45.88	–	–	–	–
SCR [14]	49.45	72.36	10.93	48.02	–	–	–	–
Methods that modify the language encoder								
DLR [40]	48.87	70.99	18.72	45.57	–	–	–	–
VGQE [41]	48.75	–	–	–	–	–	–	–
Methods based on data augmentation								
CVL [42]	42.12	45.72	12.45	48.34	–	–	–	–
RandImg [43]	55.37	83.89	41.60	44.20	–	–	–	–
CSS [15]	58.95	84.37	49.42	48.21	60.95	85.60	40.57	**47.03**
MUTANT [44]	61.72	88.90	49.68	**50.78**	–	–	–	–
SSL [45]	57.59	86.53	29.87	50.03	–	–	–	–
D-VQA [16]	61.91	88.93	52.32	50.39	–	–	–	–
KDDAug [46]	60.24	86.13	**55.08**	48.08	–	–	–	–
DQF [47]	60.47	89.76	42.49	50.10	60.70	84.44	37.14	**46.57**
Methods for reducing language bias in various ways								
AReg [48]	41.17	65.49	15.48	35.48	43.43	74.16	12.44	25.32
RUBi [21]	44.23	67.05	17.48	39.61	50.90	80.83	13.84	36.02
LMH [22]	52.45	69.81	44.46	45.54	55.27	76.47	26.66	45.68
CF-VQA(SUM) [49]	53.55	**91.15**	13.03	44.97	57.03	**89.02**	17.08	41.27
GGE [11]	57.32	87.04	27.75	49.59	–	–	–	–
AdaVQA [20]	54.02	70.83	49.00	46.29	–	–	–	–
GenB [12]	59.15	88.03	40.05	49.25	**62.74**	86.18	43.85	**47.03**
RMLVQA [19]	60.41	89.98	45.96	48.74	–	–	–	–
BILI [50]	59.76	88.20	41.76	49.79	62.59	85.29	**46.77**	46.37
DVM [51]	**61.24**	**90.02**	51.05	48.95	–	–	–	–
FAN-VQA [29]	60.99	89.05	45.12	**50.65**	–	–	–	–
CVIV [52]	60.08	88.85	40.77	50.30	–	–	–	–
HQD-EM (**Ours**)	**62.53**	89.18	**55.67**	50.44	**66.12**	**91.06**	**53.58**	46.45

Compared to recent methods, BILI [50], DVM [51], FAN-VQA [29], and CVIV [52], HQD-EM consistently surpasses them in overall accuracy. Notably, on the Number (NUM) subset, which is considered a difficult subset, HQD-EM yields substantial margins of 13.91%, 4.62%, 10.55%, and 14.90% over BILI, DVM, FAN-VQA, and CVIV, respectively, while matching or exceeding performance across the remaining categories. On the VQA-CP1 benchmark, HQD-EM similarly achieves a 3.38% absolute improvement over the previous best, demonstrating its consistent efficacy across datasets.

4.3. Results on VQA v2

As shown in Table 2, prior methods that achieve high accuracy on VQA-CP2 experience substantial degradation when evaluated on the standard VQA v2 split. For example, AdaVQA [20] drops by 7.04% in overall accuracy, underscoring that strong VQA-CP2 performance often comes at the cost of poor in-distribution generalization. This makes VQA v2 performance essential as well. In contrast, our HQD-EM maintains excellent in-distribution performance, with only a 0.71% gap between VQA-CP2 and VQA v2 overall accuracy, demonstrating its robustness and stability across differing answer distributions.

Table 2. Overall VQA v2 test accuracy and gap (VQA-CP2, VQA v2) for prior methods and HQD-EM (ours). Our HQD-EM achieves a minimal gap of 0.71%, indicating strong robustness across distribution shifts.

Model	Overall	Δ Gap
SAN [39]	52.41	27.45
GVQA [10]	48.24	16.94
UpDn [31]	63.48	23.74
S-MRL [21]	37.13	1.33
Methods that include additional annotation		
HINT [13]	63.38	16.65
SCR [14]	62.20	13.73
Methods that modify the language encoder		
DLR [40]	57.96	9.09
VGQE [41]	64.04	13.93
Methods based on data augmentation		
CSS [15]	59.91	0.96
RandImg [43]	57.24	1.87
D-VQA [16]	64.96	3.05
SSL [45]	63.73	6.14
Methods reducing language bias in various ways		
AdvReg [48]	62.75	21.58
LMH [22]	56.35	4.20
CF-VQA [49]	60.94	5.89
AdaVQA [20]	46.98	7.04
RMLVQA [19]	59.99	0.42
GenB [12]	57.89	1.26
DVM [51]	60.98	0.26
HQD-EM (Ours)	61.82	0.71

4.4. Effect of Losses

As shown in Table 3, we ablate individual components while keeping the others fixed. We do not remove $\mathcal{L}_{\text{debias}}$ independently, because it contains the pseudo labels predicted by the bias model trained with $\mathcal{L}_{\text{bias}}$. Ablating it by removing it would eliminate the entire HQD bias-learning path. Also, ablating both $\mathcal{L}_{\text{bias}}$ and $\mathcal{L}_{\text{EM}}$ simultaneously would strip away core learning signals and make the comparison uninformative, so we do not ablate that configuration. Within $\mathcal{L}_{\text{bias}}$, adding $\mathcal{L}_{\text{distill}}$ to the configuration without $\mathcal{L}_{\text{GAN}}$ yields only a marginal improvement of $+0.07$. In contrast, utilizing both $\mathcal{L}_{\text{distill}}$ and $\mathcal{L}_{\text{GAN}}$ leads to a substantial gain of $+1.17$, indicating that the adversarial signal and the distillation term complement each other and help the model more effectively learn the bias patterns captured by the HQD bias-learning path. On the $\mathcal{L}_{\text{EM}}$ side, enabling $\mathcal{L}_{\text{CE}}$ (with $\mathcal{L}_{\text{SC}}$ off) improves performance by $+0.36$, suggesting that the standard CE term remains beneficial when instance-wise margins are adjusted dynamically. Finally, activating $\mathcal{L}_{\text{SC}}$ further separates samples in the feature space and delivers the best overall score of 62.53%.

Table 3. Ablation on loss components. ✓ indicates the loss is used; an empty cell indicates that the loss is removed. "Overall" reports the full method metric.

$\mathcal{L}_{\text{bias}}$			$\mathcal{L}_{\text{debias}}$	$\mathcal{L}_{\text{EM}}$			Overall
$\mathcal{L}_{\text{GT}}$	$\mathcal{L}_{\text{GAN}}$	$\mathcal{L}_{\text{distill}}$		$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{SC}}$	$\mathcal{L}_{\text{Angular}}$	
✓		✓	✓	✓	✓	✓	61.43
✓			✓	✓	✓	✓	61.36
✓	✓	✓	✓	✓		✓	62.11
✓	✓	✓	✓			✓	61.75
✓	✓	✓	✓	✓	✓	✓	62.53

4.5. Effect on Instance Margin

As shown in Table 4, y_{bm} denotes the bias signal produced by a classifier trained with standard cross-entropy, whereas y_b is the bias signal predicted by our HQD bias model. These signals are used by HQD-EM to adjust the margin dynamically on a per-sample basis. Using y_b alone slightly outperforms using y_{bm} alone, suggesting that HQD captures bias more faithfully. More importantly, leveraging both signals yields the best result, 62.53%, indicating that they are complementary and that integrating HQD confidence into the margin leads to stronger, sample-aware regularization.

Table 4. Ablation on the EM loss targets. We compare using the margin-conditioned signals y_{bm} and y_m individually and jointly. ✓ indicates if the signal is used.

y_{bm}	y_m	Overall
✓		62.13
	✓	62.21
✓	✓	62.53

4.6. Effect of HQD Variants

As shown in Table 5, we examine the effect of each question decomposition level. Using all three levels together results in the highest accuracy, supporting our claim that Q_H captures global semantics and grammatical structure, Q_M provides partial yet salient information, and Q_L helps surface fine-grained biases. Using only Q_H, which corresponds to the original question without decomposition, achieves 62.01. Relying solely on Q_M or Q_L incurs small drops, likely because partial/masked variants miss global semantics. Among the two-level settings, configurations that include Q_H outperform the $Q_M + Q_L$ pair, suggesting Q_H provides a global anchor while Q_M and Q_L add complementary cues. Combining all three levels yields the best accuracy, 62.53%, indicating that HQD more fully exposes bias signals and enables stronger debiasing.

Table 5. Ablation on hierarchical question decomposition variants. Each row denotes the use of one or more decomposition levels. Combining all three levels provides the best performance, indicating complementary semantic (high), partial (mid), and detailed (low) signals. ✓ indicates the which semantics are used.

Q_H	Q_M	Q_L	Overall
✓			62.01
	✓		61.93
		✓	61.91
	✓	✓	61.85
✓		✓	61.92
✓	✓		62.23
✓	✓	✓	62.53

4.7. Effect of Fixing Tokens in HQD

We ablate the tokens to fix when forming $Q_L/Q_M/Q_H$. Motivated by the fact that the question-type token and head nouns often carry answer-relevant cues, we explicitly test which of these signals should be preserved during decomposition. We compare three variants: (i) fix *no* tokens; (ii) fix only the *question-type* token; (iii) fix the *question-type* token and nouns, then apply the prescribed retention ratios to the remaining tokens. As shown in Table 6, fixing only the question-type token gives the best overall accuracy (62.53%) versus fixing nothing (61.89%) or fixing type + nouns (62.08%). This indicates that preserving

the type cue is sufficient while keeping beneficial randomness. Guided by this result, we *intuitively* adopt the type-only setting for all main experiments.

Table 6. Effect of fixing tokens in HQD. All variants use identical retention ratios for $Q_L/Q_M/Q_H$ (1/3, 1/2, 1.0), with fractional counts rounded to the nearest integer and original token order preserved. Fixing only the question-type token yields the best overall accuracy.

Setting	Overall (%)
HQD (no fixed tokens)	61.89
HQD (type-only, ours)	62.53
HQD (type+nouns)	62.08

4.8. Effect of HQD-EM Across Token Lengths

As shown in Figure 3, both GenB and HQD-EM see their accuracy drop as questions grow longer, confirming that more tokens generally mean harder samples. Interestingly, HQD-EM outperforms the GenB baseline at every token-count bin, demonstrating that our hierarchical decomposition and ensemble margin loss together help the model maintain robust, bias-resilient reasoning even on complex, long-form questions. Moreover, this improvement is not confined to easier examples. HQD-EM exhibits consistent gains across all token lengths and difficulty levels, confirming that its superiority stems from genuinely better handling of harder queries rather than just boosting performance on simple cases.

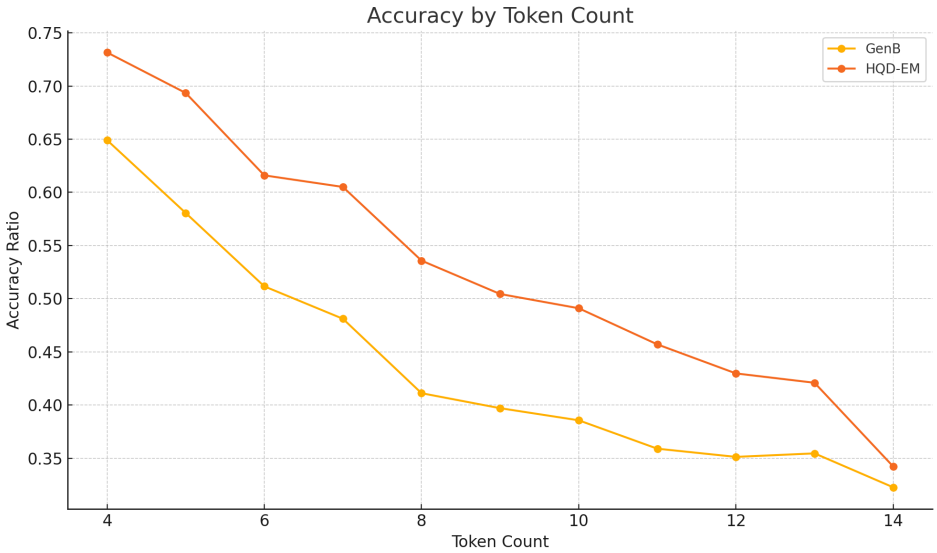


Figure 3. Accuracy ratio across different token counts for GenB and HQD-EM. HQD-EM consistently outperforms GenB, especially on longer questions.

4.9. Design Choices in HQD Module

As shown in Table 7, we systematically evaluate three fusion operations—element-wise summation ("+"), element-wise product ("×"), and a small learned MLP—applied at both the attention and question levels within the HQD module. For the MLP variant, features are first concatenated and then passed through a lightweight multi-layer perceptron that projects the fused representation back to the original feature dimensionality. The configuration using summation for both attention- and question-level fusion achieves the highest overall accuracy of 62.53%, outperforming other fusion combinations involving the learned MLP. This result underscores that simple additive fusion is particularly effective for combining visual–attention and question features, suggesting that element-wise summation plays a critical role in uncovering the optimal HQD performance.

Table 7. Impact of fusion strategies in the HQD module. +, ×, and MLP denote element-wise summation, product, and a small learned fusion network, respectively. When both attention and question level fusion use element-wise summation, the model achieves the best performance.

Attention Fusion	**Question Fusion**	Overall
×	+	62.23
+	+	62.53
+	×	62.16
×	×	62.18
+	MLP	62.09
×	MLP	61.94
MLP	MLP	62.13

4.10. Effect of Different Backbones

Through our experiments in Table 8, we run an ablation on various popular VQA architectures to show that our method is model-agnostic. In particular, we apply our method to three diverse VQA architectures, SAN [39], BAN [4], and LXMERT [7]. We include additional information about the architectures in Appendix A. We compare our methods to previous methods, GenB, DQF, and BILI, where the scores are publicly available. In every case, the backbone with HQD-EM yields a substantial increase in overall VQA-CP2 accuracy. We find that, especially in the architectures of BAN and SAN, our method outperforms the previous methods by 2.12% in SAN and 5.76% in BAN. Although BAN is a more powerful baseline than UpDn, previous methods have been unable to surpass UpDn with debiasing methods, but our method shows an improvement over UpDn in VQA-CP2. In LXMERT, our method performs favorably, and we find that the only method that outperforms Ours is DQF. DQF is a counterfactual-based method that utilizes part-of-speech tagging to generate positive and negative counterfactuals. While on the UpDn backbone, our model outperforms DQF by 2.06%, we conjecture that, as LXMERT is a much larger pretrained model, DQF may perform slightly favorably with the large number of generated counterfactuals. However, we find that in some categories, namely Yes/No, our method outperforms by 1.91%. From the overall trend of how our method works across various baseline architectures, we find that HQD-EM is architecture-agnostic and believe that HQD-EM can be a potential solution for future debiasing methods in VQA.

Table 8. Architecture ablation on the VQA-CP2 test. Δ Gap denotes the increase in overall accuracy over the base architecture. HQD-EM achieves the largest performance gain across all three backbones. † indicates our reimplementation.

Architecture	**VQA-CP2 Test**				**Δ Gap**
	All	**Yes/No**	**Num**	**Other**	
UpDn [31]	39.94	42.46	11.93	45.09	–
+ HQD-EM	62.53	89.18	55.67	50.44	+22.59
SAN [39] †	38.85	39.91	11.91	45.68	–
+ GenB [12]	56.72	88.84	19.04	50.22	+17.87
+ DQF [47]	58.12	88.92	32.58	48.99	+19.27
+ BILI [50]	56.86	88.84	22.99	49.39	+18.01
+ HQD-EM	60.24	87.54	47.30	55.29	+21.39
BAN [4] †	37.35	41.96	12.08	41.71	–
+ GenB [12]	57.37	89.11	29.52	48.37	+20.02
+ HQD-EM	63.13	90.26	55.73	50.94	+25.78
LXMERT [7]	46.23	42.84	18.91	55.51	–
+ GenB [12]	71.16	92.24	64.71	61.89	+24.93
+ DQF [47]	73.54	91.50	63.41	66.91	+27.31
+ BILI [50]	71.18	92.18	64.91	61.90	+24.95
+ HQD-EM	72.05	93.41	62.40	66.42	+25.82

4.11. Qualitative Analysis

As shown in Figure 4, we compare the attention outputs and answer predictions of the GenB baseline against those of our HQD-EM. While GenB's attention maps often scatter over irrelevant areas, leading to incorrect or unstable answers, HQD-EM repeatedly focuses on the truly informative regions of the image and arrives at the correct response. These examples highlight HQD-EM's ability to overcome spurious language biases and deliver reliable visual grounding, making it significantly more robust for real-world deployment.

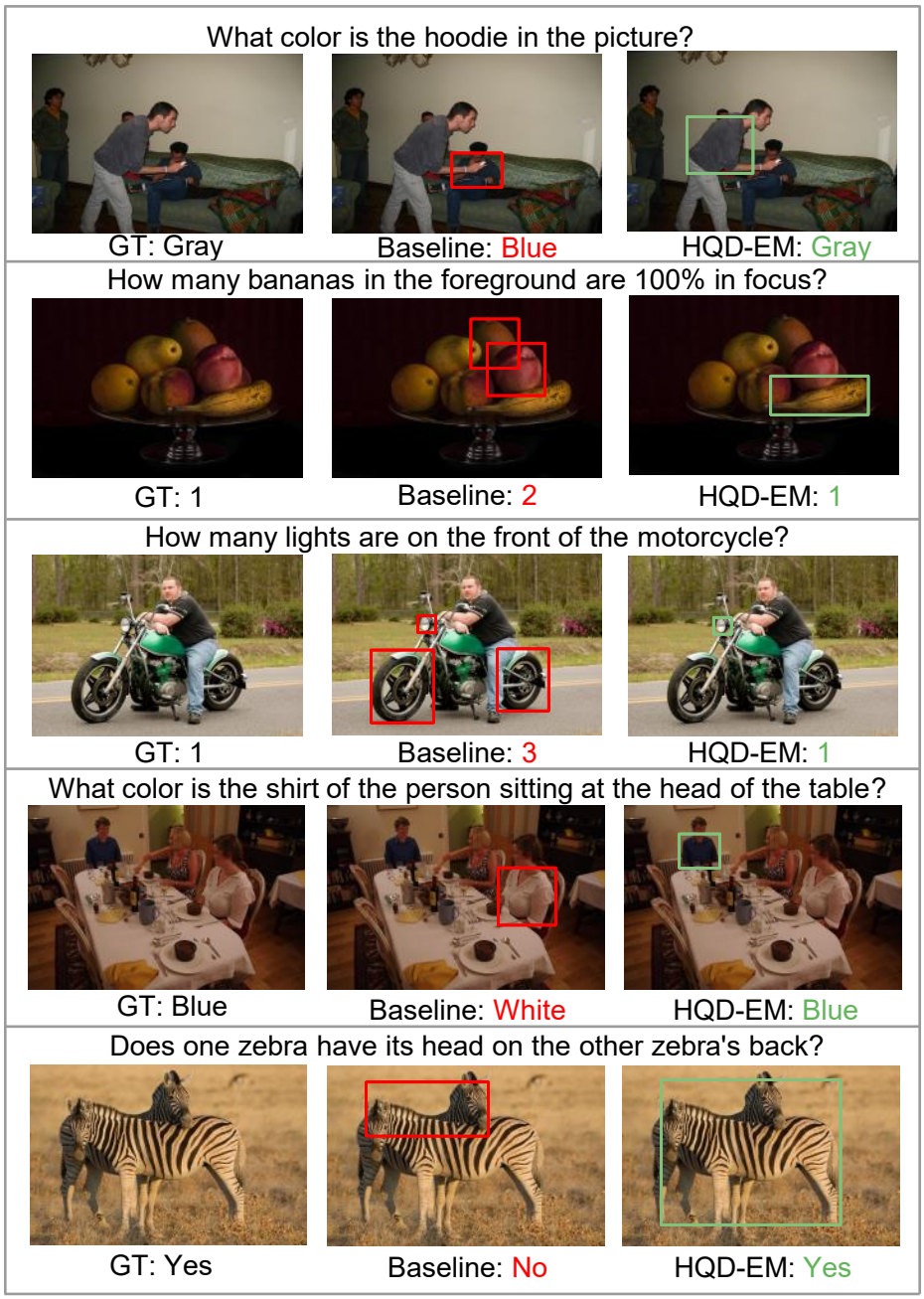


Figure 4. Qualitative comparison between the baseline GenB and HQD-EM. HQD-EM locates the correct attention regions, enabling more precise inference and increased robustness against dataset biases. The boxes indicate attention from the model. Red boxes indicate erroneously attended regions, green boxes indicate correctly attended regions.

5. Conclusions

In this work, we introduced HQD-EM, a novel framework for robust VQA that combines Hierarchical Question Decomposition and adaptive angular Margin learning. By decomposing questions into high-, mid-, and low-level variants, our HQD module captures distributed and overlapping language biases with fine granularity. Complementarily, the EM loss utilizes frequency statistics and the bias model's confidence in conjunction with the existing margin to apply dynamic, instance-specific margins, resulting in stronger generalization under biased conditions. Our extensive experiments on VQA-CP2 and VQA v2 confirm that HQD-EM not only surpasses existing methods in performance but also maintains strong accuracy in standard settings. We find that although decomposition itself can be helpful, keeping the question type is most helpful in allowing the bias model to understand biases. We also find that instead of decomposing sentences and using parts of the decomposed sentence, utilizing the high-level (the original sentence) is crucial for the bias model to capture biases. Our extensive ablation shows the importance of each component and how to utilize them best for their application while further revealing that individual question-level encoders and summation-based aggregation within HQD are crucial for maximizing bias mitigation. Overall, HQD-EM offers a unified and interpretable solution for mitigating complex multimodal biases, paving the way for more robust and trustworthy VQA systems. We believe that some limitations we have are that our method relies on sampling through rule-based means in order to make our method as light as possible. However, with existing powerful LLMs, a more dynamic method of generating and decomposing questions might be available, and we believe that may be a direction for future research.

Author Contributions: Conceptualization, J.W.C.; methodology, S.N.; software, S.N.; validation, J.W.C.; formal analysis, S.N.; investigation, S.N.; resources, J.W.C.; data curation, S.N.; writing—original draft preparation, S.N.; writing—review and editing, J.W.C.; visualization, S.N.; supervision, J.W.C.; project administration, J.W.C.; funding acquisition, J.W.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT, RS-2025-02263977, Development of Communication Platform supporting User Anonymization and Finger Spelling-Based Input Interface for Protecting the Privacy of Deaf Individuals).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in VQA: Visual Question Answering at https://visualqa.org/index.html. Accessed 1 March 2025.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

VQA	Visual Question Answering
VQA-CP	Visual Question Answering with Changing Priors
HQD-EM	Heirachical Question Decomposition with an Ensemble adaptive angular Margin

Appendix A

In this section, we include additional information about VQA architectures that we do not include in the main paper.

UpDn [31] (*Bottom-Up and Top-Down Attention*). UpDn combines object-level bottom-up visual features with top-down question-guided attention. A Faster R-CNN detector (pretrained on Visual Genome) proposes N regions; features $\{v_i\}_{i=1}^{N}$ are extracted via RoI POOLING. The question is embedded and encoded by a GRU to produce a query vector q. A single question-guided *soft* top-down attention is then applied:

$$s_i = w^\top g\big(W_v v_i + W_q q + b\big), \qquad \alpha_i = \mathrm{softmax}_i(s_i), \qquad \hat{v} = \sum_{i=1}^{N} \alpha_i v_i,$$

where $g(\cdot)$ denotes the gated-tanh nonlinearity used in the paper. VQA prediction is cast as *multi-label classification* over a fixed answer set (not a single-answer softmax):

$$\mathbf{p} = \sigma\big(W_o[\hat{v}; q] + b_o\big),$$

and is trained with a sigmoid/BCE-style loss. This separates what to detect (detector) from where to look (question-conditioned attention), yielding interpretable region attention and strong VQA performance.

SAN [39] (*Stacked Attention Networks*). SAN performs iterative evidence localization with multi-step (stacked) attention. A CNN backbone provides a spatial grid of features $\{v_i\}$; an LSTM (the original paper also considers a CNN) encodes the question into $q^{(0)}$. At stack k,

$$\alpha_i^{(k)} = \mathrm{softmax}_i\big(w^\top \tanh(W_v v_i + W_q q^{(k)})\big), \quad \hat{v}^{(k)} = \sum_i \alpha_i^{(k)} v_i, \quad q^{(k+1)} = f\big(q^{(k)}, \hat{v}^{(k)}\big),$$

where $f(\cdot)$ is a fusion function (typically with a residual connection). After 2–3 stacks, an MLP over the fused representation outputs the answer. By repeatedly updating the query with attended visual evidence, SAN refines the focus from coarse to fine regions.

BAN [4] (*Bilinear Attention Networks*). BAN models fine-grained pairwise interactions between image regions and question tokens via low-rank bilinear pooling. Let region and token feature matrices be $X \in \mathbb{R}^{d\times N}$ and $Y \in \mathbb{R}^{d\times M}$, with shared low-rank projections $U, V \in \mathbb{R}^{d\times K}$ and a glimpse-specific gating vector $p_g \in \mathbb{R}^K$. The bilinear attention for glimpse g is:

$$A_g \;=\; \mathrm{softmax}\Big((\mathbf{1}\, p_g^\top) \;\circ\; X^\top U\, V^\top Y \Big) \;\in\; \mathbb{R}^{N\times M},$$

where $\circ$ denotes the Hadamard product and $\mathbf{1}$ is an $N \times 1$ all-ones vector. The joint feature can then be aggregated (in one convenient form) as:

$$z_g \;=\; \sum_{i=1}^{N}\sum_{j=1}^{M} (A_g)_{ij}\; \big(U^\top x_i \;\odot\; V^\top y_j\big),$$

with $\odot$ being the element-wise product, x_i the i-th column of X and y_j the j-th column of Y. Outputs from several glimpses are aggregated (often with residual connections) and fed to a classifier. Compared to additive/multiplicative fusion, BAN's bilinear attention captures richer token–region alignments while remaining trainable via low-rank factorization.

LXMERT [7] (*Learning Cross-Modality Encoder Representations from Transformers*). LXMERT adopts a transformer-based two-stream *vision encoder* and *language encoder* with a *cross-modality encoder* for interaction. Region features and box/positional embeddings from a detector (e.g., Faster R-CNN) are contextualized by a self-attentive vision transformer; question tokens are contextualized by a language transformer. A cross-modality transformer then performs bi-directional co-attention to align and fuse the streams into joint representations. Pretraining uses large-scale image–text corpora with the following tasks:

1. **Masked cross-modality language modeling**
2. **Masked object prediction**: RoI *feature* regression + detected-label (category/attributes) classification
3. **Cross-modality matching**
4. **Image question answering**

For VQA fine-tuning, the fused `[CLS]` (or pooled) representation is fed to an answer classifier, leveraging general-purpose pretrained multimodal representations for robust performance.

References

1. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; Parikh, D. Vqa: Visual question answering. In Proceedings of the ICCV, Santiago, Chile, 7–13 December 2015; pp. 4971–4980.
2. Zhao, L.; Cai, D.; Zhang, J.; Sheng, L.; Xu, D.; Zheng, R.; Zhao, Y.; Wang, L.; Fan, X. Toward explainable 3d grounded visual question answering: A new benchmark and strong baseline. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *33*, 2935–2949. [CrossRef]
3. Chen, Y.C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. UNITER: UNiversal Image-TExt Representation Learning. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 104–120.
4. Kim, J.H.; Jun, J.; Zhang, B.T. Bilinear attention networks. In Proceedings of the NeurIPS, Montréal, QC, Canada, 3–8 December 2018; p. 1.
5. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *arXiv* **2019**, arXiv:1908.02265. [CrossRef]
6. Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; Dai, J. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. *arXiv* **2020**, arXiv:1908.08530. [CrossRef]
7. Tan, H.; Bansal, M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; Inui, K., Jiang, J., Ng, V., Wan, X., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 5100–5111. [CrossRef]
8. Ma, J.; Wang, P.; Kong, D.; Wang, Z.; Liu, J.; Pei, H.; Zhao, J. Robust visual question answering: Datasets, methods, and future challenges. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5575–5594. [CrossRef] [PubMed]
9. Agrawal, A.; Batra, D.; Parikh, D. Analyzing the Behavior of Visual Question Answering Models. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–4 November 2016; Su, J., Duh, K., Carreras, X., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 1955–1960. [CrossRef]
10. Agrawal, A.; Batra, D.; Parikh, D.; Kembhavi, A. Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4971–4980.
11. Han, X.; Wang, S.; Su, C.; Huang, Q.; Tian, Q. Greedy Gradient Ensemble for Robust Visual Question Answering. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 1584–1593.
12. Cho, J.W.; Kim, D.J.; Ryu, H.; Kweon, I.S. Generative Bias for Robust Visual Question Answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 11681–11690.
13. Selvaraju, R.R.; Lee, S.; Shen, Y.; Jin, H.; Ghosh, S.; Heck, L.; Batra, D.; Parikh, D. Taking a hint: Leveraging explanations to make vision and language models more grounded. In Proceedings of the ICCV, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2591–2600.
14. Wu, J.; Mooney, R.J. Self-critical reasoning for robust visual question answering. In Proceedings of the NeurIPS, Vancouver, BC, Canada, 8–14 December 2019; p. 1.
15. Chen, L.; Yan, X.; Xiao, J.; Zhang, H.; Pu, S.; Zhuang, Y. Counterfactual samples synthesizing for robust visual question answering. In Proceedings of the CVPR, Seattle, WA, USA, 13–19 June 2020; pp. 10800–10809.
16. Wen, Z.; Xu, G.; Tan, M.; Wu, Q.; Wu, Q. Debiased Visual Question Answering from Feature and Sample Perspectives. In Proceedings of the NeurIPS, Online, 6–14 December 2021; p. 1.
17. Niu, Y.; Zhang, H. Introspective distillation for robust question answering. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2021; Volume 34, pp. 16292–16304.
18. Cho, J.W.; Kim, D.J.; Jung, Y.; Kweon, I.S. Counterfactual Mix-Up for Visual Question Answering. *IEEE Access* **2023**, *11*, 95201–95212. [CrossRef]

19. Basu, A.; Addepalli, S.; Babu, R.V. RMLVQA: A Margin Loss Approach for Visual Question Answering with Language Biases. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 11671–11680.
20. Guo, Y.; Nie, L.; Cheng, Z.; Ji, F.; Zhang, J.; Bimbo, A.D. AdaVQA: Overcoming Language Priors with Adapted Margin Cosine Loss. *arXiv* **2021**, arXiv:2105.01993. [CrossRef]
21. Cadene, R.; Dancette, C.; Ben-younes, H.; Cord, M.; Parikh, D. RUBi: Reducing Unimodal Biases in Visual Question Answering. In Proceedings of the NeurIPS, Vancouver, BC, Canada, 8–14 December 2019; p. 1.
22. Clark, C.; Yatskar, M.; Zettlemoyer, L. Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. In Proceedings of the EMNLP, Hong Kong, China, 3–7 November 2019; pp. 4069–4082.
23. Johnson, J.; Hariharan, B.; Van Der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2901–2910.
24. Gurari, D.; Li, Q.; Stangl, A.J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; Bigham, J.P. VizWiz Grand Challenge: Answering Visual Questions From Blind People. In Proceedings of the CVPR, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3608–3617.
25. Kim, B.S.; Kim, J.; Lee, D.; Jang, B. Visual Question Answering: A Survey of Methods, Datasets, Evaluation, and Challenges. *ACM Comput. Surv.* **2025**, *57*, 1–35. [CrossRef]
26. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In Proceedings of the CVPR, Honolulu, HI, USA, 21–26 July 2017; pp. 6904–6913.
27. Zhang, P.; Goyal, Y.; Summers-Stay, D.; Batra, D.; Parikh, D. Yin and Yang: Balancing and Answering Binary Visual Questions. In Proceedings of the CVPR, Las Vegas, NV, USA, 27–30 June 2016; pp. 5014–5022.
28. Ouyang, N.; Huang, Q.; Li, P.; Cai, Y.; Liu, B.; Leung, H.f.; Li, Q. Suppressing Biased Samples for Robust VQA. *IEEE Trans. Multimed.* **2022**, *24*, 3405–3415. [CrossRef]
29. Bi, Y.; Jiang, H.; Hu, Y.; Sun, Y.; Yin, B. Fair Attention Network for Robust Visual Question Answering. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 7870–7881. [CrossRef]
30. Cho, J.W.; Argaw, D.M.; Oh, Y.; Kim, D.J.; Kweon, I.S. Empirical study on using adapters for debiased Visual Question Answering. *Comput. Vis. Image Underst.* **2023**, *237*, 103842. [CrossRef]
31. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the CVPR, Las Vegas, NV, USA, 27–30 June 2018; pp. 6077–6086.
32. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28.
33. Pennington, J.; Socher, R.; Manning, C. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Moschitti, A., Pang, B., Daelemans, W., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 1532–1543. [CrossRef]
34. Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), Doha, Qatar, 25–29 October 2014.
35. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems;* Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2014; Volume 27.
36. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, arXiv:1503.02531. [CrossRef]
37. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 18661–18673.
38. Liang, Z.; Jiang, W.; Hu, H.; Zhu, J. 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), Online, 16–20 November 2020.
39. Yang, Z.; He, X.; Gao, J.; Deng, L.; Smola, A. Stacked attention networks for image question answering. In Proceedings of the CVPR, Las Vegas, NV, USA, 27–30 June 2016; pp. 21–29.
40. Jing, C.; Wu, Y.; Zhang, X.; Jia, Y.; Wu, Q. Overcoming Language Priors in VQA via Decomposed Linguistic Representations. In Proceedings of the AAAI, New York, NY, USA, 7–12 February 2020; pp. 11181–11188.
41. Kv, G.; Mittal, A. Reducing language biases in visual question answering with visually-grounded question encoder. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XIII 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 18–34.

42. Abbasnejad, E.; Teney, D.; Parvaneh, A.; Shi, J.; Hengel, A.v.d. Counterfactual Vision and Language Learning. In Proceedings of the CVPR, Seattle, WA, USA, 13–19 June 2020; pp. 10041–10051.
43. Teney, D.; Kafle, K.; Shrestha, R.; Abbasnejad, E.; Kanan, C.; Hengel, A.v.d. On the Value of Out-of-Distribution Testing: An Example of Goodhart's Law. *arXiv* **2020**, arXiv:2005.09241.
44. Gokhale, T.; Banerjee, P.; Baral, C.; Yang, Y. MUTANT: A Training Paradigm for Out-of-Distribution Generalization in Visual Question Answering. In Proceedings of the EMNLP, Online, 16–20 November 2020; pp. 878–892.
45. Zhu, X.; Mao, Z.; Liu, C.; Zhang, P.; Wang, B.; Zhang, Y. Overcoming Language Priors with Self-supervised Learning for Visual Question Answering. In Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, Yokohama, Japan, 7–15 January 2021.
46. Chen, L.; Zheng, Y.; Xiao, J. Rethinking data augmentation for robust visual question answering. In Proceedings of the ECCV, Tel Aviv, Israel, 23–27 October 2022; pp. 95–112.
47. Liu, J.; Wang, G.; Fan, C.; Zhou, F.; Xu, H. Question-conditioned debiasing with focal visual context fusion for visual question answering. *Knowl.-Based Syst.* **2023**, *278*, 110879. [CrossRef]
48. Ramakrishnan, S.; Agrawal, A.; Lee, S. Overcoming language priors in visual question answering with adversarial regularization. In Proceedings of the NeurIPS, Montréal, QC, Canada, 3–8 December 2018; p. 1.
49. Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.S.; Wen, J.R. Counterfactual VQA: A Cause-Effect Look at Language Bias. In Proceedings of the CVPR, Nashville, TN, USA, 20–25 June 2021; pp. 12700–12710.
50. Zhao, L.; Li, K.; Qi, J.; Sun, Y.; Zhu, Z. Robust Visual Question Answering utilizing Bias Instances and Label Imbalance. *Knowl.-Based Syst.* **2024**, *305*, 112629. [CrossRef]
51. Peng, D.; Li, Z. Robust visual question answering via polarity enhancement and contrast. *Neural Netw.* **2024**, *179*, 106560. [CrossRef] [PubMed]
52. Pan, Y.; Liu, J.; Jin, L.; Li, Z. Unbiased Visual Question Answering by Leveraging Instrumental Variable. *IEEE Trans. Multimed.* **2024**, *26*, 6648–6662. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.