BayesVQA: Energy-Guided Bayesian Debiasing for Language-Bias-Robust Visual Question Answering

Zhiqi Huang[1], **Huanjia Zhu**[1], **Xiangwen Deng**[1], **Zhong Qinghao**[1], **Bingzhi Chen**[1*]

[1]Beijing Institute of Technology, Zhuhai
huangzhiqi@bitzh.edu.cn, chenbingzhi@bit.edu.cn

Abstract

Numerous studies have demonstrated that Visual Question Answering (VQA) models are vulnerable to language priors and dataset biases, often leading to spurious correlations between questions and answers. As a result, these models excessively rely on linguistic cues, neglecting essential visual information and causing representational distortions. To address this challenge, we propose a novel Bayesian debiasing framework termed **BayesVQA**, which integrates three carefully designed mechanisms: Energy-guided Prior Variance (EPV), Energy-guided Posterior Sampling (EPS), and Energy-guided Likelihood Reweighting (ELR). Specifically, we explicitly decompose each sample's latent representation into a biased feature and a stochastic corrective perturbation δ. Using a Bayesian formulation, we model the posterior distribution of the perturbation δ conditioned on the predictive uncertainty, quantified via calibrated energy scores. To mitigate language bias, the posterior is optimized through energy-driven variational inference with an uncertainty-adaptive prior and sampling strategy. Moreover, the ELR mechanism incorporates an energy-based weighting of the reconstruction objective and enforces an energy-coherence constraint to emphasize challenging, high-uncertainty instances and align model confidence before and after debiasing. Extensive experiments conducted across multiple standard VQA benchmarks consistently validate the superior performance of our BayesVQA method over state-of-the-art competitors under distributional shifts and challenging bias conditions.

Introduction

Visual Question Answering (VQA) tasks require models to perform joint reasoning over visual content and corresponding textual queries, producing natural-language responses for each image–question pair. Despite substantial advances (Agrawal et al. 2018; Anderson et al. 2018; Tan and Bansal 2019; Cadene et al. 2019) on standard benchmarks (Goyal et al. 2017; Hudson and Manning 2019; Agrawal et al. 2018), many VQA systems still exploit imbalances in answer-category frequencies during training (Guo et al. 2021; Basu, Addepalli, and Babu 2023; Pan et al. 2024; Vosoughi et al. 2024; Zhu et al. 2025, 2024). For example, models frequently predict the most common training-set answers when asked "*How many . . . ?*", thereby learning spurious associations between question types and responses. As shown in Fig. 1, this leads to disproportionate reliance on the linguistic modality during multimodal fusion, while critical visual cues are underutilized (Han, Wang, and Su 2021; Han et al. 2023; Zhu et al. 2025; Wang et al. 2025). This imbalance leads to pronounced semantic biases when models encounter test distributions that diverge from training data, severely compromising robustness and generalization.

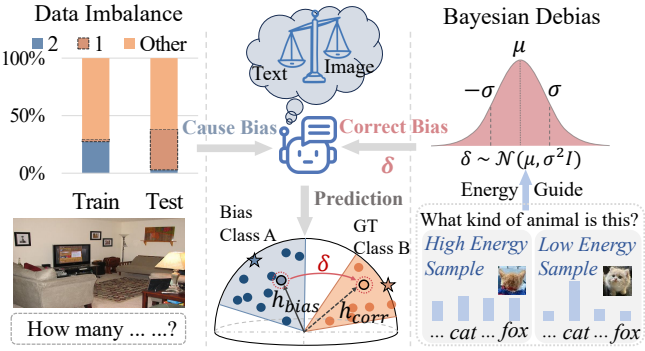


Figure 1: Imbalanced answer distributions in VQA induce language shortcut biases, where the model over-relies on question text and ignores visual evidence, leading to spurious correlations and incorrect predictions. Our energy-guided Bayesian debiasing introduces a learned corrective perturbation δ to adjust fused features and restore balanced multimodal reasoning.

Existing debiasing studies primarily attempt to mitigate the dominance of the language modality from various perspectives, aiming to enhance the contribution of the visual modality in multimodal fusion. One common class of approaches constructs explicit language bias models, introduces question-specific branches (Ramakrishnan, Agrawal, and Lee 2018; Kv and Mittal 2020; Han, Wang, and Su 2021; Han et al. 2023), or fuses multiple models to capture and counteract language priors (Zhu et al. 2025), thus reducing reliance on textual inputs. Another line of work employs data augmentation or resampling strategies, such as generating counterfactual samples by modifying questions or an-

*Corresponding author: Bingzhi Chen

swers (Chen et al. 2020; Liang et al. 2020; Niu et al. 2021), and constructing new vision-language combinations (Teney et al. 2020; Wen et al. 2021; Zhu et al. 2021) to disrupt the static correlations between question types and answers in the training data. In addition, some methods (Guo et al. 2021; Basu, Addepalli, and Babu 2023; Zhu et al. 2024) focus on the distributional structure of the feature space, using contrastive learning or cosine-margin losses to improve the discriminability of the tail categories.

Despite progress in mitigating language bias in VQA, existing approaches commonly suffer from two critical limitations. *On the one hand*, they inadequately capture and leverage the inherent sample-specific uncertainty, failing to effectively distinguish between instances with varying degrees of language-induced biases, resulting in suboptimal correction on challenging cases. *On the other hand*, their training paradigms remain overly rigid and static, neglecting the need for adaptive adjustment of corrective mechanisms based on sample difficulty and uncertainty levels, which inevitably restricts the model's robustness and generalization capability under distributional shifts. To address these limitations, our work explicitly incorporates energy-guided uncertainty modeling and Bayesian-based adaptive training dynamics into the bias-mitigation process.

In this paper, we propose a novel Bayesian debiasing framework for VQA, termed **BayesVQA**, which effectively mitigates multimodal biases, particularly in challenging out-of-distribution (OOD) scenarios. Methodologically, BayesVQA integrates three carefully designed mechanisms: **E**nergy-guided **P**rior **V**ariance (EPV), **E**nergy-guided **P**osterior Sampling (EPS), and **E**nergy-guided **L**ikelihood **R**eweighting (ELR). Specifically, our approach is grounded in a variational Bayesian inference framework augmented by an energy-based uncertainty quantification mechanism. Within this framework, predictive uncertainty, quantified through calibrated energy scores, adaptively governs three essential inference components. Firstly, EPV dynamically adjusts the prior variance of the latent corrective perturbation based on the sample's energy level, effectively regulating corrective strength in alignment with uncertainty. Secondly, EPS employs energy-scaled posterior sampling, enabling instance-specific adaptation of the posterior distribution and avoiding fixed, uniform perturbation scales. Lastly, ELR introduces an energy-weighted reconstruction loss and an energy-coherence constraint, emphasizing difficult, high-uncertainty instances while ensuring stable confidence calibration pre- and post-debiasing. By jointly considering uncertainty estimation and variational debiasing, BayesVQA achieves a balanced and adaptive integration of visual and textual modalities, preserves cross-modal semantic consistency, and substantially enhances model robustness and generalization capacity under distribution shifts. Our main contributions can be summarized as follows:

- We propose **BayesVQA**, a novel Bayesian debiasing framework explicitly designed to mitigate language-induced biases in VQA, particularly under OOD conditions. Our BayesVQA integrates uncertainty quantification into variational inference to support uncertainty-aware, adaptive balancing across modalities.
- We design three synergistic energy-guided mechanisms: (i) *EPV*, dynamically adjusting prior variance based on uncertainty; (ii) *EPS*, adaptively scaling posterior sampling per instance; and (iii) *ELR*, emphasizing high-uncertainty samples via energy-weighted reconstruction and energy-coherence constraints.
- We explicitly decompose each sample's latent representation into a biased feature and a stochastic perturbation, leveraging uncertainty-aware variational inference to effectively balance modality contributions and reduce reliance on spurious language priors.
- Extensive experiments across multiple standard VQA benchmarks verify that BayesVQA achieves state-of-the-art performance, exhibiting superior robustness and generalization under distributional shifts.

Related Work

VQA Debias Learning

A variety of specialized benchmarks have been introduced to evaluate and drive progress in debiasing VQA models. The original VQA v1 (Antol et al. 2015) and VQA v2 (Goyal et al. 2017) datasets provide in-distribution evaluation but suffer from skewed answer frequencies. To expose language priors, VQA-CP v1 and VQA-CP v2 (Agrawal et al. 2018) rearrange the train/test splits so that answer distributions are inverted between training and evaluation. VQA-CE (Dancette et al. 2021) further introduces a counterfactual evaluation protocol by constructing challenging test samples where language priors are intentionally misleading. Recently, debiasing research has diversified into several core directions: causal intervention and counterfactual reasoning to explicitly break spurious correlations (Niu et al. 2021); contrastive learning frameworks that separate unbiased from biased representations (Liang et al. 2020); adversarial regularization and ensemble methods to suppress language-only shortcuts (Ramakrishnan, Agrawal, and Lee 2018); and angular margin–based debiasing techniques that refine answer-level decision boundaries in the embedding space (Guo et al. 2021; Basu, Addepalli, and Babu 2023).

Variational Bayesian Inference

Variational Bayesian (VB) inference is a widely used method for approximate posterior estimation and uncertainty quantification in complex latent variable models (Subedar et al. 2019; Zhao et al. 2023). The Variational Autoencoder (VAE) framework employs an encoder network to learn a tractable posterior approximation over continuous latent representations, facilitating scalable inference in deep generative models. Extensions of VB to multimodal fusion jointly infer shared latent factors across heterogeneous inputs, such as audio–visual and vision–text data, yielding robustness under partial or noisy observations (Xue, Qian, and Xu 2023). Although extensive research has applied VB at the model level, such as placing priors over network weights or output distributions to improve calibration and generalization (Wilson and Izmailov 2020), these approaches rarely address bias correction at the feature fusion layer or the balancing of modality-specific gradients.

Energy Uncertainty Estimation

Energy scoring functions derived from model logits have been widely adopted in OOD detection and confidence calibration tasks, as they distinguish in-distribution from OOD inputs effectively and produce more reliable confidence estimates than raw softmax probabilities(Ming, Fan, and Li 2022). Energy-based methods further adapt uncertainty calibration at the sample-level (Zhang et al. 2023), using energy scores to modulate prediction confidence and enhance robustness under real-world distribution shifts. Moreover, energy scores correlate strongly with prediction error, making them a natural metric for sample-level weighting in training objectives. In multimodal settings, such uncertainty measures can guide adaptive fusion by dynamically emphasizing more confident modalities, thereby mitigating modality-specific noise and improving overall model robustness.

The above work collectively advances VQA models toward enhanced robustness against real-world distribution shifts and spurious linguistic priors. To the best of our knowledge, this is ***the first attempt*** to explicitly address language priors in VQA under OOD conditions by adopting an energy-guided Bayesian paradigm that integrates energy-based uncertainty quantification into variational inference, enabling adaptive multimodal bias correction.

Methodology

Preliminaries

What is Bayes? Theoretically, Bayesian inference provides a principled solution for learning from data by updating prior beliefs about unknown quantities in light of observed evidence (Bishop and Nasrabadi 2006; Murphy 2012; Berger 2013). Let θ denote model parameters and $\mathcal{D}$ the observed data, Bayes' rule combines the prior $p(\theta)$ and the likelihood $p(\mathcal{D} \mid \theta)$ to yield the posterior

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)\, p(\theta)}{p(\mathcal{D})}, \quad p(\mathcal{D}) = \int p(\mathcal{D} \mid \theta)\, p(\theta)\, d\theta, \tag{1}$$

where $p(\mathcal{D})$ is the *evidence* (marginal likelihood). Predictions marginalize parameter uncertainty via the posterior predictive distribution

$$p(y \mid x, \mathcal{D}) = \int p(y \mid x, \theta)\, p(\theta \mid \mathcal{D})\, d\theta, \tag{2}$$

which naturally quantifies both aleatoric (data) and epistemic (model) uncertainties.

What is VQA? VQA is commonly formulated as multi-class classification over a fixed answer vocabulary $\mathcal{A}$. Given a dataset $\mathcal{D} = \{(v_i, q_i, a_i)\}_{i=1}^{N}$, where v_i is an image, q_i the associated question, and $a_i \in \{0,1\}^{|\mathcal{A}|}$ the one-hot ground-truth label, the model encodes visual and textual inputs as

$$h_{v,i} = e_v(v_i), \qquad h_{q,i} = e_q(q_i), \tag{3}$$

with $h_{v,i} \in \mathbb{R}^{d_v}$ and $h_{q,i} \in \mathbb{R}^{d_q}$, where $e(\cdot)$ denotes encoder function. A fusion module produces a joint representation,

$$h_i = \mathrm{Fuse}(h_{v,i},\, h_{q,i}) \in \mathbb{R}^d, \tag{4}$$

which is mapped by a classifier $c : \mathbb{R}^d \to \mathbb{R}^{|\mathcal{A}|}$ to logits $f_i = c(h_i)$, $|\mathcal{A}|$ is the number of output classes. Training typically minimizes the cross-entropy loss between the softmax probabilities and the target labels:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{|\mathcal{A}|} a_{i,k}\, \log \frac{\exp\big(f_{i,k}\big)}{\sum_{t=1}^{|\mathcal{A}|} \exp\big(f_{i,t}\big)}. \tag{5}$$

For one-hot targets, Eq. (5) reduces to the negative log-likelihood of the ground-truth class per example.

Problem Statement

Motivation Assumptions. VQA is prone to spurious *language priors*, whereby the question q can overly determine the answer distribution irrespective of the image. To explicitly capture and suppress such priors, we assume the existence of a latent correction variable δ_i that produces a debiased representation via an additive residual:

$$h_i^{\mathrm{corr}} \;=\; h_i \;+\; \delta_i, \qquad h_i, \delta_i \in \mathbb{R}^d. \tag{6}$$

Intuitively, the corrective perturbation δ_i is considered as a latent variable to quantify the uncertainty embedded in h_i.

To achieve this, our BayesVQA approach proposes a novel VQA debiasing training scheme from a probabilistic perspective, enabling uncertainty-aware adjustments that explicitly account for language-driven shortcuts rather than relying on a fixed point correction. Specifically, we leverage **Bayesian posterior inference** to model a proper distribution of the corrective perturbation δ_i and improve calibration and robustness under distribution shift.

Answer-Aware Gating. Here, we selectively augment the fused representation with an answer-aware feature using a lightweight gating mechanism. Given h_i and $h_{a,i}$, we compute a per-dimension gate and form a gated representation that blends the two features,

$$\mathrm{gate}_i \;=\; \sigma\big(W_g[\,h_i; h_{a,i}\,] + b_g\big) \in (0,1)^d, \tag{7}$$

$$h_i \;=\; \mathrm{gate}_i \odot h_i + \big(1 - \mathrm{gate}_i\big) \odot h_{a,i}, \tag{8}$$

with $W_g \in \mathbb{R}^{d \times 2d}$ and $b_g \in \mathbb{R}^d$, where $\sigma(\cdot)$ denotes Sigmoid function. Consequently, the gated representation h_i is used to condition the variational posterior, i.e., $q_\phi(\delta_i \mid h_i')$, thereby providing answer-consistent and visually-supported context for the correction δ_i.

Bayesian Posterior Inference

Rather than prescribing a deterministic correction, we infer a distribution over plausible corrections conditioned on the observed features, i.e., $p(\delta_i \mid h_i)$, thereby quantifying both the magnitude and direction of the additive adjustment in Eq. (6). This posterior view aligns with our objective: it enables uncertainty-aware debiasing of h_i by marginalizing over latent corrections instead of relying on a single point estimate. Based on Bayes' rule in Eq. (1), the posterior of δ is given by:

$$p(\delta_i \mid h_i) = \frac{p(h_i \mid \delta_i)\, p(\delta_i)}{p(h_i)} = \frac{p(h_i \mid \delta_i)\, p(\delta_i)}{\int p(h_i \mid \delta_i)\, p(\delta_i)\, d\delta_i}, \tag{9}$$

where $p(\delta)$ represents our prior belief about the corrective perturbation, and $p(h_i \mid \delta)$ denotes the likelihood of observing the fused representation given a particular correction.

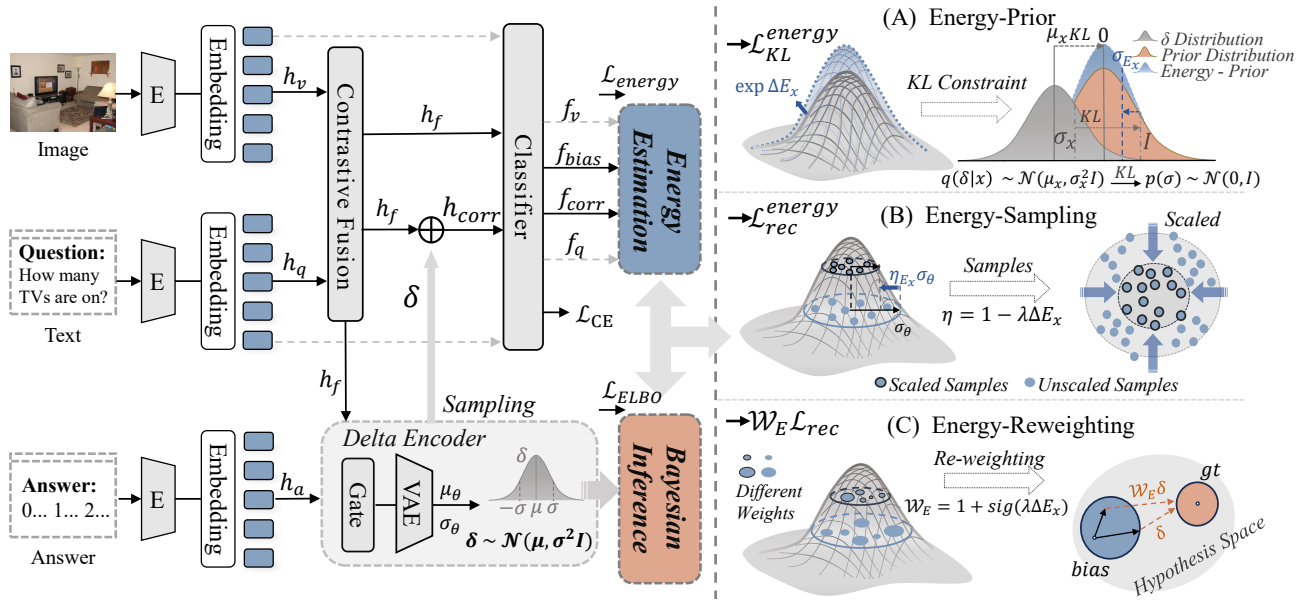


Figure 2: Illustration of the proposed BayesVQA framework. We first employ variational Bayesian inference to model the posterior distribution of the corrective perturbation δ. A DeltaEncoder network learns the mean and variance of a Gaussian posterior over δ from the fused multimodal feature, and we apply the reparameterization trick to sample δ. Building on this, we incorporate each sample's energy-based uncertainty score $E(h)$ into both the prior variance and the sampling scale, allowing the perturbation magnitude to adaptively expand or contract according to instance difficulty. This uncertainty-driven modulation not only enhances compensation for weak-modality features but also preserves overall stability and robustness in feature correction.

Variational Bayesian Inference. The marginal likelihood $p(h_i) = \int p(h_i \mid \delta)\, p(\delta)\, d\delta$ is generally intractable, which makes exact posterior inference $p(\delta \mid h_i)$ via Bayes' rule unavailable in practice. Therefore, we adopt amortized variational inference. Specifically, we posit a parameterized variational distribution $q_\phi(\delta \mid h_i)$, realized by an inference network (DELTAENCODER) (Schwartz et al. 2018) , and fit it by minimizing the KL divergence to the true posterior:

$$\phi^\star = \arg\min_{\phi} \mathrm{KL}\big(q_\phi(\delta \mid h_i) \,\|\, p(\delta \mid h_i)\big). \tag{10}$$

Here, minimizing $\mathrm{KL}(q_\phi \,\|\, p)$ is equivalent to maximizing the ELBO (Blei, Kucukelbir, and McAuliffe 2017), i.e.,

$$\mathcal{L}_{\mathrm{ELBO}}(\phi; h_i) = \mathcal{L}_{\mathrm{rec}}(\phi; h_i) - \mathcal{L}_{\mathrm{KL}}(\phi; h_i), \tag{11}$$

$$\text{where} \quad \mathcal{L}_{\mathrm{rec}}(\phi; h_i) = \mathbb{E}_{q_\phi(\delta|h_i)}\big[\log p(h_i \mid \delta)\big], \tag{12}$$

$$\mathcal{L}_{\mathrm{KL}}(\phi; h_i) = \mathrm{KL}\big(q_\phi(\delta \mid h_i) \,\|\, p(\delta)\big), \quad p(\delta) \sim \mathcal{N}(0, I). \tag{13}$$

In practice, we fix the prior $p(\delta)$ to the standard normal for traceability. For the reconstruction term $\mathcal{L}_{\mathrm{rec}}$, Eq. (10) cannot be directly differentiable, which blocks the backpropagation of the gradient. To address this, we employ the Monte Carlo samples reparameterization trick,

$$\widehat{\mathcal{L}}_{\mathrm{rec}}(\phi; h_i) = \frac{1}{m} \sum_{k=1}^{m} \log p\big(h_i \mid \delta^{(k)}\big), \tag{14}$$

$$\delta^{(k)} = \mu_\phi(h_i) + \sigma_\phi(h_i) \odot \varepsilon^{(k)}, \quad \varepsilon^{(k)} \sim \mathcal{N}(0, I), \tag{15}$$

where m is the number of samples per input, and $\mu_\phi(\cdot), \sigma_\phi(\cdot)$ are outputs of the DELTAENCODER. This reparameterization yields low-variance, fully differentiable gradients w.r.t. ϕ.

Energy-Guided Bayesian Debiasing

Overview of BayesVQA. Conventional variational inference typically adopts a fixed prior and a fixed-form variational family, limiting its capacity to adapt to heteroscedastic, sample-dependent difficulty and risk. In the context of VQA debiasing, certain fused representations are dominated by spurious language priors or exhibit elevated uncertainty. In our work, we directly adopt a supervised contrastive fusion objective over the unimodal representations h_v and h_q to explicitly align vision–language semantics and regularize the joint representation space (Zou et al. 2023). After answer-aware gating, our BayesVQA approach operationalizes an energy-driven Bayesian scheme in which the per-sample energy on the corrected representation is mapped to three coefficients that steer inference, including (i) Energy-Guided Prior Variance, (ii) Energy-Guided Posterior Sampling, (iii) Energy-Guided Likelihood Reweighting. The overall pipeline of BayesVQA is depicted in Fig. 2.

Energy Estimation & Normalization. As a central guiding signal, our BayesVQA approach aims to integrate an energy-based uncertainty measure into variational posterior inference. Note that lower energy typically corresponds to higher softmax confidence. To quantify per-sample uncertainty *after* correction, we evaluate a temperature-calibrated energy on the obtained representations (Zhang et al. 2023),

$$E(h) = -t \cdot \log \sum_{k=1}^{|\mathcal{A}|} \exp(f_k(h)/t), \tag{16}$$

where $f_k(h)$ denotes the k-th logit, $t = 1$ denotes a temperature parameter.

For cross-batch comparability, we define a centered (normalized) energy token as the deviation of each per-sample energy from the mini-batch mean, i.e.,

$$\Delta E = E_i^{\text{corr}} - \frac{1}{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|} E_j^{\text{corr}}, \tag{17}$$

where $|\mathcal{B}|$ is the mini-batch size. ΔE is then used to modulate energy-guided components (e.g., prior variance, posterior sampling scale, and reconstruction weight). To avoid circular dependencies and ensure stable ELBO optimization, we stop gradients through ΔE.

Energy-Guided Prior Variance. In standard variational inference, the corrective perturbation δ is assigned a fixed standard normal prior (cf. Eq. (13)), i.e., $p(\delta) \sim \mathcal{N}(0, I)$ with identity *covariance*. Such a prior treats all samples identically, regardless of their difficulty. To encode uncertainty, we replace it with an energy-guided prior variance generated by the normalized energy token, i.e.,

$$\text{Var}_i = \exp(-\lambda_1 \cdot \Delta E), \tag{18}$$

where λ_1 is a scaling coefficient. Hence, we can define the energy-conditioned prior as

$$\widetilde{p}(\delta_i) \sim \mathcal{N}\big(0,\ \text{Var}_i \cdot I\big), \tag{19}$$

and turning Eq. (13) with an energy-guided prior:

$$\mathcal{L}_{\text{KL}} \to \mathcal{L}_{\text{KL}}^{\text{energy}}(\phi; h_i) = \text{KL}\big(q_\phi(\delta_i \mid h_i) \,\|\, \widetilde{p}(\delta_i)\big). \tag{20}$$

With Eqs. (18) and (20), our work adopts a conservative-when-uncertain scheme: high-energy (*low-confidence*) samples use a *contracted* prior variance to enforce stronger regularization and limit the magnitude of corrections, while low-energy (*high-confidence*) samples use an *expanded* prior variance to encourage exploration.

Energy-Guided Posterior Sampling. During posterior sampling, the reparameterized posterior in Eq. (15) typically assumes a stationary sampling variance (e.g., a constant or globally shared scale), thereby treating all samples identically. To adaptively adjust sampling dispersion based on uncertainty, our BayesVQA utilizes an energy-scale factor that expands or contracts the dispersion of posterior draws,

$$\eta_i = 1 - \lambda_1 \cdot \Delta E, \tag{21}$$

and sample as

$$\delta_E^{(k)} = \mu_\theta(h_i) + \big(\eta_i \cdot \sigma_\theta(h_i)\big) \odot \epsilon^{(k)}, \epsilon^{(k)} \sim \mathcal{N}(0, I). \tag{22}$$

As a result, we augment the reparameterization in Eq. (15) with an energy-scale factor:

$$\widehat{\mathcal{L}}_{\text{rec}} \to \mathcal{L}_{\text{rec}}^{\text{energy}}(\phi; h_i) \ = \ \frac{1}{m}\sum_{k=1}^{m} \log p\big(h_i \mid \delta_E^{(k)}\big), \tag{23}$$

In practice, the proposed posterior sampling strategy further rescales the dispersion of reparameterized posterior draws according to per-sample energy, contracting it for high-energy (low-confidence) instances and expanding it for low-energy (high-confidence) instances.

Energy-Guided Likelihood Reweighting. To adaptively emphasize uncertain samples during reconstruction, we employ an energy-based reweighting scheme by defining a dynamic per-sample importance weight:

$$w_i = \ 1 + \sigma\big(\lambda_1 \cdot \Delta E\big). \tag{24}$$

Accordingly, we can further augment the Monte Carlo estimator associated with Eq. (15) by attaching the corresponding importance weight w_i:

$$\widehat{\mathcal{L}}_{\text{rec}} \to w_i \cdot \mathcal{L}_{\text{rec}}^{\text{energy}}(\phi; h_i) \ = \ w_i \cdot \frac{1}{m}\sum_{k=1}^{m} \log p\big(h_i \mid \delta_E^{(k)}\big). \tag{25}$$

This preserves the differentiability of the pathwise gradient estimator and, via the per-sample weight w_i, adaptively upweights high-uncertainty (high-energy) instances while downweighting confident ones, thereby concentrating corrective updates on challenging cases and improving stability while mitigating overfitting.

Energy Variational Bayes. Based on the above analysis, the objective of variational Bayesian inference (cf. Eq. (11)) can be reformulated as the *energy-guided* ELBO:

$$\mathcal{L}_{\text{ELBO}}^{\text{energy}}(\phi; h_i) = w_i \cdot \mathcal{L}_{\text{rec}}^{\text{energy}}(\phi; h_i) - \mathcal{L}_{\text{KL}}^{\text{energy}}(\phi; h_i). \tag{26}$$

Benefitting from Eq. (26), coupling the energy measure with variational inference enables per-sample adaptation according to sample-specific uncertainty, thereby strengthening bias mitigation on bias instances and yielding more balanced, robust multimodal representations.

Training and Optimization

Energy Coherence Loss. To preserve reliable predictions under uncertainty and to encourage higher confidence after debiasing, an energy coherence loss is also designed to align uncertainty before and after correction. Specifically, the energy coherence loss (ECL) is defined as:

$$\mathcal{L}_{\text{ECL}} = \mathbb{E}\Big[\max\big(0,\ E^{\text{corr}} - E^{\text{bias}}\big)\Big]. \tag{27}$$

Functionally, $\mathcal{L}_{\text{ECL}}$ serves as an uncertainty monotonicity regularizer that constrains the corrected energy not to exceed its original energy, i.e., $E^{\text{corr}} \le E^{\text{bias}}$.

Cross-Modal Alignment. Let $E_i^v = E\big(f_i^v\big)$ and $E_i^q = E\big(f_i^q\big)$ denote the per-sample energies computed from the visual-only and question-only branches, respectively (cf. Eq. (16)). To align these two modality-specific uncertainty estimates, we minimize the mean-squared error between them:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{2B}\sum_{i=1}^{B}\Big[\big(E_i^{\text{v}} - \widehat{E}^{\text{corr}}\big)^2 + \big(E_i^{\text{q}} - \widehat{E}^{\text{corr}}\big)^2\Big], \tag{28}$$

where $\widehat{E}^{\text{corr}}$ is the batch-average energies term as defined in Eq. (17). Minimizing $\mathcal{L}_{\text{MSE}}$ discourages overconfidence in either modality alone and encourages balanced reliance on both visual and language cues during fusion, thereby improving overall robustness and calibration.

Datasets		VQA-CP v2				VQA-CP v1			
Methods		Overall	Y/N	Num	Others	Overall	Y/N	Num	Others
UpDn (Anderson et al. 2018)	CVPR'18	39.74	42.27	11.93	46.05	37.96	42.79	12.41	42.53
LMH (Clark, Yatskar, and Zettlemoyer 2019)	EMNLP'19	52.15	70.29	44.10	44.86	55.73	78.59	24.68	45.47
GGE-iter (Han, Wang, and Su 2021)	ICCV'21	57.12	87.35	26.16	49.77	59.82	85.52	28.93	46.67
AdaVQA (Guo et al. 2021)	IJCAI'21	54.02	70.83	49.00	46.29	61.20	91.17	41.34	39.38
COB (Jha et al. 2023)	WACV'23	57.53	88.36	28.81	49.27	60.98	87.41	32.02	46.34
GENB (Cho et al. 2023)	CVPR'23	59.15	88.03	40.05	49.25	62.74	86.18	43.85	47.03
GGD (Han et al. 2023)	TPAMI'23	59.37	88.23	38.11	49.82	-	-	-	-
PWVQA (Vosoughi et al. 2024)	TMM'24	59.06	88.26	52.89	45.45	-	-	-	-
CVIV (Pan et al. 2024)	TMM'24	60.08	88.85	40.77	50.30	-	-	-	-
PDGH (Liu et al. 2025)	AAAI'25	61.68	89.29	53.13	**50.32**	64.56	89.56	47.35	46.01
BayesVQA	Ours	**62.08**	**89.69**	**55.88**	49.66	**66.91**	**91.74**	**54.31**	**47.30**

Table 1: Comparisons with the state-of-the-art methods on the VQA-CP v2 and VQA-CP v1 datasets.

Total Loss of BayesVQA. Overall, the total training objective of the proposed method is composed as follows:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{CE}} - \mathcal{L}_{\text{ELBO}} + \lambda_2 \mathcal{L}_{\text{ECL}} + \mathcal{L}_{\text{MSE}}, \quad (29)$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss based on the corrected representation h^{corr}. And λ_2 is a hyperparameter. During inference, we retain only the original fused representation h for prediction without the perturbation δ, the corrected feature h^{corr}, or any energy-related modules, yielding a simple and efficient inference procedure.

Experiments

Dataset, Metric, and Implementation Details

We evaluated the robustness of our approach against real-world biases on various bias-sensitive benchmarks, such as VQA-CP v1, VQA-CP v2 (Agrawal et al. 2018), and VQA-CE (Dancette et al. 2021). Additionally, we evaluated the in-distribution (ID) performance of the proposed BayesVQA method on the VQA v2 (Goyal et al. 2017) dataset. All experiments follow the standard VQA evaluation metric (Antol et al. 2015). To assess the efficiency of our proposed model, we conduct comparative experiments with the most relevant methods. We implemented our BayesVQA model in PyTorch with a single RTX 4090 GPU. The batch size B is set to 512. The learning rate is set to 0.002.

Comparisons with State-of-the-Arts

Evaluation on VQA-CP v2 and VQA-CP v1. Table 1 reports both the overall and per-category accuracy on the VQA-CP v2 and VQA-CP v1 datasets. BayesVQA achieves state-of-the-art performance, with an overall accuracy of 62.08% on VQA-CP v2 and 66.91% on VQA-CP v1, surpassing the second-best methods by 0.40% and 2.35%, respectively. Notably, BayesVQA leads in the "Yes/No" and "Number" categories—question types that are vulnerable to language bias. On VQA-CP v2, BayesVQA achieves 89.69% on "Yes/No" and 55.88% on "Number", indicating strong mitigation of language bias and improved visual grounding. The concurrent gain on "Other" also reflects balanced generalization to diverse, unconstrained questions.

Evaluation on VQA-CE and VQA-V2. Table 3 reports the performance of BayesVQA and several baselines on the VQA-CE dataset. Our method achieves the highest overall accuracy of 61.53%, outperforming prior approaches like GenB and RMLVQA. Although the counter category accuracy is slightly lower than some baselines, BayesVQA shows notable improvement on the easy subset, reflecting its strong ability to leverage reliable visual cues. These results demonstrate the effectiveness of our energy-guided Bayesian debiasing framework in improving generalization and mitigating language bias.

On the in-distribution VQA-v2 dataset, BayesVQA attains accuracy competitive with strong baselines, indicating that the debiasing neither overcorrects nor degrades standard performance and preserves reliable multimodal reasoning.

Extensive Experiments with Different Architectures. To assess generalizability, we integrate BayesVQA into three additional backbones: SAN (Yang et al. 2016), S-MRL (Cadene et al. 2019), and LXMERT (Tan and Bansal 2019). As reported in Table 3, SAN+BayesVQA boosts overall accuracy from 26.88% to 49.41%, a remarkable 22.53-point gain. Similarly, S-MRL+BayesVQA rises from 38.46% to 59.74%, improving by 21.28 points. These results demonstrate that our energy-guided Bayesian debiasing consistently enhances diverse VQA architectures. Preliminary results on LXMERT also indicate substantial improvements, further validating BayesVQA's adaptability across varied multimodal models.

Ablation Studies

To thoroughly examine the individual contributions of each proposed component, we conduct comprehensive ablation studies on the VQA-CP v2 benchmark. As illustrated in Table 4, our experiments yield three primary observations: First, combining EPV and EPS consistently enhances model

Methods	All	Y/N	Num	Other	Increased
SAN	26.88	35.34	11.34	24.70	**22.53** ↑
SAN+BayesVQA	**49.41**	**88.02**	**39.98**	**31.77**	
S-MRL	38.46	42.85	12.81	43.20	**21.28** ↑
S-MRL+BayesVQA	**59.74**	**88.93**	**49.50**	**47.27**	
LXMERT	48.66	47.49	22.24	56.52	**17.61**↑
LXMERT+BayesVQA	**66.27**	**92.30**	**33.32**	**61.67**	

Table 2: Performance of our BayesVQA with different backbone architectures.

Datasets	VQA-CE			VQA-V2			
Methods	Overall	Counter	Easy	Overall	Y/N	Num	Others
CSS	53.55	34.36	62.08	59.91	73.25	**39.77**	**55.11**
GENB	57.87	34.80	68.15	-	-	-	-
RMLVQA	58.05	**35.01**	68.21	59.99	76.68	37.54	53.26
BayesVQA	**61.53**	33.12	**74.22**	**60.09**	**77.73**	34.43	53.48

Table 3: Comparisons on VQA-CE and VQA v2 datasets.

Methods	EPV	EPS	ELR	ECL	Overall-CP
Variant-I	✓	-	-	-	61.32
Variant-II	✓	✓	-	-	61.55
Variant-III	-	-	✓	-	61.58
Variant-IV	✓	-	-	✓	61.69
BayesVQA (Ours)	✓	✓	✓	✓	**62.08**

Table 4: Ablation experiments on the VQA-CP v2 dataset.

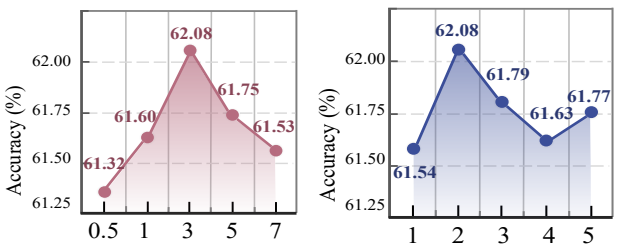
performance, demonstrating that incorporating uncertainty-aware Bayesian inference effectively facilitates adaptive bias correction. Second, ELR provides further performance improvements by dynamically modulating the reconstruction loss, emphasizing challenging instances, enhancing feature discriminability, and reducing residual biases. Finally, ECL refines the prior distribution to ensure cross-modal semantic consistency, which substantially boosts reasoning accuracy.

Parameter Analysis

We perform an in-depth parameter analysis of the proposed BayesVQA method by investigating its behavior under different hyperparameter configurations. Our study focuses on two key hyperparameters: λ_1 in Eqn. (18), and λ_2 in Eqn. (29). Systematic experiments, as shown in Figure 3, reveal that the model achieves optimal performance when $\lambda_1 = 3$, and $\lambda_2 = 0.02$. This finding underscores that an optimal combination of these hyperparameters is critical for maximizing the performance of the BayesVQA model.

Qualitative Analysis

Figure 4 visualizes attentions for BayesVQA versus a strong baseline; red boxes mark the highest-scoring regions. The proposed BayesVQA consistently attends to question-relevant visual evidence, whereas the baseline often locks



(a) λ_1 (b) $\lambda_2 \times 100$

Figure 3: Comparison of Accuracy on the VQA-CP v2 dataset with different parameter configurations.

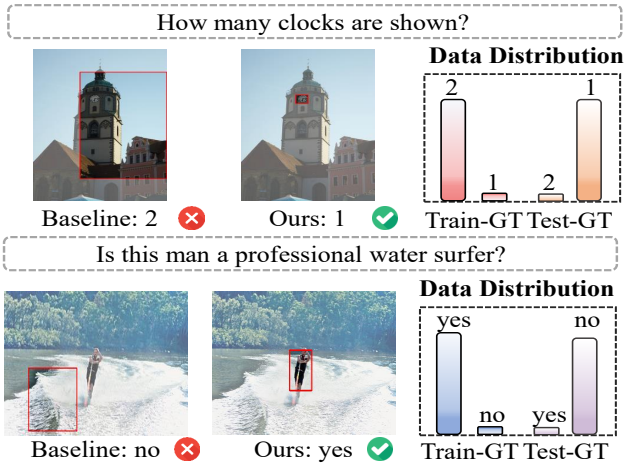


Figure 4: **Qualitative analysis of BayesVQA.** The red boxes show the highest-scoring region of models. Compared to the baseline, our BayesVQA method focuses on the correct visual regions, indicating successful alleviation of bias-related challenges and achieving robust debiased learning.

onto irrelevant areas driven by language priors. This behavior stems from our energy-guided variational Bayesian inference, which adaptively calibrates corrective perturbations according to sample-level uncertainty—dynamically shaping the prior, posterior sampling, and likelihood weighting. As a result, BayesVQA suppresses spurious language biases, strengthens visual grounding, and yields more robust, interpretable multimodal reasoning.

Conclusion

This study proposed a novel Bayesian debiasing framework, BayesVQA, which leverages energy-based uncertainty to guide the inference process for robust multimodal reasoning under out-of-distribution scenarios. By integrating energy signals into prior modeling, posterior sampling, and likelihood reweighting within the variational Bayesian inference framework, our BayesVQA achieves adaptive bias correction tailored to each sample's confidence. Extensive experiments demonstrate that our method consistently outperforms existing approaches on several challenging VQA benchmarks, highlighting its effectiveness and strong generalization in complex real-world settings.

Acknowledgments

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation (Nos. 2025A1515010225, 2025A1515060001), in part by the National Natural Science Foundation of China (No. 62302172), and in part by the Key Scientific Research Platforms and Projects of Regular Colleges and Universities in Guangdong Province (No. 2025ZDZX3043).

References

Agrawal, A.; Batra, D.; Parikh, D.; and Kembhavi, A. 2018. 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)*, 4971–4980.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6077–6086.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2425–2433.

Basu, A.; Addepalli, S.; and Babu, R. V. 2023. RMLVQA: A Margin Loss Approach for Visual Question Answering With Language Biases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 11671–11680.

Berger, J. O. 2013. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Bishop, C. M.; and Nasrabadi, N. M. 2006. *Pattern recognition and machine learning*. Springer.

Blei, D. M.; Kucukelbir, A.; and McAuliffe, J. D. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518): 859–877.

Cadene, R.; Dancette, C.; Ben-younes, H.; Cord, M.; and Parikh, D. 2019. Rubi: Reducing unimodal biases in visual question answering. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, volume 32.

Chen, L.; Yan, X.; Xiao, J.; Zhang, H.; Pu, S.; and Zhuang, Y. 2020. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 10800–10809.

Cho, J. W.; Kim, D.-J.; Ryu, H.; and Kweon, I. S. 2023. Generative bias for robust visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 11681–11690.

Clark, C.; Yatskar, M.; and Zettlemoyer, L. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4069–4082.

Dancette, C.; Cadene, R.; Teney, D.; and Cord, M. 2021. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1574–1583.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6904–6913.

Guo, Y.; Nie, L.; Cheng, Z.; Ji, F.; Zhang, J.; and Del Bimbo, A. 2021. AdaVQA: Overcoming Language Priors with Adapted Margin Cosine Loss. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

Han, X.; Wang, S.; and Su, C. 2021. Greedy gradient ensemble for robust visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1584–1593.

Han, X.; Wang, S.; Su, C.; Huang, Q.; and Tian, Q. 2023. General greedy de-bias learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45: 1–17.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6700–6709.

Jha, A.; Patro, B.; Van Gool, L.; and Tuytelaars, T. 2023. Barlow constrained optimization for visual question answering. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1084–1093.

Kv, G.; and Mittal, A. 2020. Reducing language biases in visual question answering with visually-grounded question encoder. In *European Conference on Computer Vision (ECCV)*, 18–34. Springer.

Liang, Z.; Jiang, W.; Hu, H.; and Zhu, J. 2020. 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)*, 3285–3292.

Liu, Y.; Zhu, J.; Wen, C.; Lu, G.; Lin, H.; and Chen, B. 2025. Towards Robust Visual Question Answering via Prompt-Driven Geometric Harmonization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, 5721–5729.

Ming, Y.; Fan, Y.; and Li, Y. 2022. Poem: Out-of-distribution detection with posterior sampling. In *International Conference on Machine Learning (ICML)*, 15650–15665. PMLR.

Murphy, K. P. 2012. *Machine learning: a probabilistic perspective*. MIT press.

Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.-S.; and Wen, J.-R. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 12700–12710.

Pan, Y.; Liu, J.; Jin, L.; and Li, Z. 2024. Unbiased visual question answering by leveraging instrumental variable. *IEEE Transactions on Multimedia (TMM)*, 26: 6648–6662.

Ramakrishnan, S.; Agrawal, A.; and Lee, S. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, volume 31.

Schwartz, E.; Karlinsky, L.; Shtok, J.; Harary, S.; Marder, M.; Kumar, A.; Feris, R.; Giryes, R.; and Bronstein, A. 2018. Delta-encoder: an effective sample synthesis method for few-shot object recognition. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, volume 31.

Subedar, M.; Krishnan, R.; Meyer, P. L.; Tickoo, O.; and Huang, J. 2019. Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 6301–6310.

Tan, H.; and Bansal, M. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5100–5111.

Teney, D.; Kafle, K.; Shrestha, R.; Abbasnejad, E.; Kanan, C.; and van den Hengel, A. 2020. On the value of out-of-distribution testing: an example of goodhart's law. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 407–417.

Vosoughi, A.; Deng, S.; Zhang, S.; Tian, Y.; Xu, C.; and Luo, J. 2024. Cross modality bias in visual question answering: A causal view with possible worlds vqa. *IEEE Transactions on Multimedia (TMM)*.

Wang, Q.; Yu, Y.; Yuan, Y.; Mao, R.; and Zhou, T. 2025. VideoRFT: Incentivizing Video Reasoning Capability in MLLMs via Reinforced Fine-Tuning. *arXiv preprint arXiv:2505.12434*.

Wen, Z.; Xu, G.; Tan, M.; Wu, Q.; and Wu, Q. 2021. Debiased visual question answering from feature and sample perspectives. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, volume 34, 3784–3796.

Wilson, A. G.; and Izmailov, P. 2020. Bayesian deep learning and a probabilistic perspective of generalization. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, volume 33, 4697–4708.

Xue, D.; Qian, S.; and Xu, C. 2023. Variational causal inference network for explanatory visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2515–2525.

Yang, Z.; He, X.; Gao, J.; Deng, L.; and Smola, A. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 21–29.

Zhang, Q.; Wu, H.; Zhang, C.; Hu, Q.; Fu, H.; Zhou, J. T.; and Peng, X. 2023. Provable dynamic fusion for low-quality multimodal data. In *International Conference on Machine Learning (ICML)*, 41753–41769. PMLR.

Zhao, C.; Mei, S.; Ni, B.; Yuan, S.; Yu, Z.; and Wang, J. 2023. Variational adversarial defense: A bayes perspective for adversarial training. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(5): 3047–3063.

Zhu, H.; Liu, Y.; Fang, X.; Lu, G.; and Chen, B. 2025. Cause-Effect Driven Optimization for Robust Medical Visual Question Answering with Language Biases. *arXiv preprint arXiv:2506.17903*.

Zhu, J.; Liu, Y.; Zhu, H.; Lin, H.; Jiang, Y.; Zhang, Z.; and Chen, B. 2024. Combating Visual Question Answering Hallucinations via Robust Multi-Space Co-Debias Learning. In *Proceedings of the ACM International Conference on Multimedia (MM)*, 955–964.

Zhu, X.; Mao, Z.; Liu, C.; Zhang, P.; Wang, B.; and Zhang, Y. 2021. Overcoming language priors with self-supervised learning for visual question answering. In *Proceedings of the International Conference on International Joint Conferences on Artificial Intelligence (IJCAI)*, 1083–1089.

Zou, H.; Shen, M.; Chen, C.; Hu, Y.; Rajan, D.; and Chng, E. S. 2023. UniS-MMC: Multimodal Classification via Unimodality-supervised Multimodal Contrastive Learning. In *Findings of the Association for Computational Linguistics (ACL)*, 659–672.