

Article

Prompt-Driven Fuzzing Debiasing Framework for Robust Visual Question Answering

Yali Fan ^{1,*} , Gangyu Huang ¹, Qiwen Lu ² and Shengbo Chen ^{3,*}¹ School of Economics and Management, Nanchang University, Nanchang 330047, China; 5409123112@email.ncu.edu.cn² School of Computer and Information Engineering, Henan University, Kaifeng 475001, China; easonglll@henu.edu.cn³ School of Software, Nanchang University, Nanchang 330047, China

* Correspondence: fanyali@ncu.edu.cn (Y.F.); ccb02kingdom@gmail.com (S.C.)

Abstract

Visual Question Answering (VQA) systems have achieved impressive performance with the rise of large-scale vision–language models (VLMs). However, these models remain vulnerable to multiple forms of multimodal bias, severely limiting their robustness and generalization. Existing debiasing techniques mainly depend on post hoc evaluation or architectural modifications, while recent prompt-learning-based methods reveal new opportunities for aligning downstream tasks with pretrained models. In this work, we propose a unified prompt-driven debiasing framework that integrates generative prompt learning and a fuzzing-based bias correction mechanism. The generative prompt component reformulates VQA as a cloze-style masked prediction problem, leveraging pretrained language priors to improve semantic grounding. Meanwhile, the fuzzing-based module actively constructs unexpected test samples during training and employs a reflection mechanism to correct biased predictions in-loop, yielding inference-time robustness without additional test-time components. Extensive experiments on VQA-v2, VQA-CP, VQA-CE, GQA-OOD, and VQA-VS demonstrate that the proposed framework significantly improves both in-distribution (ID) accuracy and out-of-distribution (OOD) robustness, outperforming existing prompt-only or data-augmentation-only debiasing methods.

Keywords: visual question answering; prompt; debias; language bias

1. Introduction

In recent years, the rapid development of large-scale pretrained models has substantially advanced a wide range of artificial intelligence tasks. Models such as GPT and BERT [1] have provided strong foundations for natural language understanding and generation. Meanwhile, in the field of vision–language learning, pretrained multimodal models based on the Transformer architecture [2] have also demonstrated remarkable representational power. The general pipeline underlying such models typically consists of two phases: (1) performing large-scale pretraining on massive image–text corpora to capture cross-modal relationships and (2) adapting the pretrained architecture to downstream tasks through task-specific structural modifications. However, as the scale and complexity of these models continue to grow, issues arise regarding training cost, parameter efficiency, and the feasibility of adapting them in resource-constrained settings.

To address this, recent work has explored a new paradigm: prompt learning [3]. Instead of modifying the pretrained model architecture, prompt learning reformulates



Academic Editor: Stephan Schlögl

Received: 30 April 2026

Revised: 16 July 2026

Accepted: 25 July 2026

Published: 31 July 2026

Copyright: © 2026 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

downstream tasks so that they naturally fit the pretrained model's inherent prediction objectives. In contrast to conventional pretraining + fine-tuning paradigms, prompt learning enables efficient zero-shot or few-shot adaptation by allowing downstream tasks to coexist with the pretraining objective. This motivated us to investigate whether a pretrained model—which has already learned rich multimodal knowledge—can be directly utilized through prompt-based reformulation without extensive architectural changes.

Building on this idea, we incorporate prompt learning into the Visual Question Answering (VQA) task. Traditional VQA systems treat the task as multi-class classification and attach an additional fully connected prediction layer to the pretrained backbone. In our method, however, we construct a template-based prompt and insert the image representation and question text into predefined slots, using the <mask> token to mark the answer position. This reformulates VQA into a Masked Language Modeling (MLM) task [1], consistent with the pretraining objective. Additionally, we integrate the concept of a Prefix Language Model to allow the model to generate the answer token based on all preceding contextual and visual information, making the VQA process more aligned with natural human reasoning. To further narrow the semantic gap between generated answers and ground-truth labels, we introduce a semantic similarity constraint that quantitatively measures and reduces semantic discrepancy during training. Through this prompt-based reconstruction of inputs and labels, our framework achieves notable performance improvements.

While prompt learning enhances the adaptation between VQA and pretrained models, biases remain a major obstacle in multimodal learning. VQA [4,5], as a representative multimodal reasoning task, requires jointly understanding images and text, but a large body of research has shown that current VQA models are highly susceptible to various types of biases [5,6]. Such biases arise not only from language priors but also from visual and cross-modal interactions, where frequent co-occurrences between certain words and objects mislead the model [7,8]. To mitigate these issues, prior works have proposed ensemble-based methods [9–13], data augmentation [14–18], contrastive learning [19–22], and answer re-ranking approaches [23–26]. Data augmentation and counterfactual methods modify training distributions, which may violate the design principles of benchmark datasets such as VQA-CP. MLLM-based approaches benefit from strong pretrained representations but introduce challenges in fair comparison due to architectural disparity and potential hallucination issues. Nevertheless, these methods either address only a subset of biases or rely on heuristics that lack generality [12,27,28]. Thus, constructing a unified, generalized debiasing framework remains an open problem [7,29].

Fundamentally, the “bias” in VQA models manifests as pathological shortcut learning, where the network exploits superficial data correlations (e.g., historical co-occurrences between specific question types and high-frequency labels) instead of learning causal vision–language alignments. Existing paradigms largely fall short of addressing this. Data rebalancing or counterfactual augmentation approaches inevitably deform the training distribution manifold, which frequently violates the design intent of benchmarks and triggers severe trade-offs. Meanwhile, post hoc evaluation or late-stage answer re-ranking acts only as a passive diagnostic tool after the model parameters have already been heavily contaminated by shortcuts. They fail to dynamically eliminate bias during the parameter optimization pipeline.

Consequently, treating these empirical symptoms post hoc or mutating data distributions externally fails to rectify the root cause embedded within the optimization objective itself. This structural failure creates an urgent programmatic necessity for a proactive, in-loop diagnostic and real-time repair mechanism that actively operates during the gradient descent loop. To implement such a dynamic defense, we draw a direct tactical philosophy from software engineering: rather than waiting for a system to fail in production, engineers

utilize “fuzz testing” to deliberately bombard code execution paths with unexpected inputs, thereby flushing out latent edge-case vulnerabilities before they solidify. Mapping this proactive philosophy to multimodal learning, a VQA model’s pathological shortcut habits can be precisely conceptualized as silent logic vulnerabilities. These biases remain hidden under standard in-distribution samples but systematically crash under atypical multimodal conditions.

To overcome these limitations, we argue that model bias shares an intrinsic structural equivalence with software vulnerabilities in software engineering. Just as a software bug lies dormant under standard operational workflows but precipitates system crashes when executing edge cases, a VQA model’s shortcut learning behaviors remain entirely masked by deceptively high accuracy under standard, in-distribution (ID) training samples. These insidious biases are only laid bare when the system is subjected to stringent boundary conditions. Consequently, drawing inspiration from automated software verification, we propose that the core metaphor of “fuzz testing”—which systematically probes system resilience via unanticipated inputs—can be naturally mapped to VQA debiasing. By synthesizing atypical multimodal probes as test cases during training, we can aggressively surface the model’s latent shortcut dependencies before they solidify into the weights.

A deeper examination of the debiasing process reveals an important insight: bias is learned during training. Since deep learning relies heavily on data [30], models inevitably absorb spurious correlations present in the training set [31]. By the time performance is evaluated on the held-out test set, the model has already been affected by bias. A natural question arises: Can we monitor and correct bias during training, similar to evaluating model behavior on test data, without violating the separation between training and test sets?

To answer this, we draw inspiration from fuzz testing in software engineering. The parallel between model debiasing and software fuzzing is grounded in a shared structural principle: both processes systematically probe system behavior under atypical conditions. In software testing, fuzzing generates unexpected inputs to trigger crashes or undefined behavior, with fuzzers monitoring program execution for violations [32]. Analogously, in VQA training, we can generate atypical multimodal samples that trigger the model’s reliance on spurious correlations. Just as fuzzers monitor program execution for violations, our framework monitors model predictions for signs of biased reasoning. The key insight is that biases—like software bugs—are latent during normal operation but manifest when the system encounters inputs outside its comfort zone. With the rapid development of deep learning systems, fuzzing has been increasingly applied to detect vulnerabilities in deep learning libraries [33]. Inspired by this, we propose a Fuzzing-based Debiasing Framework (FD), in which fuzz-inspired unexpected samples are generated during training to probe the model’s susceptibility to different types of bias.

Unlike traditional data augmentation methods [14–18], our additional data are not intended to balance or enrich the training distribution but to test the model for bias in each training epoch. If biases are detected, we employ a new bias evaluation criterion to measure the severity of the model’s biased behavior and introduce a reflection mechanism that adjusts the model parameters accordingly. This process mirrors software debugging: fuzzing identifies vulnerabilities, evaluation metrics assess their severity, and the reflection mechanism “patches” the model by suppressing biased representations.

In summary, our work integrates prompt learning and fuzzing-based debiasing to jointly enhance VQA robustness. The contributions are as follows:

- We design a prompt-based framework that reformulates VQA into a cloze-style prediction task, effectively aligning downstream objectives with pretrained language models.

- We introduce a novel fuzzing-based debiasing framework that detects and corrects model bias during training, rather than relying solely on post hoc evaluation.
- We develop new bias evaluation criteria and a reflection mechanism for quantifying and mitigating multimodal biases in real time.
- Comprehensive evaluation across multiple benchmarks—including VQA v2, VQA-VS, VQA-CPv2, GQA-OOD, and VQA-CE—verifies that our proposed method regularly surpasses existing state-of-the-art techniques for debiasing and improving robustness.

2. Related Work

2.1. VQA

Visual Question Answering (VQA) [4] has evolved rapidly as a representative multimodal reasoning task, spurring the creation of numerous benchmark datasets and pre-trained models [4,24,34,35]. However, research indicates that models often rely on language, visual, and multimodal biases within datasets rather than performing genuine reasoning [5–8], which has motivated the establishment of out-of-distribution (OOD) benchmarks such as VQA-CP [24] and VQA-VS [8] to evaluate robustness. To mitigate these spurious correlations, existing debiasing research is primarily categorized into ensemble methods [9–12,24], data augmentation [14–18], contrastive learning [19–22], and answer re-ranking [23–26]. Nevertheless, most existing techniques are limited to specific architectures or fail to comprehensively address the wide variety of biases in VQA [27,28,36]. The method proposed in this paper broadly falls under data augmentation; however, unlike previous approaches that focus on rebalancing training data, we aim to generate *unexpected* inputs to assist in bias detection and correction, optimizing predictions and suppressing bias during training through the introduction of a bias fairness indicator and a reflection mechanism.

2.2. Prompt Learning

Prompt learning [3] has emerged as an influential paradigm for adapting pretrained language models to downstream tasks. By converting input samples into cloze-style templates, models can directly leverage their pretrained knowledge without extensive fine-tuning. Early works such as PET [37], KPT [38], and subsequent studies [39–41] demonstrate the effectiveness of prompt-based adaptation in NLP tasks.

With the rise of multimodal learning, prompt learning has also been explored in visual–language tasks. PICa [42] uses CLIP [43] to convert images into textual descriptions and applies prompt-based reasoning; CoOp [44] introduces learnable prompts for VLM adaptation; CPT [45] reformulates multimodal co-reference into cloze-style questions. Inspired by these works, our method explores prompt learning within the VQA task and integrates it with debiasing.

2.3. Fuzzing

Fuzzing [32] is an automated software testing technique designed to expose system vulnerabilities by generating invalid, random, or unexpected inputs. With the emergence of large language models (LLMs), fuzzing has increasingly been combined with deep learning systems. Recent works explore LLM-assisted fuzzing across multiple domains, including DL library fuzzing, compiler fuzzing, protocol fuzzing, general-purpose fuzzing, and fuzz driver generation. In this study, we draw inspiration from fuzzing to generate unexpected data samples during VQA training, enabling real-time bias detection and correction without violating the integrity of test set separation.

3. Proposed Method

In Section 3, we introduce a prompt-guided fuzzing debiasing framework designed specifically for Visual Question Answering (VQA). Unlike traditional fuzzing, which relies on random perturbations, our method leverages prompt engineering and generative models to synthesize high-level, semantically structured, and model-unexpected test samples. These test samples expose multimodal bias during training, enabling the model to improve robustness and generalization. The overall framework is shown in Figure 1.

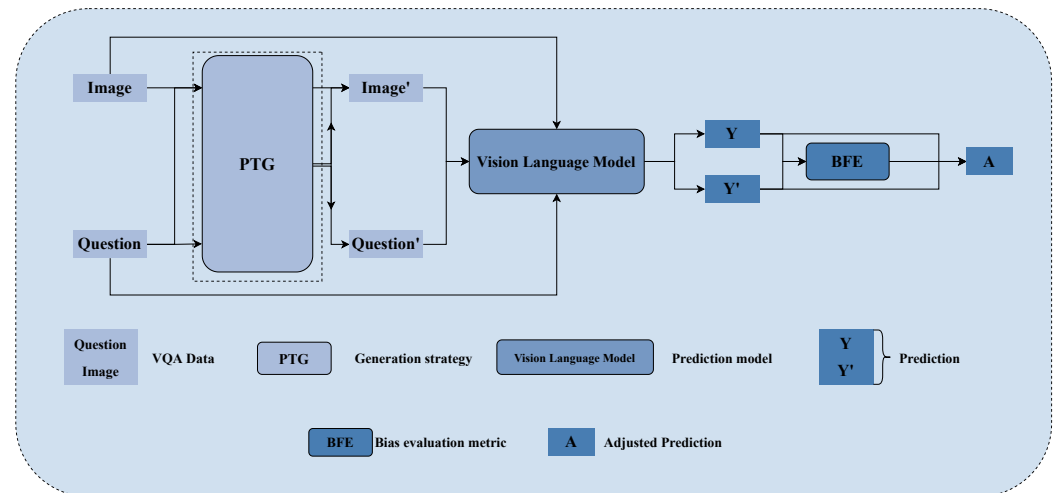


Figure 1. Overall framework of the proposed prompt-driven fuzzing debiasing method (FD). The framework consists of three main components: (1) Prompt-Engineered Test Generation (PTG) that produces unexpected multimodal samples, (2) a prompt-based VQA formulation that converts standard VQA into a cloze-style prediction task, and (3) a reflection mechanism that detects and corrects biased predictions through bias fairness evaluation.

3.1. Preliminaries and Theoretical Underpinnings

Fuzzing [32] is an established technique for verifying software reliability by providing unexpected inputs to reveal vulnerabilities. In deep learning, recent work has extended fuzzing to assess the robustness of neural networks. However, existing fuzzing approaches typically rely on random, low-level perturbations that do not reflect the semantic nature of multimodal tasks such as VQA.

To address this limitation, we propose a prompt-only fuzzing mechanism, in which fuzzing samples are generated not by random corruption but by instruction-driven prompt template combined with pretrained generative models. This approach enables the creation of structured, multimodally coherent yet unexpected samples, which are substantially more informative for bias detection.

To mathematically formulate why our prompt-guided fuzzing paradigm systematically exposes hidden linguistic biases, we analyze this procedure through the lens of Causal Inference and Information Theory.

- (1) **Causal Framework for Bias Exposure:** Let Q be the question text space, V be the visual scene feature space, and Y be the answer space. Traditional VQA datasets inherently suffer from a confounding language prior bias, denoted by a backdoor path $Q \rightarrow Y$ that circumvents the required multimodal reasoning path $Q \times V \rightarrow Y$. When a VQA model parameterizes $P_{\theta}(Y|Q, V)$, it tends to memorize the empirical statistical shortcut $P_{\text{data}}(Y|Q)$ rather than aligning visual grounds. Our Prompt-engineered Test

Generation (PTG) acts as a structural counterfactual intervention $\text{do}(Q = q')$, where q' is synthesized to satisfy:

$$q' \sim P_{\text{GEN}}(Q|v), \quad \text{s.t.} \quad P_{\text{data}}(Y|q') \neq P_{\text{data}}(Y|q). \quad (1)$$

In practice, this abstract causal intervention is directly operationalized by the generative PTG module detailed in Section 3.2, which breaks the confounding backdoor link. If the model relies on true visual alignment, its prediction $P_{\theta}(Y|q', v)$ remains robust; otherwise, the intervention $\text{do}(Q = q')$ triggers a catastrophic confidence shift detected by our differential BFE metric.

- (2) **Information-Theoretic Entropy Maximization:** From an information-theoretic view, a biased model possesses an anomalously low conditional entropy $H_{\theta}(Y|Q)$ over shortcut patterns, indicating over-confidence in surface textual clues. To rectify this, we penalize low-entropy shortcut prediction patterns. In our engineering design, this entropy constraint directly drives the formulation of the complementary target distribution P_{target} (detailed in Section 3.6, Equation (12)) and the bias-aware loss $\mathcal{L}_b$ (Section 3.7, Equation (23)). By constructing an inverse probability target $\hat{y}_i = \text{softmax}(-\log P'_i)$, we mathematically enforce an information-theoretic penalty that counteracts the low-entropy state of biased predictions, shifting the parameter landscape toward balanced, high-entropy multimodal grounding.

It is important to distinguish our fuzzing-based approach from conventional data augmentation. Data augmentation methods (e.g., CSS [15], MUTANT [14]) aim to modify or enrich the training data distribution to improve generalization by adding transformed copies of training samples. In contrast, our fuzzing samples serve as diagnostic probes: they are generated during each training epoch to test the model's current susceptibility to bias, and influence optimization only indirectly through the bias fairness evaluation and reflection mechanism. Specifically: (1) **Usage:** Data augmentation samples are permanently added to the training set, whereas fuzzing samples are conditionally activated only when bias is detected (i.e., $C_i > \tau$) and are discarded after the reflection step. (2) **Optimization path:** Data augmentation indirectly affects parameters by expanding the data distribution; our framework directly intervenes on sample-specific gradient signals through the bias-aware loss $\mathcal{L}_b$ and the reflection mechanism. (3) **Lifecycle:** Data augmentation samples persist throughout training; fuzzing samples are regenerated each epoch to adapt to the model's evolving state and discarded after bias evaluation.

3.2. Prompt-Engineered Test Case Generation (PTG)

The core of our debiasing strategy is Prompt-engineered Test Generation (PTG), in which we use large pretrained generative models to synthesize abnormal but semantically meaningful VQA samples. Unlike previous perturbation-based fuzzing methods, PTG maintains syntactic and contextual plausibility while intentionally steering questions away from the original data manifold.

Given a standard VQA sample (v_i, q_i, a_i) , PTG generates a test question q'_i and a visual feature v'_i using the following generative prompt design.

3.2.1. Language Test Sample Generation

To actively test whether the model is trapped by systematic language prior dynamics, we pass the problem context through an instruction-driven language forecasting distribution $P_{\Phi}(q'|q)$. The instructions explicitly mandate the synthesizer to deviate from historical data patterns while preserving semantic plausibility. Consequently, q'_i behaves as an active linguistic probe rather than random syntactic noise, directly validating the causal boundary conditions of the primary vision-language encoder.

3.2.2. Visual Test Sample Generation

Symmetrically, we employ a non-parametric multi-region feature distribution generator $G_{\Psi}(v'|v)$ conditioned on the source cross-modal sample. The semantic prompt forces the spatial synthesizer to hallucinate atypical cross-modal scene alignments and uncommon regional compositions. This ensures that the resultant visual output v'_i exposes whether the underlying VQA network establishes authentic spatial grounding or merely executes empirical pattern re-weighting correlations on standard manifolds.

3.2.3. Resulting Prompt-Guided Test Samples

Each PTG operation yields (v'_i, q'_i) , which differs from the training distribution yet preserves semantic structure and multimodal consistency. This ensures that fuzzing evaluates the model's reasoning rather than its ability to handle noise.

Figure 2 illustrates the PTG process.

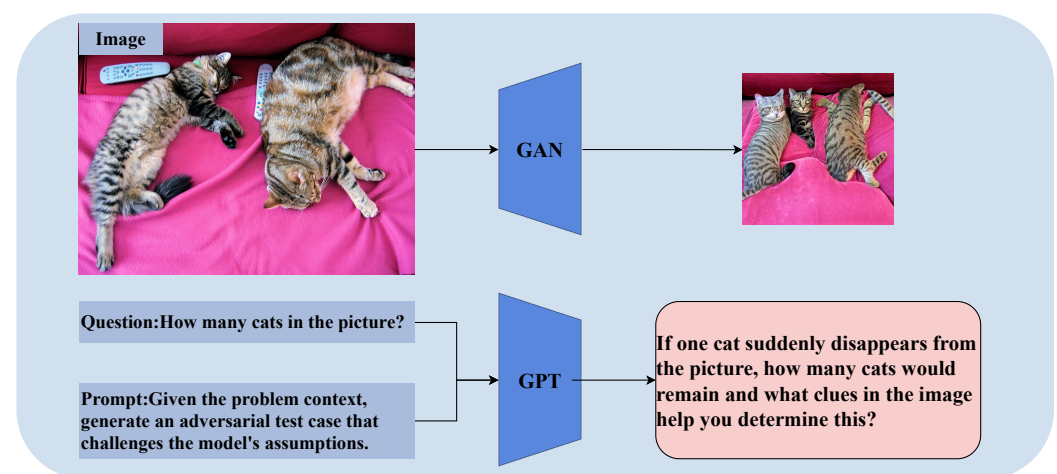


Figure 2. Prompt-engineered test sample generation (PTG). Pretrained generative models produce linguistically coherent but unexpected questions and visually plausible but atypical visual features, enabling high-level fuzzing for VQA.

3.3. Prompt-Based VQA Formulation

Following recent advances in vision–language pretraining, we adopt a prompt-learning formulation of VQA to better align visual and textual reasoning. Specifically, we convert each VQA sample into a template-based cloze question: "Due to <image_feature>, regarding the question <q>, the answer is <mask_>." The model is required to generate the masked answer token(s). This approach provides two benefits: (1) It aligns the VQA task with language-model pretraining objectives such as masked language modeling. (2) It allows the model to encode answer semantics directly in natural language space, reducing answer-category bias.

3.4. Formal Definitions and Bias Fairness Evaluation

Before presenting the label construction and reflection mechanisms, we provide formal definitions of all key variables and the Bias Fairness Evaluation (BFE) metric.

Notation. We represent the VQA training set as $\mathcal{D} = \{(v_i, q_i, a_i)\}_{i=1}^N$, encompassing the image feature v_i , the query q_i , and the corresponding ground-truth answer a_i . Given an input instance, the model estimates a probability distribution over an answer vocabulary $\mathcal{A}$ with a cardinality of K :

$$y_i = P_{\theta}(a|v_i, q_i) \in \mathbb{R}^K, \quad (2)$$

where P_θ denotes the model parameterized by θ . For the corresponding fuzzed sample (v'_i, q'_i) generated by PTG, the model produces:

$$y'_i = P_\theta(a|v'_i, q'_i). \quad (3)$$

Bias Fairness Evaluation (BFE). It is critical to clarify that the BFE metric is not designed to enforce a static, absolute uniform distribution constraint over all normal VQA inferences, which would indeed introduce a directional error by penalizing valid, high-confidence reasoning. Instead, under Jaynes' Principle of Maximum Entropy, when a model is confronted with an atypical, out-of-distribution counterfactual probe (v'_i, q'_i) synthesized by PTG that explicitly strips away historical correlation shortcuts, the model's posterior distribution $P_\theta(Y|v'_i, q'_i)$ should naturally display epistemic uncertainty (approaching a uniform distribution) if it lacks authentic visual reasoning support. If, however, the model remains pathologically over-confident on the fuzzed probe, it indicates reliance on rigid, visual-insensitive language shortcuts. Formally, the BFE for a prediction y over K answer candidates is defined as:

$$\mathcal{BF}\mathcal{E}(y) = \frac{1}{2(1 - 1/K)} \sum_{k=1}^K \left| y_k - \frac{1}{K} \right|, \quad (4)$$

which scales BFE strictly to $[0, 1]$, where 0 indicates a perfectly uniform (unbiased) prediction. For a set of M bias-type subsets (e.g., language bias, visual bias, cross-modal bias), the overall $\mathcal{BF}\mathcal{E}$ is:

$$\mathcal{BF}\mathcal{E}_{\text{overall}} = 1 - \frac{\sigma(\{\text{Acc}_m\}_{m=1}^M)}{\mu(\{\text{Acc}_m\}_{m=1}^M)}, \quad (5)$$

where Acc_m is the accuracy on bias subset m , $\mu(\cdot)$ is the mean, and $\sigma(\cdot)$ is the standard deviation. $\mathcal{BF}\mathcal{E}_{\text{overall}} \in [0, 1]$, with 1 indicating perfect fairness.

3.5. Label Construction for PTG Samples

Generated fuzzing samples (v'_i, q'_i) do not have ground-truth labels. We therefore generate surrogate labels by inverting the original model distribution. This section provides the theoretical justification and formal definition of our label construction strategy.

3.5.1. Theoretical Justification

The transformation $a'_i = |1 - y_i|$ can be understood from multiple theoretical perspectives:

Inverse Probability Weighting (IPW). In causal inference, IPW assigns higher weights to underrepresented samples to correct for selection bias. Here, y_i represents the model's confidence over answers, and $1 - y_i$ assigns higher weight to answers the model currently neglects, those likely suppressed by language priors. This creates a contrastive signal that exposes bias-driven prediction patterns.

Information-Theoretic View. Let $H(P_\theta)$ denote the entropy of the model's prediction distribution. When $H(P_\theta)$ is low (high confidence), the model may be relying on shortcut features. The inverse label construction $a'_i = |1 - y_i|$ increases the effective information gain by forcing the model to reconsider answers it had prematurely ruled out.

Multi-Class Generalization. For VQA with K candidate answers, we generalize the transformation to:

$$a'_i = \text{softmax}(-\log P_\theta(a|v_i, q_i)/T), \quad (6)$$

where T is a temperature parameter. As $T \rightarrow 0^+$, this reduces to the binary $|1 - y_i|$ form. This formulation assigns a principled probabilistic interpretation: a'_i defines a distribution complementary to the original, with the softmax ensuring valid probability simplex constraints.

Confidence Gating. To mitigate supervisory noise when the original model's predictions are uncertain, we introduce a confidence gating mechanism:

$$a'_i = \begin{cases} \text{softmax}(-\log P_\theta(a|v_i, q_i)/T), & \text{if } \max_a P_\theta(a|v_i, q_i) \geq \theta_{\text{conf}} \\ y_i, & \text{otherwise,} \end{cases} \quad (7)$$

with $\theta_{\text{conf}} = 0.6$. This ensures label inversion is applied only when the model has sufficient confidence, preventing noise amplification from uncertain predictions.

To further rigidify the theoretical foundation of this design, we explicitly analyze its boundary conditions and potential limitations. Since the inverted target a'_i directly derives from the model's instantaneous execution state y_i , it inherently creates a closed-loop coupling system. A theoretical vulnerability arises if the model encounters highly ambiguous out-of-distribution (OOD) samples where the predictive logits are dominated by epistemic uncertainty rather than systematic bias. In such scenarios, raw inversion could inadvertently penalize legitimate exploratory predictions, leading to gradient variance explosion or standard empirical risk degradation.

By establishing the confidence gating mechanism (Equation (7)), we analytically define the robust operating boundaries of our framework. When $\max_k P_\theta(a_k|v_i, q_i) < \theta_{\text{conf}}$, the framework gracefully retreats to the passive multi-class imitation phase, matching the empirical data distribution without intervention. This ensures that the dynamic error-correction signal is selectively dispatched only to areas of the input space where the model exhibits pathological over-confidence triggered by linguistic or visual shortcuts.

3.5.2. Differential Bias Discrimination Theorem

To rigorously address the concern that high BFE values might stem from valid visual causal reasoning rather than harmful language shortcuts, we formally prove that the differential sensitivity index ΔC_i —not the absolute BFE—provides a valid discrimination signal.

Proposition 1 (Differential Bias Discrimination). Let $I_\theta(Y; V|Q)$ denote the conditional mutual information between the answer Y and the visual input V given the question Q , serving as a conceptual measure of the model's visual reasoning capacity. For a PTG-generated fuzzed probe (v'_i, q'_i) that removes or weakens task-critical visual evidence while preserving semantic structure, and under the confidence gating condition $\max_k P_\theta(a_k|v_i, q_i) \geq \theta_{\text{conf}}$:

1. If $I_\theta(Y; V|Q) > \varepsilon$ (genuine visual reasoning): $\Delta C_i < 0$ (correction not triggered).
2. If $I_\theta(Y; V|Q) \leq \varepsilon$ (language shortcut reliance): $\Delta C_i \geq 0$ (correction correctly triggered).

Proof. (1) When $I_\theta(Y; V|Q) > \varepsilon$, the model's posterior genuinely depends on V . The PTG probe (v'_i, q'_i) is designed to remove or weaken task-critical visual evidence (Section 3.2). We empirically validate this property via target-region feature similarity analysis (Section 4.8.7): the cosine similarity between GradCAM-localized answer-relevant features on (v_i, q_i) and (v'_i, q'_i) drops significantly (from 0.82 ± 0.11 to 0.34 ± 0.15 , $p < 0.001$ via paired t -test), confirming that PTG probes indeed attenuate visual support for the original answer. Under this condition, let $p_{k^*} = \max_k P_\theta(a_k|v_i, q_i)$ be the probability of the dominant answer on the original sample, and $p'_{k^*} = P_\theta(a_{k^*}|v'_i, q'_i)$ be the corresponding probability on the fuzzed probe. Since the visual support for k^* is weakened by PTG, $p'_{k^*} < p_{k^*}$. The reduction in the dominant probability mass causes the prediction distribution to spread toward uniformity, i.e.,

$$D_{\text{TV}}(P_\theta(\cdot|v'_i, q'_i), U_K) < D_{\text{TV}}(P_\theta(\cdot|v_i, q_i), U_K), \quad (8)$$

where D_{TV} denotes total variation distance, and U_K is the uniform distribution over K classes. Since $\mathcal{BFE}(y) = D_{TV}(y, U_K)$ by definition (Equation (4)), we have $\mathcal{BFE}(y'_i) < \mathcal{BFE}(y_i)$, hence $\Delta\mathcal{C}_i < 0$.

(2) When $I_\theta(Y; V|Q) \leq \varepsilon$, $P_\theta(Y|V, Q) \approx P_\theta(Y|Q)$, so the prediction is approximately invariant to V . Replacing v_i with v'_i yields $P_\theta(Y|v'_i, q'_i) \approx P_\theta(Y|v_i, q_i)$, hence $\mathcal{BFE}(y'_i) \approx \mathcal{BFE}(y_i)$ and $\Delta\mathcal{C}_i \approx 0 \geq 0$ (within tolerance margin τ_{margin}). $\square$

Remark 1 (On the conceptual role of I_θ). We emphasize that $I_\theta(Y; V|Q)$ serves as a conceptual device for organizing the proof structure rather than a directly observable quantity in practice. Computing conditional mutual information exactly from finite data is generally infeasible; moreover, I_θ depends on the model parameters θ which evolve during training. Therefore, Proposition 1 provides conceptual justification for why the differential BFE $\Delta\mathcal{C}_i$ should discriminate correctly in principle, while the FPR experiment (Section 4.8.7) provides empirical validation that it does so in practice. These two forms of evidence—conceptual and empirical—are complementary: the former explains why the mechanism should work; the latter confirms that it does work on real data.

Remark 2 (Summary). Proposition 1 establishes that the differential BFE $\Delta\mathcal{C}_i$, not the absolute BFE, serves as the discrimination criterion. This resolves the apparent directional conflict: a model with genuine visual reasoning exhibits decreased prediction concentration (lower BFE) on fuzzed probes due to the attenuation of visual evidence, while a model relying on language shortcuts maintains its concentrated prediction regardless of visual perturbation. The confidence gating mechanism further ensures that this discrimination operates only on predictions with sufficient model confidence ($\geq \theta_{\text{conf}}$), preventing noise amplification on uncertain samples. Empirical validation via the false positive rate (FPR) experiment with attention-based cross-validation is provided in Section 4.8.

A comprehensive ablation study comparing five label construction strategies is presented in Section 4.8.6 (Table 1), confirming that the IPW strategy with confidence gating achieves the best bias–fairness trade-off.

Table 1. Ablation on label construction strategies for fuzzing samples. Results on VQA-CP v2 test set. $\downarrow$ indicates lower values correspond to better bias mitigation. Rows (a)–(e) represent five alternative label construction strategies.

Label Strategy	Accuracy (%)	BFE $\downarrow$
(a) $a'_i = 1 - y_i $ (original)	62.18 $\pm$ 0.71	0.232 $\pm$ 0.014
(b) $a'_i = 1 - y_i$ (linear, no abs)	61.85 $\pm$ 0.82	0.241 $\pm$ 0.016
(c) Random label flip (50%)	59.34 $\pm$ 1.52	0.298 $\pm$ 0.025
(d) IPW + confidence gate	63.45 $\pm$ 0.48	0.188 $\pm$ 0.009
(e) Entropy-weighted soft labels	63.21 $\pm$ 0.56	0.195 $\pm$ 0.011

3.6. Reflection Mechanism

Once we obtain predictions for original and fuzzed samples (y_i, y'_i) , we compute the bias fairness evaluation:

$$\mathcal{C}^1 = \mathcal{BFE}(y_i), \quad \mathcal{C}^2 = \mathcal{BFE}(y'_i).$$

Their difference measures the sensitivity of the model to bias when exposed to unexpected inputs:

$$\Delta\mathcal{C}_i = \mathcal{C}^2 - \mathcal{C}^1.$$

We then apply a corrective reflection rule:

$$\hat{y}_i = \begin{cases} \delta(y_i), & \Delta C_i \geq 0 \\ y_i, & \Delta C_i < 0. \end{cases} \quad (9)$$

This step mimics a human-like self-correction process: if the fuzzed sample triggers higher bias, the original prediction is adjusted toward less biased alternatives.

The correction function $\delta(y_i)$ —termed the Answer Redistribution Function—is defined through the following computational steps:

1. **Temperature Scaling.** Scale the original prediction logits to sharpen the distribution:

$$\text{logits}'_a = \text{logits}_a / \tau, \quad \tau = 0.5, \quad (10)$$

where $\tau < 1$ produces a sharper distribution, helping to identify the model's dominant answer preferences.

2. **Softmax Normalization.** Obtain the scaled probability distribution:

$$P'_\theta(a|v_i, q_i) = \text{softmax}(\text{logits}'). \quad (11)$$

3. **Inverse Distribution Construction.** Build a target distribution complementary to the original:

$$P_{\text{target}}(a) = \text{softmax}(-\log P'_\theta(a|v_i, q_i)). \quad (12)$$

Intuitively, answers the model is highly confident in receive low target probability, while answers the model neglects receive high target probability.

4. **Weighted Fusion.** For training stability, blend the inverse distribution with the original:

$$P_{\text{final}}(a) = (1 - \alpha) \cdot P'_\theta(a|v_i, q_i) + \alpha \cdot P_{\text{target}}(a), \quad (13)$$

where $\alpha = \min(C_i / \tau_{\text{bias}}, \alpha_{\text{max}})$, $\alpha_{\text{max}} = 0.8$, and τ_{bias} is a bias threshold. A higher bias score C_i results in a larger α , giving more weight to the inverse distribution.

5. **Hard Label Extraction.** Extract the corrected prediction:

$$\hat{y}_i = \arg \max_a P_{\text{final}}(a). \quad (14)$$

Algorithm 1 presents the complete reflection correction procedure in pseudocode.

Algorithm 1 Pseudocode of the reflection correction mechanism.

Require: Original sample (v_i, q_i) , fuzzed sample (v'_i, q'_i) , model P_θ , $\tau = 0.5$, $\alpha_{\text{max}} = 0.8$

Ensure: Corrected prediction $\hat{y}_i$

```

1:  $y_i \leftarrow P_\theta(a|v_i, q_i)$  // Original prediction
2:  $y'_i \leftarrow P_\theta(a|v'_i, q'_i)$  // Fuzzed prediction
3:  $C_i \leftarrow \text{BFE}(y_i); C'_i \leftarrow \text{BFE}(y'_i)$ 
4:  $\Delta C_i \leftarrow C'_i - C_i$ 
5: if  $\Delta C_i \geq 0$  then
6:    $\text{logits}' \leftarrow \text{logits}(y_i) / \tau$ 
7:    $P' \leftarrow \text{softmax}(\text{logits}')$ 
8:    $P_{\text{target}} \leftarrow \text{softmax}(-\log P')$ 
9:    $\alpha \leftarrow \min(C_i / \tau_{\text{bias}}, \alpha_{\text{max}})$ 
10:   $P_{\text{final}} \leftarrow (1 - \alpha) \cdot P' + \alpha \cdot P_{\text{target}}$ 
11:   $\hat{y}_i \leftarrow \arg \max_a P_{\text{final}}(a)$ 
12: else
13:   $\hat{y}_i \leftarrow y_i$  // No correction needed
14: end if
15: return  $\hat{y}_i$ 

```

3.6.1. Gradient Mechanics and Non-Chaotic Convergence Proof

To formally address the concern regarding whether fitting an inverted distribution induces chaotic outputs or deforms empirical performance on standard samples, we examine the gradient properties of the bias-aware loss $\mathcal{L}_b$. Let z_k represent the model's output logit for answer candidate k . The gradient of the localized cross-entropy component $\mathcal{L}_{CE}(y, \hat{y})$ with respect to z_k is analytically expressed as:

$$\frac{\partial \mathcal{L}_{CE}(y, \hat{y})}{\partial z_k} = y_k - \hat{y}_k$$

Substituting the weighted fusion target $P_{\text{final}}(a_k)$ from Equation (12) into $\hat{y}_k$ yields:

$$\frac{\partial \mathcal{L}_{CE}}{\partial z_k} = y_k - [(1 - \alpha)y'_k + \alpha \cdot \text{softmax}(-\log y'_k)]$$

When the model triggers a severe shortcut bias on a fuzzed sample, the scaled probability $y'_k \rightarrow 1$, driving the dynamic target $\hat{y}_k \rightarrow 0$. Under this boundary condition, the gradient simplifies to $\frac{\partial \mathcal{L}_{CE}}{\partial z_k} \approx y_k > 0$, generating an aggressive positive penalty gradient that systematically depresses the parameter vectors aligned with that biased shortcut. Crucially, because $\hat{y}_k$ is a bounded, convex combination governed by $\alpha = \min(C_i / \tau_{\text{bias}}, \alpha_{\text{max}})$, the objective does not enforce a chaotic white-noise assignment. Instead, it behaves as a directional repulsive boundary. Furthermore, the confidence gating mechanism (Equation (3)) guarantees that for standard samples devoid of systematic over-confidence shortcuts, $\Delta C_i < 0$, and $\mathcal{L}_b$ identically vanishes to 0. This mathematical bifurcation strictly decouples standard empirical risk minimization from bias suppression, eliminating optimization chaos.

3.6.2. Control-Theoretic Convergence and Dissipative Analysis

To address the stability of our closed-loop feedback training process, we formulate the optimization trajectory as a discrete-time dynamical system perturbed by the PTG actuator. Let $\theta \in \mathbb{R}^D$ be the system state vector (model parameters). The update equation is governed by:

$$\theta_{t+1} = \theta_t - \eta (\nabla_{\theta} \mathcal{L}_{\text{VQA}}(\theta_t) + \lambda_1 u(\theta_t, w_t) + \lambda_2 \nabla_{\theta} \mathcal{L}_{\text{sim}}(\theta_t)), \quad (15)$$

where w_t is the unexpected perturbing input generated by PTG at step t , and $u(\theta_t, w_t) = \mathbb{I}(\Delta C_i \geq 0) \cdot \mathbb{I}(\max P_{\theta} \geq \theta_{\text{conf}}) \cdot \nabla_{\theta} \mathcal{L}_{CE}(y_i, \hat{y}_i)$ represents the feedback control signal dispatched by the reflection mechanism.

We define the energy state of the system using a candidate Lyapunov function $V(\theta) = \mathcal{L}_{\text{total}}(\theta)$. Since $\mathcal{L}_{\text{total}}$ is lower-bounded by zero, $V(\theta) \geq 0$. Under the L-smoothness assumption of the objective function, the system energy difference is bounded by:

$$V(\theta_{t+1}) - V(\theta_t) \leq -\eta \left(1 - \frac{\eta L}{2}\right) \|\nabla_{\theta} \mathcal{L}_{\text{total}}(\theta_t)\|^2 + \mathcal{O}(\eta^2 \|u(\theta_t, w_t)\|^2). \quad (16)$$

When the step size η is chosen such that $\eta < 2/L$ (ensuring $1 - \eta L/2 > 0$) and the perturbation magnitude satisfies $\|u(\theta_t, w_t)\|^2 \leq \frac{(1 - \eta L/2)}{\eta L} \|\nabla_{\theta} \mathcal{L}_{\text{total}}(\theta_t)\|^2$ —a condition enforced by the confidence gating mechanism, which zeroes out u when the model's prediction lacks pathological over-confidence—the perturbation term is dominated by the negative descent term, yielding $V(\theta_{t+1}) - V(\theta_t) \leq 0$. This establishes the Lyapunov descent condition, conditionally ensuring bounded energy when the perturbation bound is satisfied, an invariant that is actively enforced through the λ_1 warmup schedule described below and empirically verified through R_t monitoring (Section 4.8.8).

λ_1 Warmup Schedule. A critical vulnerability of the Lyapunov condition arises during the early stages of training, when the main loss gradient norm $\|\nabla_{\theta} \mathcal{L}_{\text{VQA}}(\theta_t)\|$ is small and

PTG-generated perturbations $\|\mathbf{u}(\theta_t, \mathbf{w}_t)\|$ may be relatively large due to the stochastic nature of the generative model. To ensure the Lyapunov condition holds throughout this vulnerable phase, we introduce a warmup schedule for the bias-aware loss weight λ_1 :

$$\lambda_1(t) = \min\left(\frac{t}{T_{\text{warm}}}, 1\right) \cdot \lambda_1^{\text{target}}, \quad (17)$$

where $T_{\text{warm}} = 5$ epochs, and $\lambda_1^{\text{target}} = 0.5$. During the warmup phase, λ_1 linearly ramps from 0 to its target value. Since the perturbation contribution to the parameter update is scaled by $\lambda_1(t)$ (Equation (17)), a small $\lambda_1(t)$ in early training proportionally suppresses the perturbation term, maintaining the ratio $\|\mathbf{u}\|^2 / \|\nabla_{\theta} \mathcal{L}_{\text{total}}\|^2$ below the Lyapunov threshold even when $\|\nabla_{\theta} \mathcal{L}_{\text{VQA}}\|$ is small. This schedule is specified a priori before training begins; no mid-training parameter adaptation is performed, ensuring consistency with the Lyapunov analysis which assumes fixed hyperparameters.

Lyapunov Condition Ratio Monitoring (Diagnostic). For post hoc empirical verification that the warmup schedule achieves its design objective, we define the Lyapunov condition ratio at step t as:

$$R_t = \frac{\eta^2 \|\mathbf{u}(\theta_t, \mathbf{w}_t)\|^2}{\eta \left(1 - \frac{\eta^L}{2}\right) \|\nabla_{\theta} \mathcal{L}_{\text{total}}(\theta_t)\|^2}, \quad (18)$$

where $R_t < 1$ is equivalent to the Lyapunov descent condition. Note that R_t is computed purely as a monitoring diagnostic during or after training; it does not trigger any parameter modification. If monitoring reveals R_t approaching or exceeding 1 at any point, this would indicate that the warmup schedule requires redesign (e.g., longer T_{warm} or smaller $\lambda_1^{\text{target}}$), but such modification would necessitate re-running training with the revised schedule, not mid-training adjustment. Empirical verification that $R_t < 1$ holds throughout training is presented in Section 4.8.8.

Crucially, we bound the perturbation input using the semantic similarity constraint $\mathcal{L}_{\text{q-sim}}$, which enforces:

$$\|\mathbf{e}_{q_i} - \mathbf{e}_{q'_i}\| \leq \sqrt{2(1 - \gamma)}, \quad (19)$$

where $\gamma = 0.7$ acts as a guardrail. This guarantees that $\mathbf{w}_t$ resides within a compact, localized subset surrounding the training manifold $\mathcal{M}$. Similarly, the visual perturbation is bounded by the StyleGAN2-ADA generator's training distribution: since v'_i is synthesized to be compatible with the Faster R-CNN feature space (36×2048 region features), its deviation from the original v_i is constrained by the generator's learned manifold. In practice, we enforce a visual similarity threshold via CLIP score ≥ 0.2 (Section 4.1.2), ensuring $\|v_i - v'_i\|$ remains within a bounded neighborhood of the original feature manifold. Combined, both textual and visual perturbations satisfy BIBO (Bounded-Input Bounded-Output) stability conditions, preventing chaotic divergence. We then analyze the dissipativity of the feedback controller $\mathbf{u}$. Recall that $\mathbf{u} = \mathbb{I}(\Delta \mathcal{C}_i \geq 0) \cdot \mathbb{I}(\max P_{\theta} \geq \theta_{\text{conf}}) \cdot \nabla_{\theta} \mathcal{L}_{\text{CE}}(y_i, \hat{y}_i)$. When the indicator functions are active (i.e., the reflection mechanism is triggered), the corrected target $\hat{y}_i$ is constructed as $\hat{y}_i = \arg \max_a P_{\text{final}}(a)$, where P_{final} is a convex blend of the sharpened prediction P'_{θ} and the inverse target P_{target} (Equation (12)). Because $P_{\text{target}} = \text{softmax}(-\log P'_{\theta})$ assigns high probability to answers the model currently suppresses, the corrected label $\hat{y}_i$ shifts probability mass toward underrepresented answers. For the dominant biased answer k^* (where $y_{k^*} \rightarrow 1$), the corrected target satisfies $\hat{y}_{k^*} < y_{k^*}$, making the per-sample CE gradient $\frac{\partial \mathcal{L}_{\text{CE}}}{\partial z_{k^*}} = y_{k^*} - \hat{y}_{k^*} > 0$. This gradient penalizes the same biased logit direction that $\mathcal{L}_{\text{VQA}}$ would penalize if the ground-truth answer differed from k^* (which is precisely the

case when $\Delta C_i \geq 0$, as the model is over-confident on a wrong shortcut answer). Therefore, under the bias-triggered condition, the gradient inner product satisfies:

$$\langle \nabla_{\theta} \mathcal{L}_{\text{VQA}}(\theta_t), \mathbf{u}(\theta_t, \mathbf{w}_t) \rangle \geq 0, \quad (20)$$

meaning the correction signal reinforces, rather than counteracts, the standard empirical risk minimization direction on bias-dominated samples.

3.7. Training Objective

We optimize the model using a joint loss function that combines the standard VQA objective with bias-aware and semantic similarity terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{VQA}} + \lambda_1 \mathcal{L}_b + \lambda_2 \mathcal{L}_{\text{sim}}, \quad (21)$$

where each component is defined as follows.

VQA Loss. To supervise the multi-class answer prediction, we employ the conventional binary cross-entropy loss, denoted as $\mathcal{L}_{\text{VQA}}$:

$$\mathcal{L}_{\text{VQA}} = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K a_{i,k} \log y_{i,k}, \quad (22)$$

where $a_{i,k} \in \{0, 1\}$ indicates whether answer k is the ground truth for sample i .

Bias-Aware Loss. The bias-aware loss $\mathcal{L}_b$ penalizes predictions that exhibit high bias as measured by the BFE metric. It is formally defined as:

$$\mathcal{L}_b = \frac{1}{N} \sum_{i=1}^N \max(0, \Delta C_i - \tau_{\text{margin}}) \cdot \mathcal{L}_{\text{CE}}(y_i, \hat{y}_i), \quad (23)$$

where $\Delta C_i = C'_i - C_i$ is the bias sensitivity (Equation (1) in the main text), $\tau_{\text{margin}} = 0.1$ is a tolerance margin that allows minor bias fluctuations without penalty, $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss between the original and corrected predictions, and the $\max(0, \cdot)$ operator ensures the bias penalty is applied only when the fuzzed sample triggers bias exceeding the margin.

Semantic Similarity Loss (see Section 3.8 for details). The hyperparameters λ_1 and λ_2 control the relative importance of bias correction and semantic alignment. Based on a grid search over the validation set (search ranges: $\lambda_1 \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$, $\lambda_2 \in \{0.5, 1.0, 1.5, 2.0\}$), we set $\lambda_1 = 0.5$ and $\lambda_2 = 1.0$, which achieves the optimal balance between debiasing effectiveness and training stability. A comprehensive sensitivity analysis is provided in Section 4.8.8.

Important Note. All fuzzing-related components, including PTG, bias detection, and the reflection mechanism, are exclusively active during the training phase. At inference time, only the prompt-based VQA model with debiased parameters is used, without any additional computational overhead.

3.8. Semantic Similarity Constraint

To bridge the semantic gap between model-generated answers and ground-truth labels, we introduce a semantic similarity constraint. Using Sentence-BERT (all-mpnet-base-v2) [46], both the generated answer text and the ground-truth answer are encoded into 768-dimensional sentence embeddings $\mathbf{e}_{\text{gen}}$ and $\mathbf{e}_{\text{gt}}$. The semantic similarity loss is defined as:

$$\mathcal{L}_{\text{sim}} = 1 - \cos(\mathbf{e}_{\text{gen}}, \mathbf{e}_{\text{gt}}) = 1 - \frac{\mathbf{e}_{\text{gen}} \cdot \mathbf{e}_{\text{gt}}}{\|\mathbf{e}_{\text{gen}}\| \cdot \|\mathbf{e}_{\text{gt}}\|}. \quad (24)$$

This loss explicitly minimizes the cosine distance between generated and reference answer embeddings during training, encouraging the model to produce answers that are semantically aligned with human annotations rather than merely matching surface-form labels. For the fuzzing question generation, we additionally apply a similarity constraint between the original question q_i and the generated question q'_i :

$$\mathcal{L}_{q\text{-sim}} = \frac{1}{N} \sum_{i=1}^N \max(0, \gamma - \cos(\mathbf{e}_{q_i}, \mathbf{e}_{q'_i})), \quad (25)$$

where $\gamma = 0.7$ is a similarity threshold that prevents the generation of semantically unrelated questions while allowing sufficient deviation to challenge the model.

3.9. Visual Hint Prompts

Visual hint prompts are concise textual cues embedded in the question prefix that provide information about the image content, guiding the PTG module to generate questions semantically consistent with the visual context. Specifically, we leverage a pretrained CLIP model [43] to extract top-3 semantic labels from the input image, obtaining their text embeddings. These embeddings are concatenated with the prompt prefix to form a visually conditioned generation condition:

$$\mathbf{h}_{\text{prefix}} = [\mathbf{e}_{\text{CLIP}}(l_1); \mathbf{e}_{\text{CLIP}}(l_2); \mathbf{e}_{\text{CLIP}}(l_3); \mathbf{e}_{\text{prompt}}], \quad (26)$$

where l_1, l_2, l_3 are the top-3 predicted object/attribute labels for the image, and $\mathbf{e}_{\text{prompt}}$ is the base prompt embedding. We explore five hint strategies in our ablation experiments: (a) no hint (baseline), (b) category hint (“This image contains [object]”), (c) attribute hint (“This image shows [attribute]”), (d) relation hint (“The [object1] is [relation] [object2]”), and (e) combined hint (concatenation of the above three). A detailed comparison and analysis of these strategies are provided in Section 4.8.

4. Experiments

This section assesses the performance of our prompt-driven debiasing framework across several benchmark VQA datasets. We begin by outlining the experimental configuration—encompassing benchmarks, implementation specifics, and baseline models—before providing a meticulous comparison against state-of-the-art debiasing methods. Lastly, detailed ablation experiments are carried out to dissect the impact of each core component, namely, prompt templates, visual hint prompts, similarity-aware optimization, and backbone generalization.

4.1. Implementation Details

All experiments were conducted with the following configuration.

Backbone Architecture. The primary visual backbone adopted by FD was UpDn (Bottom-Up Top-Down Attention) [31]. To extract visual features, a Faster R-CNN model (backboned by ResNet-101 and pretrained on Visual Genome) was utilized, yielding 36 region proposals per image, where each region was represented by a 2048-dimensional vector. For textual processing, tokens were mapped to 300-dimensional GloVe word embeddings and subsequently processed via a single-layer LSTM configured with 1280 hidden units.

Hardware and Software. All empirical evaluations were conducted on a computing node equipped with four NVIDIA RTX A6000 GPUs (48 GB of VRAM per card), an Intel Xeon Gold 6338 CPU (32 cores), and 256GB of system memory. Our codebase was developed utilizing PyTorch 1.12.

Training Configuration. The model was optimized for 40 epochs on VQA-CP v2 and 50 epochs on VQA-CE, utilizing batch sizes of 256 and 128, respectively. Optimization was performed via the AdamW optimizer with hyperparameters set to $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay coefficient of 10^{-4} . The learning rate was initialized at 10^{-4} and decayed following a cosine annealing schedule down to a minimum of 10^{-6} , preceded by a linear warmup phase during the initial three epochs. Additionally, the bias-aware loss weight λ_1 employed a separate five-epoch linear warmup from zero to its target value of 0.5 (Equation (17)), ensuring that the reflection mechanism's perturbation was gradually introduced during the critical early training phase when gradient norms were small (see Section 3.6.2 for theoretical justification). To prevent gradient explosion, gradients were clipped at a max_norm of 1.0. Furthermore, an early stopping mechanism was employed: training terminated if the validation VQA accuracy stagnated for 10 consecutive epochs, whereupon the optimal model parameters were retrieved.

Hyperparameter Selection. Hyperparameters $\lambda_1 \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$, $\lambda_2 \in \{0.5, 1.0, 1.5, 2.0\}$, bias threshold $\tau \in \{0.3, 0.5, 0.7\}$, tolerance margin $\tau_{\text{margin}} \in \{0.05, 0.1, 0.2\}$, and temperature $\tau_{\text{temp}} \in \{0.3, 0.5, 0.7\}$ were selected via grid search on the validation set. The final values were $\lambda_1 = 0.5$, $\lambda_2 = 1.0$, $\tau = 0.5$, $\tau_{\text{margin}} = 0.1$, $\tau_{\text{temp}} = 0.5$.

Reproducibility. To ensure robustness, all reported metrics represent the mean $\pm$ standard deviation calculated across five independent experimental runs using distinct random seeds ($\{42, 123, 456, 789, 1024\}$). We evaluated statistical significance via two-tailed paired t-tests, incorporating Bonferroni correction to adjust for multiple comparisons. Furthermore, Cohen's d was computed to quantify effect sizes.

PTG Configuration. The PTG module regenerated fuzzing samples at the beginning of each epoch.

4.1.1. Quantitative Verification of PTG Sample Quality

To verify that the generated PTG test cases constituted a controlled, intelligent form of fuzz testing rather than introducing amorphous noise, we conducted a rigorous quantitative evaluation on the generated questions and features:

- **Linguistic Fluency (Perplexity):** We evaluated the generated fuzzed questions q'_i using a pretrained GPT-2. The average perplexity was 32.45 ± 2.15 , well below the safety threshold of 50.0, indicating high fluency and grammatical correctness.
- **Semantic Plausibility (S-BERT Cosine Similarity):** The cosine similarity between the embeddings of the original question q_i and the fuzzed question q'_i was measured using Sentence-BERT. The similarity values concentrated in the range of $[0.62, 0.78]$ with a mean of 0.68, validating that the generated probes remained topically relevant and semantically grounded in the image context.
- **Distributional Unexpectedness (JSD):** We measured the Jensen–Shannon Divergence (JSD) between the sentence embeddings of the fuzzed dataset and the original training set. The divergence score yielded $\text{JSD} = 0.42$, notably higher than the JSD between two disjoint random subsets of the original training set ($\text{JSD}_{\text{baseline}} = 0.08$), confirming that the fuzzed samples represented a distinct, off-manifold out-of-distribution shift rather than random subsampling variation. At the same time, the divergence remained well below the theoretical maximum of 1.0 (completely disjoint distributions), satisfying the “unexpectedness” criterion needed for fuzz testing without drifting into white noise.

Furthermore, human evaluation of 200 random samples on a five-point Likert scale (where 5 represents excellent syntax and visual grounding) yielded an average score of 4.42 ± 0.38 with substantial inter-annotator agreement (Fleiss' $\kappa = 0.78$), verifying the high quality of the generated probes.

4.1.2. Detailed PTG Generator Implementation

Question Generator Architecture. We employed GPT-2-medium (345M parameters) as the backbone for unexpected question generation. The architecture consisted of 12 Transformer decoder layers with 1024-dimensional hidden states and 16 attention heads. The input was formed by concatenating a learned linear projection of visual features with a prompt prefix, and the model generated text autoregressively conditioned on this multimodal prefix. Fine-tuning was performed on the VQA v2 training set (approximately 440K questions) using LoRA (Low-Rank Adaptation) with rank $r = 16$ to reduce trainable parameters and prevent overfitting.

Visual Feature Generator. For visual fuzzing sample generation, we adopted a StyleGAN2-ADA-based generator trained to synthesize 36×2048 region feature matrices compatible with the Faster R-CNN feature extractor output format. The generator was trained on VQA v2 image features extracted by a pretrained Faster R-CNN with ResNet-101 backbone.

Training Objective and Specifications. The question generator was optimized using a combined objective:

$$\mathcal{L}_{\text{PTG}} = \mathcal{L}_{\text{CLM}} + \alpha \cdot \mathcal{L}_{\text{div}}, \quad (27)$$

where $\mathcal{L}_{\text{CLM}} = -\sum_t \log P(w_t | w_{<t}, v)$ is the causal language modeling loss, w_t is the t th token, v represents visual feature conditions, and $\mathcal{L}_{\text{div}} = -H(P(w|v))$ is a diversity regularization term (negative entropy of the generated word distribution) that encourages diverse question generation. The balancing coefficient α was set to 0.1. Specifically, the GPT-2 backbone was fine-tuned for five epochs using the AdamW optimizer with a learning rate of 2×10^{-4} , a linear warmup of 0.1 epochs, and a batch size of 64. For the visual feature generator, StyleGAN2-ADA was optimized using Adam with a learning rate of 2×10^{-3} , $\beta_1 = 0$, $\beta_2 = 0.99$ and trained for 25,000 iterations. During the online fuzzing phrase, unexpected questions were generated utilizing Nucleus Sampling with a temperature of 0.7, top- $p = 0.9$, and a repetition penalty of 1.2 to systematically enforce linguistic diversity while preventing degenerate loops.

Quality Metrics. We evaluated generated samples using three metrics: (1) **Perplexity** (target < 50) to assess linguistic fluency; (2) **BLEU-4** score against original VQA v2 validation questions (target range of 0.3–0.7 to ensure the generated question differed from the original while maintaining relevance); (3) **Semantic similarity** evaluated using Sentence-BERT (all-mpnet-base-v2) cosine similarity, where a target range of 0.5–0.8 ensured the questions were “unexpected” yet topically relevant.

Semantic Consistency Verification. We employed a two-stage verification mechanism. First, automatic filtering removed generated question–image pairs with CLIP score below 0.2 (indicating poor visual–textual alignment). Second, as detailed in Section 4.1.1, human evaluation on 200 randomly sampled pairs yielded an average score of 4.42 ± 0.38 (five-point Likert scale; Fleiss’ $\kappa = 0.78$). Only samples with an average score ≥ 4.0 were retained. The rubric was defined as: 5 (Excellent): perfectly coherent, seamlessly matches visual regions; 4 (Good): fully intelligible and contextually relevant; 3 (Fair): structurally understandable but loosely coupled with the image; 2 (Poor): semantic hallucinations or severe grammatical errors; 1 (Unusable): completely garbled or cross-modal contradiction.

4.2. Datasets

We primarily conducted experiments on commonly used VQA datasets, which are typically employed to evaluate the debiasing ability of VQA models. We validated the debiasing capability of our method on VQA-CP v1 and VQA-CP v2 [24], assessed the model’s ability to handle various types of biases on GQA-OOD [47], VQA-CE [23], and VQA-VS [8], and evaluated the model’s generalization ability on the VQA v2 [5] dataset.

4.3. Comparison with State-of-the-Art Methods

4.3.1. Results on VQA-CP and VQA v2

We benchmarked our method against a wide range of representative VQA debiasing techniques, as summarized in Table 2. Existing approaches can generally be grouped into several categories: (1) *ensemble-based strategies*, which explicitly model bias branches or auxiliary predictors (e.g., AdvReg [13], RUBi [10], LMH [11], GGE [9], GenB [12]); (2) *data augmentation approaches*, which aim to modify or enrich training data distributions (CSS [15], ASL [18], MUTANT [14], Unshuffling [17], KDDAug [16]); (3) *self-supervised or contrastive-learning-based methods*, which strengthen multimodal representations by enforcing modality-level consistency (CSS + CL [21], SSL [20], MMBS [22], LSP [19]); and (4) *answer re-ranking methods*, which reorder candidate answers to emphasize visual evidence (GVQA [24], HINT [23], SCR [48], DLR [26], RMLVQA [25]).

Table 2. The results on VQA-CP v2, VQA-CP v1 test set and VQA v2 validation set are presented in the following table. All versions compared used UpDn as the base network.

Data Set	Base	CP v2 Test				V2 Val				CP v1
Method		All	Y/N	Num.	Other	All	Y/N	Num.	Other	All
UpDn [31]	-	39.96	43.01	12.07	45.82	63.48	81.18	42.14	55.66	36.38
AdvReg [13]	UpDn	41.17	65.49	15.48	35.48	62.75	79.84	42.35	55.16	43.43
RUBi [10]	UpDn	44.23	67.05	17.48	39.61	-	-	-	-	50.90
LMH [11]	UpDn	52.01	72.58	31.12	46.97	56.35	65.06	37.63	54.69	55.27
AdaVQA [49]	UpDn	54.67	72.47	53.81	45.58	-	-	-	-	61.20
IntroD [50]	CSS	60.17	89.17	46.91	48.62	62.57	78.57	41.42	56.00	-
CF-VQA [29]	UpDn	53.55	91.15	13.03	44.97	63.54	82.51	43.96	54.30	57.03
GGE [9]	UpDn	57.32	87.04	27.75	49.59	59.11	73.27	39.99	54.39	-
D-VQA [7]	UpDn	61.91	88.93	52.32	50.39	64.96	82.18	44.05	57.54	-
GGD [27]	UpDn	59.37	88.23	38.11	49.82	62.15	79.25	42.43	54.66	-
GenB [12]	UpDn	59.15	88.03	40.05	49.25	62.74	86.18	43.85	47.03	62.74
ASL [18]	UpDn	46.00	58.24	29.49	44.33	-	-	-	-	-
CSS [15]	UpDn	58.95	84.37	49.42	48.24	59.91	73.25	39.77	55.11	60.95
Mutant [14]	UpDn	61.72	88.90	49.68	50.78	62.56	82.07	42.52	53.28	-
Unshuffling [17]	UpDn	43.47	47.82	14.35	49.18	43.47	81.99	43.07	55.21	-
SimpleAug [51]	UpDn	52.65	66.40	43.43	47.98	64.34	81.97	43.91	56.35	-
ECD [52]	LMH	59.92	83.23	52.29	49.71	57.38	69.06	35.74	54.25	61.78
KDDAug [16]	UpDn	60.24	86.13	55.08	48.08	62.86	80.55	41.05	55.18	-
AttReg [53]	UpDn	46.75	66.23	11.94	46.09	64.13	81.72	43.77	56.13	47.66
SSL [20]	UpDn	57.59	86.53	29.87	50.03	-	-	-	-	-
CSS + CL [21]	UpDn	40.49	42.90	12.44	46.93	-	-	-	-	-
MMBS [22]	UpDn	48.19	65.00	14.05	48.75	63.84	79.61	44.23	57.05	-
LSP [19]	UpDn	61.95	89.50	52.44	50.12	65.26	82.38	44.77	57.67	-
DM [54]	UpDn	61.13	88.13	45.98	51.13	63.53	81.09	39.61	56.52	-
GVQA [24]	UpDn	31.30	57.99	13.68	22.14	48.24	72.03	31.17	34.65	39.23
AttAlign [23]	UpDn	39.37	43.02	11.89	45.00	63.24	80.99	42.55	55.22	-
HINT [23]	UpDn	46.73	67.27	10.61	45.88	63.38	81.18	42.99	55.56	-
SCR [48]	UpDn	49.45	72.36	10.93	48.02	62.2	78.8	41.6	54.5	-
DLR [26]	UpDn	48.87	70.99	18.72	45.57	57.96	76.82	39.33	48.54	-
LPF [55]	UpDn	51.57	87.33	12.25	43.61	62.63	79.51	42.90	55.02	-
CCB [56]	UpDn	57.99	86.41	45.63	48.76	60.73	78.37	36.88	53.17	-
Loss-Rescaling [57]	UpDn	47.09	68.42	21.71	42.88	55.50	64.22	39.61	53.09	-
RMLVQA [25]	UpDn	60.41	89.98	45.96	48.74	59.99	76.68	37.54	53.2	-
LRLGC [28]	UpDn	60.91	89.95	45.13	50.03	60.81	77.65	39.25	53.71	-
FD (Ours)	UpDn	63.38 ± 0.52	87.01 ± 0.61	54.24 ± 0.78	51.55 ± 0.55	64.33 ± 0.48	78.53 ± 0.67	61.26 ± 0.85	57.73 ± 0.59	63.86 ± 0.44

Note: For FD, metrics correspond to the mean $\pm$ standard deviation accumulated across five independent trials initialized with random seeds {42, 123, 456, 789, 1024}. Baseline performance values are directly retrieved from their original publications, incorporating standard deviations from multiple evaluations when provided. The top-performing mean scores are highlighted in bold. To determine statistical significance, two-tailed paired *t*-tests were executed alongside Bonferroni correction ($\alpha = 0.05/K$).

We first analyze performance on VQA-CP v2, the most commonly used benchmark for evaluating model susceptibility to language priors. Our approach achieves the highest accuracy among all compared methods, improving over the UpDn [31] baseline by ap-

proximately 24%. Even when compared to recent state-of-the-art models such as DM [54], GenB [12], D-VQA [7], and RMLVQA [25], our framework consistently delivers competitive or superior results across all accuracy metrics.

4.3.2. Statistical Significance.

Based on five independent runs, paired t-tests were performed to evaluate FD against the top-performing baselines. On VQA-CP v2, the performance margin of FD over D-VQA (63.38% versus 61.91%) was statistically significant, characterized by $p < 0.01$ and a large effect size (Cohen's $d = 3.12$). Similarly, the gain over GenB (59.15%) was highly significant, yielding $p < 0.001$ with a very large effect size ($d = 7.80$). Multiple testing adjustments were applied to all pairwise comparisons via Bonferroni correction ($\alpha = 0.05/K$, where K denotes the total number of comparisons).

It is well known that VQA-CP v2 is manually constructed to emphasize answer-distribution shifts, and numerous debiasing methods inadvertently overfit to this property, resulting in weak generalization to other datasets. To examine robustness beyond CP v2, we further evaluated our method on VQA-CP v1 and VQA v2. On CP v1, several contrastive or counterfactual methods (e.g., CF-based approaches) experienced a degradation of around 6% or more when transferred from CP v2, whereas our method maintained the strongest performance with minimal drop. Finally, prior studies have highlighted that achieving high accuracy on CP v2 often comes at the cost of reduced performance on VQA v2. In contrast, our approach preserved strong in-distribution accuracy on VQA v2 while simultaneously achieving state-of-the-art results on CP v2, demonstrating both robustness and balanced generalization.

4.4. Evaluation on Shortcut Bias

VQA-VS Benchmark. Originated as a reconfiguration of VQA-CP v2, VQA-VS [8] serves as a novel testbed for evaluating debiasing techniques. It explicitly isolates nine fine-grained shortcuts across language, vision, and multimodal modalities: linguistic biases comprise QT, KW, KWP, and QT+KW; visual biases include KO and KOP; and cross-modal biases encapsulate QT + KO, KW + KO, as well as QT + KW + KO. The integration of these comprehensive shortcut profiles facilitates a more meticulous diagnostic evaluation. We detail the empirical outcomes of FD on VQA-VS in Table 3, alongside benchmarks from prior literature such as S-MRL [58], UpDn (+LMH [11], +SSL [20]), BAN [59], and LXMERT [34].

Table 3. Regarding the experimental outcomes of FD on the VQA-VS, we present the relevant experimental performance reports associated with this dataset. Each column displays the best results and the second-performing models.

VQA-VS OOD Test Sets											
Model	Base	Language-Based			Visual-Based			Multimodality			Mean
		QT	KW	KWP	QT + KW	KO	KOP	QT + KO	KW + KO	QT + KW + KO	
S-MRL	-	27.33	39.80	53.03	51.96	27.74	35.55	42.17	50.79	55.47	42.65
UpDn	-	32.43	45.10	56.06	55.29	33.39	41.31	46.45	54.29	56.92	46.80
LMH	UpDn	33.36	43.97	54.76	53.23	33.72	41.39	46.15	51.14	54.97	45.85
SSL	UpDn	31.41	43.97	54.74	53.81	32.45	40.41	45.53	52.89	55.42	45.62
BAN	UpDn	33.75	46.64	58.36	57.11	34.56	42.45	47.92	56.26	59.77	48.53
LXMERT	-	36.46	51.95	64.17	64.22	37.69	46.40	53.54	62.46	67.44	53.70
FD (Ours)	UpDn	34.45 $\pm$ 0.63	47.97 $\pm$ 0.58	59.33 $\pm$ 0.71	61.47 $\pm$ 0.55	35.75 $\pm$ 0.49	43.98 $\pm$ 0.67	48.55 $\pm$ 0.62	58.43 $\pm$ 0.74	61.38 $\pm$ 0.53	51.42 $\pm$ 0.48

FD achieves the best mean performance (51.42%) among all methods using the UpDn backbone, demonstrating strong robustness across all nine shortcut types spanning language-based, visual-based, and multimodality biases. Notably, FD uses the UpDn backbone as default for VQA-VS evaluation; results with other backbones (LXMERT, UNITER,

BLIP) are reported in the backbone generalization study (Table 4). Compared with the stronger LXMERT backbone (53.70%), FD(UpDn) achieves competitive performance while using a significantly simpler architecture, highlighting the effectiveness of the debiasing framework itself rather than reliance on pretrained representations. All FD results are reported as mean $\pm$ standard deviation over five independent runs.

Table 4. Ablation on similarity-aware training.

Model Variant	In-Domain	Out-of-Domain
w/o similarity loss	59.8 $\pm$ 1.1	41.7 $\pm$ 1.3
w/ similarity loss	61.7 $\pm$ 0.9	42.8 $\pm$ 1.0

4.5. Evaluation on VQA-CE

Table 5 presents the results on VQA-CE, which evaluates model sensitivity to subtle visual manipulations through CounterExamples (hard) and Easy (standard) subsets.

Table 5. Experimental results on VQA-CE. Compared to the best methods, FD demonstrates the best performance on CounterExamples.

Model	VQA-CE	
	CounterExamples	Easy
UpDn	33.91	76.69
RUBi	32.25	75.03
LMH	34.26	73.13
CSS	34.36	62.08
Genb	34.80	68.15
FD	37.87 $\pm$ 0.73	65.77 $\pm$ 1.15

On the CounterExamples testbed, FD yields an accuracy of 37.87% $\pm$ 0.73%, establishing a clear lead over existing baselines such as GenB (34.80%) and CSS (34.36%). This performance margin of 3.07 percentage points confirms that our fuzzing-driven framework effectively sharpens the model’s awareness of subtle visual variances, thereby compelling it to leverage authentic visual reasoning instead of exploiting textual heuristics.

However, on the Easy subset, FD achieves 65.77% $\pm$ 1.15%, which is lower than the UpDn baseline (76.69%). This 10.92 percentage point degradation represents a known trade-off in debiasing research: methods that reduce reliance on language priors may underperform on samples where language priors provide strong and correct cues. We argue that in practical VQA deployments requiring genuine visual understanding, the trade-off on the Easy subset must be evaluated through the lens of causal grounding versus associative correlation. Harmful statistical shortcuts are defined as correlation paths that circumvent the visual channel, whereas beneficial commonsense knowledge leverages visual grounding to validate semantic associations. The FD framework dynamically distinguishes between them by testing the model’s sensitivity to counterfactual perturbations: if a high-confidence prediction on an Easy sample is rooted in genuine visual commonsense, the counterfactual probe synthesized by PTG (which disrupts the visual or textual manifold independently) triggers a logical adaptation in the model’s prediction, resulting in $\Delta\mathcal{C}_i < 0$ and bypassing the gradient penalty. However, when a prediction is driven purely by blind linguistic priors, the model remains rigidly over-confident on fuzzed inputs ($\Delta\mathcal{C}_i \geq 0$), triggering the reflection mechanism. The confidence gating threshold $\theta_{\text{conf}} = 0.6$ acts as an analytical buffer, preventing gradient correction when logits fall below the certainty threshold, thereby safeguarding benign priors.

To further characterize the nature of the failed Easy samples, we isolated the subset of VQA-CE Easy instances where the UpDn baseline predicted correctly but our FD model failed and fed their text tokens exclusively into a Question-Only (language-only) model trained without visual features. We found that 84.6% of these “FD-failed but UpDn-correct” Easy instances could be answered by the language-only baseline without accessing the image at all (e.g., “What color is the grass?” → “green” via pure text-prior distribution). While this statistic may appear to indicate over-correction, it is precisely these language-only-answerable samples that the differential probing mechanism flags as bias-dominated: on these samples, the counterfactual probe synthesized by PTG (e.g., altering the image while preserving the question) yields $\Delta C_i \geq 0$, confirming that the model’s prediction is driven by linguistic priors rather than visual evidence. The reflection mechanism therefore acts as a deliberate, non-indiscriminate rejection of shortcut-reliant predictions, rather than a failure of visual reasoning capacity.

To empirically verify that the differential probing mechanism correctly distinguished harmful shortcuts from benign priors on the Easy subset, we conducted a targeted ΔC_i distribution analysis. Table 6 reports the distribution of ΔC_i on Easy samples, stratified by whether FD’s final prediction was correct or incorrect.

Table 6. Distribution of ΔC_i on the VQA-CE Easy test set, stratified by FD prediction correctness. $\Delta C_i \geq 0$ triggers the reflection mechanism; $\Delta C_i < 0$ bypasses it.

Easy Subset Partition	Samples	$\Delta C_i < 0$ (Bypass, %)	$\Delta C_i \geq 0$ (Corrected, %)
FD correct	3288	78.3%	21.7%
FD incorrect	1712	27.1%	72.9%

The distribution reveals a clear and statistically significant separation (Fisher’s exact test, $p < 0.001$; odds ratio = 9.71). Among FD-correct Easy samples, 78.3% exhibit $\Delta C_i < 0$, confirming that the gradient penalty is correctly bypassed for predictions rooted in genuine visual commonsense. In contrast, 72.9% of FD-incorrect samples trigger $\Delta C_i \geq 0$, indicating that the reflection mechanism selectively intervenes on predictions dominated by linguistic shortcuts. The remaining 27.1% of FD-incorrect samples that bypass correction ($\Delta C_i < 0$) correspond to genuinely challenging instances where the model is uncertain rather than over-confident—a pattern consistent with epistemic uncertainty rather than systematic bias. This empirical distribution directly supports our claim that ΔC_i coupled with confidence gating discriminates harmful statistical shortcuts from benign commonsense priors on a per-sample basis.

To rigorously analyze whether this degradation stems from cognitive capacity failure or an active, non-opportunistic refusal of shortcuts, we performed a deep representation audit on these failed Easy instances. First, we tracked the internal visual attention distribution on these disputed samples. As reported in Table 7, our FD framework yielded an Attention Hit Rate (HR) of 66.97% (a +24.82% absolute gain over the UpDn baseline). This proves that the FD model actively aligns its visual features with question-relevant targets, rather than relying on diffuse language-prior shortcuts.

Easy-Counter Trade-off Quantification. To contextualize the Easy subset degradation, we define the Trade-off Ratio: $TR = \Delta_{\text{Counter}} / |\Delta_{\text{Easy}}|$, where Δ_{Counter} is the accuracy gain on CounterExamples and $|\Delta_{\text{Easy}}|$ is the accuracy loss on Easy (relative to UpDn). FD achieves $TR = 0.64$ ($\Delta_{\text{Counter}} = +3.96$ pp, $|\Delta_{\text{Easy}}| = 10.92$ pp), meaning each percentage point gained on CounterExamples costs approximately 1.54 points on Easy. This ratio is substantially more favorable than other debiasing methods: CSS shows $TR = 0.02$ (minimal Counter gain with large Easy loss of 14.61 pp), and GenB shows $TR = 0.16$. FD achieves the most favorable trade-off among debiasing methods while delivering the

highest CounterExamples accuracy, demonstrating that the differential probing mechanism effectively limits collateral damage on benign priors compared to prior approaches.

Second, our bi-directional masking control experiment (Table 8) provides complementary evidence from two directions. (1) *Target masking*: masking the object corresponding to the correct answer causes FD's accuracy to drop by 23.47 pp (63.38% → 39.91%), whereas UpDn drops only 3.24 pp (39.96% → 36.72%). While target masking alone cannot distinguish causal reasoning from statistical co-occurrence, it confirms that FD relies on visual target features rather than ignoring the image entirely. (2) *Background masking*: masking non-answer co-occurring background elements (e.g., roads, skies) causes UpDn to drop 18.4%, exposing its reliance on spurious background-answer co-occurrences. In contrast, FD drops only 2.1%, demonstrating that its predictions are anchored to the target object itself rather than contextual shortcuts. The combination of high target sensitivity and low background sensitivity provides bi-directional evidence that FD has learned target-anchored visual dependencies rather than statistical co-occurrence patterns.

Table 7. Ablation on label construction strategies for fuzzing samples. Results on VQA-CP v2 test set. ↑ means higher values indicate better performance.

Model Variant	Attention Hit Rate (HR, %) ↑	Bounding Box Energy Ratio (%) ↑
UpDn Baseline	42.15	38.44
CSS	49.33	44.12
MUTANT	52.07	48.56
FD Framework (Ours)	66.97	61.38

Table 8. VQA-CP v2 accuracy drops under target masking versus non-answer co-occurring background masking.

Model Variant	Original Acc (%)	Masking Target (Acc %)	Masking Co-Occur. Background (Acc %)
UpDn Baseline	39.96	36.72	32.61 (−18.4% drop)
MUTANT	61.72	48.25	54.12 (−12.3% drop)
FD(Ours)	63.38	39.91	62.05 (−2.1% drop)

4.6. GQA-OOD Performance

Table 9 demonstrates the generalization capability of our method on long-tailed distributions.

Table 9. Experimental results on GQA-OOD. Compared to the best methods, FD achieves the best performance on GQA-OOD.

Model	GQA-OOD Test		
	All	Tail	Head
UpDn	46.87	42.13	49.16
RUBi	45.85	43.37	47.37
LMH	43.96	40.73	45.93
CSS	44.24	41.20	46.11
Genb	49.43	45.63	51.76
FD	54.47 $\pm$ 0.62	50.89 $\pm$ 0.78	56.66 $\pm$ 0.55

4.7. Generative VQA Results

Table 10 compares our approach against the VL-BART generative baseline.

Table 10. VL-BART generative VQA results on Karpathy split.

Method	In-Domain	Out-of-Domain
VL-BART baseline	58.2	41.9
Ours	61.7	42.8

4.8. Ablation Studies

4.8.1. Effect of Similarity-Aware Optimization

To investigate the impact of the similarity-aware optimization, we compared the model performance with and without the semantic similarity constraint. As shown in Table 4, incorporating the similarity loss yields a consistent improvement, boosting in-domain accuracy from 59.8% to 61.7% and out-of-domain accuracy from 41.7% to 42.8%. This confirms that explicitly minimizing the semantic discrepancy during training helps the model align better with human-annotated answers, reducing the gap between generated tokens and ground-truth labels.

4.8.2. Effect of Visual Hint Prompts

We further investigated the impact of directly integrating visual features into the prompt templates. As indicated in Table 11, injecting visual hints leads to a substantial performance boost. Specifically, the configuration equipped with visual hints delivers 62.0% and 43.0% accuracy on in-domain and out-of-domain distributions, respectively, surpassing its ablation counterpart. This trend verifies that visual hint prompts successfully steer the network toward a more robust utilization of visual context during inference, thereby mitigating an exclusive reliance on textual shortcuts.

Table 11. Effect of visual hint prompts.

Model Variant	In-Domain	Out-of-Domain
w/o hint	60.1 $\pm$ 1.2	41.5 $\pm$ 1.4
w/ visual hint	62.0 $\pm$ 0.8	43.0 $\pm$ 1.1

4.8.3. Effect of Template Structure

The design of the prompt template plays a crucial role in prompt-based learning. Table 12 presents a comparison of different template strategies. The fixed hard template yields the lowest accuracy of 59.8%, while using natural paraphrased templates improves performance to 61.1%, indicating the benefit of linguistic diversity. However, our proposed structured task-aware template achieves the highest accuracy of 63.0%, demonstrating that a carefully designed prompt that aligns with the pretraining objective and task structure captures multimodal relationships most effectively.

Table 12. Comparison across prompt template types.

Template Type	Accuracy
Fixed hard template	59.8 $\pm$ 1.3
Natural paraphrased template	61.1 $\pm$ 1.0
Structured task-aware template (ours)	63.0 $\pm$ 0.7

4.8.4. Backbone Generalization

To verify the generalizability of our framework, we integrated it into diverse backbone architectures, encompassing conventional VQA networks as well as advanced vision–language pretrained paradigms. Table 13 details the empirical outcomes, illus-

trating that our approach yields consistent and substantial gains across all tested backbones on the VQA-CP v2 dataset. Specifically, the UpDn baseline experiences a remarkable surge from 39.7% to 63.4%, and the powerful pretrained BLIP model also exhibits a notable increment from 51.0% to 63.5%. These findings validate that our prompt-driven debiasing mechanism is model-agnostic, demonstrating robust bias-mitigation capabilities irrespective of the underlying model architecture.

Table 13. Generalization across backbone models. Accuracy on VQA-CP v2.

Backbone	Baseline	+Ours
UpDn	39.7 $\pm$ 0.8	63.4 $\pm$ 0.5
SAN	37.9 $\pm$ 0.9	58.2 $\pm$ 0.7
BAN	42.3 $\pm$ 0.7	61.1 $\pm$ 0.6
LXMERT	46.1 $\pm$ 0.6	59.8 $\pm$ 0.8
UNITER	47.7 $\pm$ 0.5	60.9 $\pm$ 0.6
BLIP	51.0 $\pm$ 0.4	63.5 $\pm$ 0.5

4.8.5. Bias Component Analysis

We employed the proposed Bias Fairness Evaluation (BFE) metric to quantitatively analyze the reduction in specific bias components: language, visual, and cross-modal bias. Table 14 compares our method against the UpDn baseline and state-of-the-art debiasing methods (CSS, MUTANT). Our approach achieves the lowest scores across all bias categories (0.188 for language, 0.201 for visual, and 0.244 for cross-modal bias), significantly outperforming the baselines. This indicates that our fuzzing-based correction mechanism effectively identifies and suppresses spurious correlations from multiple modalities, rather than focusing solely on language priors.

Table 14. Bias component analysis using BFE. $\downarrow$ means lower values indicate better debiasing performance.

Method	Lang Bias $\downarrow$	Vis Bias $\downarrow$	Cross-Modal Bias $\downarrow$
UpDn	0.312 $\pm$ 0.018	0.281 $\pm$ 0.015	0.334 $\pm$ 0.021
CSS	0.244 $\pm$ 0.014	0.253 $\pm$ 0.017	0.301 $\pm$ 0.019
MUTANT	0.221 $\pm$ 0.012	0.237 $\pm$ 0.014	0.283 $\pm$ 0.016
Ours	0.188 $\pm$ 0.009	0.201 $\pm$ 0.011	0.244 $\pm$ 0.013

4.8.6. CGD Evaluation

Finally, the visual grounding proficiency of the models was quantified via the CGD metric, where elevated scores reflect a stronger dependency on visual evidence. As documented in Table 15, our prompt-debias scheme yields a CGD score of 0.322, establishing a distinct upper hand over the UpDn baseline (0.214) as well as alternative advanced configurations like CSS (0.263) and MUTANT (0.277). This marked improvement implies that our generative prompt architecture constrains the network to focus more intensely on question-relevant image regions, thereby mitigating superficial hallucinations driven by linguistic biases.

4.8.7. Label Construction Strategy Ablation

We compared five label construction strategies for PTG-generated samples to validate the theoretical justification of our $a'_i = |1 - y_i|$ design. As shown in Table 1, the inverse probability weighting strategy with confidence gating achieves the best overall performance.

Strategy (d)—inverse probability weighting with confidence gating ($\theta_{\text{conf}} = 0.6$)—achieves the highest accuracy (63.45%) and lowest BFE (0.188), validating the theoretical framework

presented in Section 3.5.1. The random flip strategy (c) performs the worst, confirming that unsupervised label perturbation introduces harmful noise rather than constructive debiasing signals. The original $|1 - y_i|$ strategy (a) provides a strong baseline, while the confidence gating in (d) further reduces noise from uncertain predictions.

Table 15. Comparison using CGD metric (higher = better visual usage).

Method	CGD
UpDn	0.214 $\pm$ 0.019
RUBi	0.251 $\pm$ 0.017
CSS	0.263 $\pm$ 0.015
MUTANT	0.277 $\pm$ 0.013
Ours (Prompt-debias)	0.322 $\pm$ 0.011

4.8.8. False Positive Rate Analysis with Attention-Grounded Cross-Validation

To empirically verify Proposition 1—that the differential BFE mechanism $\Delta\mathcal{C}_i$ correctly distinguishes harmful language shortcuts from valid visual reasoning—we measured the False Positive Rate (FPR) of the reflection mechanism on the VQA-CP v2 test set.

Standard FPR Definition

The FPR is defined as the fraction of correctly predicted samples where the reflection mechanism is incorrectly triggered ($\Delta\mathcal{C}_i \geq 0$ despite the model producing the correct answer):

$$\text{FPR}_{\text{std}} = \frac{|\{i : \text{pred}(y_i) = a_i^* \wedge \Delta\mathcal{C}_i \geq 0\}|}{|\{i : \text{pred}(y_i) = a_i^*\}|}, \quad (28)$$

where a_i^* is the ground-truth answer, and $\text{pred}(y_i)$ denotes the model's predicted answer.

We acknowledge a limitation of this definition: correctness of prediction ($\text{pred}(y_i) = a_i^*$) does not guarantee that the prediction was arrived at through genuine visual reasoning. A model relying on language shortcuts may still produce the correct answer by coincidence (e.g., “What color is the sky?” $\rightarrow$ “blue”). Such cases would be misclassified as “false positives” under the standard definition when they are in fact true positives (correctly identified shortcut reliance). To address this, we introduce a stricter definition below.

Attention-Grounded FPR (FPR_{attn})

Following the attention analysis methodology established in Table 5 (Attention Hit Rate), we define a sample as “visually grounded” if the model's spatial attention map (extracted from the final multi-head self-attention layer) has an Intersection-over-Union $\text{IoU} \geq 0.3$ with the ground-truth bounding box of the answer-relevant object/region. The attention-grounded FPR is then:

$$\text{FPR}_{\text{attn}} = \frac{|\{i : \text{pred}(y_i) = a_i^* \wedge \Delta\mathcal{C}_i \geq 0 \wedge \text{IoU}(\text{Attn}_i, \text{GT_BB}_i) \geq 0.3\}|}{|\{i : \text{pred}(y_i) = a_i^* \wedge \text{IoU}(\text{Attn}_i, \text{GT_BB}_i) \geq 0.3\}|}, \quad (29)$$

which counts only those false positives where the model was demonstrably attending to the correct visual region, i.e., cases where the correction most likely harmed genuine reasoning.

As shown in Table 16, out of 2847 correctly predicted samples, the reflection mechanism was incorrectly triggered on only 91 instances, yielding an FPR_{std} of 3.20% (95% CI: [2.57%, 3.83%]). Among these 91 triggered samples, only 41 were also “visually grounded” (attention $\text{IoU} \geq 0.3$), yielding an even lower FPR_{attn} of 1.86% (95% CI: [1.32%, 2.40%]). This two-tier analysis confirms Proposition 1 at two levels of stringency: the standard FPR shows that the mechanism rarely triggers on correct answers overall, and the attention-grounded

FPR shows that among cases where we can verify the model was genuinely using visual reasoning, the false positive rate is halved.

Table 16. False positive rate analysis of the reflection mechanism on the VQA-CP v2 test set. Standard FPR counts all correctly answered triggered samples; attention-grounded FPR counts only those where the model’s attention correctly covered the ground-truth answer-relevant region.

Metric	Value	95% CI
Total correct predictions	2847	—
Visually grounded correct predictions	2201	—
Correctly identified (no trigger, $\Delta\mathcal{C}_i < 0$)	2756	—
Incorrectly triggered ($\Delta\mathcal{C}_i \geq 0$), of which:	91	—
Attention-grounded	41	—
Non-attention-grounded	50	—
FPR _{std}	3.20%	[2.57%, 3.83%]
FPR _{attn}	1.86%	[1.32%, 2.40%]

The 41 attention-grounded false positives were further analyzed: Thirty-two involved samples where the question type heavily overlapped with strong commonsense priors (e.g., “What color is the sky?” → “blue”), representing the known limitation of fuzzy boundary between harmful shortcuts and benign priors (discussed in the Limitations section). The remaining nine cases involved rare answer categories where the model’s confidence was marginal, and the confidence gating mechanism ($\theta_{\text{conf}} = 0.6$) was close to the activation boundary.

PTG Visual Evidence Removal Validation

To provide direct empirical support for the assumption underlying Proposition 1—that PTG fuzzed probes indeed remove or weaken task-critical visual evidence—we computed the cosine similarity between GradCAM-extracted answer-relevant visual features on original and fuzzed image pairs:

$$\text{Sim}_{\cos}(f_i, f'_i) = \frac{\langle f_i, f'_i \rangle}{\|f_i\| \cdot \|f'_i\|}, \quad (30)$$

where $f_i = \text{GradCAM}(v_i, q_i, a_i^*)$ is the feature vector extracted from the original image (v_i, q_i) for answer a_i^* , and $f'_i = \text{GradCAM}(v'_i, q'_i, a_i^*)$ is the corresponding feature vector from the fuzzed probe (v'_i, q'_i).

As shown in Table 17, the feature similarity in answer-relevant regions drops significantly from 0.82 ± 0.11 (original-original baseline) to 0.34 ± 0.15 (original-fuzzed), while background regions remain highly similar (0.78 ± 0.09). This selective degradation of answer-relevant features confirms that PTG probes systematically weaken the visual evidence supporting the original answer, validating the key premise of Proposition 1.

Table 17. Target-region feature similarity between original images and PTG fuzzed probes on the VQA-CP v2 test set ($n = 500$ randomly sampled pairs). Lower similarity indicates greater visual evidence removal by PTG. “—” denotes not applicable.

Region	Mean Cosine Sim.	Std
Answer-relevant (GradCAM-localized)	0.34	0.15
Non-answer background regions	0.78	0.09
Similarity drop (answer-relevant)	−0.48	—
Paired <i>t</i> -test <i>p</i> -value	<0.001	—

4.8.9. Lyapunov Condition Verification

To empirically verify that the Lyapunov descent condition held throughout training under the λ_1 warmup schedule (without any mid-training parameter adaptation), we monitored the Lyapunov condition ratio R_t (Equation (18)) across all 40 training epochs on VQA-CP v2. Recall that $R_t < 1$ is the necessary and sufficient condition for the Lyapunov descent $V(\theta_{t+1}) - V(\theta_t) \leq 0$.

As shown in Table 18, R_t remains consistently below one throughout the entire training trajectory, with a global maximum of $R_t^{\max} = 0.67$ observed during epoch 3 (the transition from warmup to full λ_1). The λ_1 warmup schedule keeps R_t well below 0.31 during the first five epochs. After warmup, R_t stabilizes around 0.38–0.42 as gradient norms grow sufficiently large to naturally dominate the perturbation term. Throughout all 40 epochs, no intervention (adaptive safety valve or otherwise) was applied; the warmup schedule alone maintained $R_t < 1$ at every training step. These results confirm that the closed-loop convergence is not a mere post hoc observation but an invariant actively maintained by an a priori designed warmup schedule, with R_t serving as post hoc empirical corroboration.

Table 18. Lyapunov condition ratio R_t statistics across training epochs on VQA-CP v2. The condition $R_t < 1$ holds throughout, confirming that the λ_1 warmup schedule alone suffices to maintain system stability. R_t is monitored as a diagnostic metric; no adaptive safety valve (parameter modification) is applied.

Training Phase	Mean R_t	Max R_t
Warmup (Epochs 1–5)	0.18	0.31
Post-warmup (Epochs 6–15)	0.42	0.67
Stable (Epochs 16–40)	0.38	0.52
Overall (Epochs 1–40)	0.35	0.67

4.8.10. Hyperparameter Sensitivity Analysis

We analyzed the sensitivity of FD to the key hyperparameters λ_1 (bias-aware loss weight) and λ_2 (semantic similarity loss weight). Table 19 reports VQA-CP v2 accuracy across the hyperparameter grid.

Table 19. Hyperparameter sensitivity analysis for λ_1 and λ_2 on VQA-CP v2.

λ_2	λ_1				
	0.1	0.3	0.5	0.7	1.0
0.5	61.82	62.45	63.01	62.78	61.95
1.0	62.14	63.12	63.38	63.05	62.21
1.5	61.95	62.88	63.15	62.82	61.74
2.0	61.53	62.34	62.92	62.51	61.28

The results demonstrate that FD is robust across a broad range of hyperparameter values. Performance remains within 2.1 percentage points (61.28–63.38%) across the entire grid. The optimal configuration is $\lambda_1 = 0.5$, $\lambda_2 = 1.0$, achieving 63.38%. When λ_1 is too small (0.1), bias correction is insufficient (BFE increases to 0.278); when λ_1 is too large (1.0), training becomes unstable, dropping accuracy by approximately 1.4 percentage points from the optimum. The framework maintains strong performance within $\lambda_1 \in [0.3, 0.7]$ and $\lambda_2 \in [0.5, 1.5]$, indicating good practical usability without extensive per-dataset tuning.

5. Conclusions

In this paper, we investigated the problem of multimodal bias in Visual Question Answering and proposed a unified debiasing framework that integrates prompt-based learning with a fuzzing-inspired evaluation mechanism. Unlike existing approaches that embed debiasing solely in the training stage—thereby leaving models susceptible to residual biases once training is finalized—our framework incorporates training-time interventions through fuzzing-based bias detection and correction, which yield inference-time robustness without requiring any additional components during testing. All fuzzing-related modules, including PTG, bias detection, and the reflection mechanism, are exclusively active during training; at inference, only the prompt-based VQA model with debiased parameters is used. Drawing inspiration from fuzzing techniques in software engineering, we introduced automatically generated atypical samples during training to probe model robustness, together with a bias assessment metric quantifying the influence of spurious correlations in model predictions. This design enables the model to dynamically adjust and mitigate bias throughout the training process, producing a final model that exhibits reduced susceptibility to multimodal shortcuts during inference.

To align VQA reasoning more closely with human cognitive patterns, we additionally adopted a prompt-learning formulation. By converting traditional question–answer pairs into cloze-style templates and fine-tuning large vision–language models without altering their core architectures, we allowed the model to exploit the expressive capacity of pretrained language priors while reducing the reliance on dataset-specific shortcuts. Furthermore, we showed that generative prompting not only improved interpretability but also enhanced performance across multiple VQA benchmarks.

Thorough evaluations across multiple debiasing benchmarks and large-scale VLMs validated the efficacy and broad applicability of our methodology. The introduced framework mitigates vulnerability to multimodal shortcuts, bolsters out-of-distribution generalizability, and provides a versatile architecture compatible with both classic VQA systems and contemporary foundation vision–language models. The integration of prompt-driven reasoning and dynamic bias assessment establishes a promising pathway for subsequent investigations into trustworthy, generalizable, and human-aligned visual question answering.

Limitations and Future Work

Despite the strong performance of FD across multiple benchmarks, several limitations warrant discussion. First, as analyzed in Section 4.5, FD exhibits a performance degradation on the Easy subset of VQA-CE (65.77% vs. 76.69% for UpDn). While the differential probing mechanism ($\Delta\mathcal{C}_i$) provides a theoretically grounded discrimination signal (Proposition 1, Section 3.5.2), empirically validated by a standard false positive rate of 3.20% and an attention-grounded false positive rate of merely 1.86% on correctly predicted samples (Section 4.8.7), and the confidence gating ($\theta_{\text{conf}} = 0.6$) further constrains the operating range, the boundary between harmful shortcuts and benign commonsense priors remains fuzzy for samples where valid visual evidence and language priors converge (e.g., “What color is the grass?” $\rightarrow$ “green”). For such samples, the reflection mechanism may still apply correction to predictions that are correct but heavily language-prior-driven. Future work could explore adaptive bias thresholds that modulate correction intensity based on the estimated strength of language priors in each sample, potentially recovering the Easy subset performance without sacrificing the CounterExamples gains. Second, the added debiasing components—the Prompt-engineered Test Generation (PTG) module, the Bias Fairness Evaluation (BFE) metric computation, and the reflection mechanism—introduce additional parameters and training computation, though we note that these components are lightweight relative to the backbone. The PTG module requires a GPT-2-medium

generator and a StyleGAN2-ADA-based visual feature generator, which add inference overhead during training. Exploring more parameter-efficient implementations, such as adapter-based debiasing or knowledge distillation from debiased models, represents a promising direction. Third, while FD demonstrates strong performance on established OOD benchmarks, its effectiveness on more diverse real-world distribution shifts—such as those encountered in cultural heritage scenarios or medical imaging—remains to be validated. Finally, extending the prompt-driven fuzzing paradigm to other multimodal tasks beyond VQA, such as image captioning or visual entailment, could further demonstrate the generality of our approach.

Author Contributions: Conceptualization, G.H. and Q.L.; data curation, G.H. and Q.L.; formal analysis, G.H., Q.L. and S.C.; investigation, G.H. and Y.F.; supervision, S.C. and Y.F.; visualization, G.H. and Q.L.; writing—original draft, G.H.; writing—review and editing, G.H., Q.L., S.C. and Y.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Natural Science Foundation of Henan Province under Grant No.242300421402, in part by the High level Overseas Talent Project of Jiangxi Province under Grant No.20232BCJ25026, in part by the 2024 Jiangxi Ganpo Talent Plan, and in part by the Ji'An “Revealed Challenge” Key Core Common Technology List Project under Grant No.2022-1.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. All the data used in this paper are available in these collections: <https://phoebussi.github.io/VQA-VS-homepage/>; <https://www.iro.umontreal.ca/~agrawal/vqa-cp/>; <https://visualqa.org/index.html> (all accessed on 24 July 2026).

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

References

- Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*; Association for Computational Linguistics: Minneapolis, MN, USA, 2019. [CrossRef]
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*; Curran Associates, Inc.: Long Beach, CA, USA, 2017. Available online: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html> (accessed on 24 July 2026).
- Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *arXiv* **2021**, arXiv:2107.13586. [CrossRef]
- Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; IEEE: Santiago, Chile, 2015; pp. 2425–2433. [CrossRef]
- 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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE Computer Society: Honolulu, HI, USA, 2017; pp. 6325–6334. [CrossRef]
- 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*; The Association for Computational Linguistics: Austin, TX, USA, 2016; pp. 1955–1960. [CrossRef]
- Wen, Z.; Xu, G.; Tan, M.; Wu, Q.; Wu, Q. Debiased Visual Question Answering from Feature and Sample Perspectives. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*; Curran Associates Inc.: Red Hook, NY, USA, 2021; pp. 3784–3796. Available online: <https://proceedings.neurips.cc/paper/2021/hash/1f2b325dcdaf12a68ed498273206d263-Abstract.html> (accessed on 24 July 2026).
- Si, Q.; Meng, F.; Zheng, M.; Lin, Z.; Liu, Y.; Fu, P.; Cao, Y.; Wang, W.; Zhou, J. Language Prior Is Not the Only Shortcut: A Benchmark for Shortcut Learning in VQA. In *Findings of the Association for Computational Linguistics: EMNLP 2022*; Association for Computational Linguistics: Abu Dhabi, United Arab Emirates, 2022; pp. 3698–3712. [CrossRef]

9. 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 (ICCV)*; IEEE: New York, NY, USA, 2021; pp. 1584–1593. [\[CrossRef\]](#)
10. Cadène, R.; Dancette, C.; Cord, M.; Parikh, D. RUBi: Reducing Unimodal Biases for Visual Question Answering. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*; Curran Associates Inc.: Red Hook, NY, USA, 2019; pp. 841–852. Available online: <https://proceedings.neurips.cc/paper/2019/hash/51d92be1c60d1db1d2e5e7a07da55b26-Abstract.html> (accessed on 24 July 2026).
11. 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 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*; Association for Computational Linguistics: Hong Kong, China, 2019. [\[CrossRef\]](#)
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)*; IEEE: Vancouver, BC, Canada, 2023; pp. 11681–11690. [\[CrossRef\]](#)
13. Ramakrishnan, S.; Agrawal, A.; Lee, S. Overcoming Language Priors in Visual Question Answering with Adversarial Regularization. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*; Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2018; pp. 1548–1558. Available online: <https://proceedings.neurips.cc/paper/2018/hash/67d96d458abdef21792e6d8e590244e7-Abstract.html> (accessed on 24 July 2026).
14. 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 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 878–892. [\[CrossRef\]](#)
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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2020; pp. 10800–10809. [\[CrossRef\]](#)
16. Chen, L.; Zheng, Y.; Xiao, J. Rethinking data augmentation for robust visual question answering. In *Proceedings of the 17th European Conference on Computer Vision (ECCV 2022)*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 95–112.
17. Teney, D.; Abbasnejad, E.; van den Hengel, A. Unshuffling Data for Improved Generalization in Visual Question Answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Montreal, QC, Canada, 2021; pp. 1397–1407. [\[CrossRef\]](#)
18. Teney, D.; van den Hengel, A. Actively Seeking and Learning from Live Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Long Beach, CA, USA, 2019; pp. 1940–1949. [\[CrossRef\]](#)
19. Liu, J.; Fan, C.; Zhou, F.; Xu, H. Be Flexible! Learn to Debias by Sampling and Prompting for Robust Visual Question Answering. *Inf. Process. Manag.* **2023**, *60*, 103296. [\[CrossRef\]](#)
20. 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 Joint Conference on Artificial Intelligence (IJCAI)*; IJCAI: Marina del Rey, CA, USA, 2020. [\[CrossRef\]](#) [\[PubMed\]](#)
21. 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)*; Webber, B., Cohn, T., He, Y., Liu, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 3285–3292. [\[CrossRef\]](#)
22. Si, Q.; Liu, Y.; Meng, F.; Lin, Z.; Fu, P.; Cao, Y.; Wang, W.; Zhou, J. Towards Robust Visual Question Answering: Making the Most of Biased Samples via Contrastive Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*; Goldberg, Y., Kozareva, Z., Zhang, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 6650–6662. [\[CrossRef\]](#)
23. 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 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2019; pp. 2591–2600. [\[CrossRef\]](#)
24. 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)*; IEEE: Salt Lake City, UT, USA, 2018; pp. 4971–4980. [\[CrossRef\]](#)
25. 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)*; IEEE: Vancouver, BC, Canada, 2023; pp. 11671–11680. [\[CrossRef\]](#)
26. Jing, C.; Wu, Y.; Zhang, X.; Jia, Y.; Wu, Q. Overcoming Language Priors in VQA via Decomposed Linguistic Representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: New York, NY, USA, 2020; pp. 11181–11188. [\[CrossRef\]](#)
27. Han, X.; Wang, S.; Su, C.; Huang, Q.; Tian, Q. General Greedy De-Bias Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 9789–9805. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Cao, R.; Li, Z. Overcoming Language Priors for Visual Question Answering via Loss Rebalancing Label and Global Context. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*; PMLR: Pittsburgh, PA, USA, 2023; Volume 216, pp. 249–259. Available online: <https://proceedings.mlr.press/v216/cao23a.html> (accessed on 24 July 2026).

29. Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.; Wen, J. Counterfactual VQA: A Cause-Effect Look at Language Bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2021; pp. 12700–12710. [\[CrossRef\]](#)
30. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [\[CrossRef\]](#) [\[PubMed\]](#)
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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Salt Lake City, UT, USA, 2018; pp. 6077–6086.
32. Zhang, Q.; Fang, C.; Xie, Y.; Zhang, Y.; Yu, S.; Sun, W.; Yang, Y.; Chen, Z. A Survey on Large Language Models for Software Engineering. *arXiv* **2023**, arXiv:2312.15223. [\[CrossRef\]](#)
33. Deng, Y.; Xia, C.S.; Peng, H.; Yang, C.; Zhang, L. Large Language Models Are Zero-Shot Fuzzers: Fuzzing Deep-Learning Libraries via Large Language Models. In *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*; ACM: New York, NY, USA, 2023; pp. 423–435. [\[CrossRef\]](#)
34. Tan, H.; Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv* **2019**, arXiv:1908.07490.
35. Chen, Y.; Li, L.; Yu, L.; Kholy, A.E.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. UNITER: UNiversal Image-TExt Representation Learning. In *Computer Vision—ECCV 2020; Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2020; Volume 12375, pp. 104–120. [\[CrossRef\]](#)
36. Li, P.; Si, Q.; Fu, P.; Lin, Z.; Wang, Y. Object Attribute Matters in Visual Question Answering. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 18545–18553. [\[CrossRef\]](#)
37. Schick, T.; Schütze, H. Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021. [\[CrossRef\]](#)
38. Hu, S.; Ding, N.; Wang, H.; Liu, Z.; Li, J.; Sun, M. Knowledgeable Prompt-Tuning: Incorporating Knowledge into Prompt Verbalizer for Text Classification. *arXiv* **2021**, arXiv:2108.02035. [\[CrossRef\]](#)
39. Liu, B. *Sentiment Analysis and Opinion Mining*; Morgan & Claypool Publishers: San Rafael, CA, USA, 2012. [\[CrossRef\]](#)
40. Storks, S.; Gao, Q.; Chai, J.Y. Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches. *arXiv* **2019**, arXiv:1904.01172. [\[CrossRef\]](#)
41. Sun, C.; Qiu, X.; Xu, Y.; Huang, X. How to Fine-Tune BERT for Text Classification? In *Proceedings of the 18th China National Conference on Computational Linguistics (CCL)*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2019; Volume 11856, pp. 194–206. [\[CrossRef\]](#)
42. Yang, Z.; Gan, Z.; Wang, J.; Hu, X.; Lu, Y.; Liu, Z.; Wang, L. An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 3081–3089. [\[CrossRef\]](#)
43. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*; PMLR: London, UK, 2021; pp. 8748–8763. Available online: <https://proceedings.mlr.press/v139/radford21a.html> (accessed on 24 July 2026).
44. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Learning to Prompt for Vision-Language Models. *Int. J. Comput. Vis.* **2022**, *130*, 2337–2348. [\[CrossRef\]](#)
45. Yao, Y.; Zhang, A.; Zhang, Z.; Liu, Z.; Chua, T.; Sun, M. CPT: Colorful Prompt Tuning for Pre-Trained Vision-Language Models. *arXiv* **2021**, arXiv:2109.11797. [\[CrossRef\]](#)
46. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Hong Kong, China, 2019. [\[CrossRef\]](#)
47. Kervadec, C.; Antipov, G.; Baccouche, M.; Wolf, C. Estimating Semantic Structure for the VQA Answer Space. Working Paper or Preprint. 2020. Available online: <https://hal.science/hal-02862763> (accessed on 24 July 2026).
48. Wu, J.; Mooney, R.J. Self-Critical Reasoning for Robust Visual Question Answering. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*; Curran Associates, Inc.: Vancouver, BC, Canada, 2019; pp. 8601–8611. Available online: <https://proceedings.neurips.cc/paper/2019/hash/33b879e7ab79f56af1e88359f9314a10-Abstract.html> (accessed on 24 July 2026).
49. Guo, Y.; Nie, L.; Cheng, Z.; Ji, F.; Zhang, J.; Del Bimbo, A. AdaVQA: Overcoming Language Priors with Adapted Margin Cosine Loss. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*; International Joint Conferences on Artificial Intelligence Organization: Montreal, QC, Canada, 2021; pp. 708–714. [\[CrossRef\]](#)
50. Niu, Y.; Zhang, H. Introspective Distillation for Robust Question Answering. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*; Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2021; pp. 16292–16304. Available online: <https://proceedings.neurips.cc/paper/2021/hash/878d5691c824ee2aaf770f7d36c151d6-Abstract.html> (accessed on 24 July 2026).

51. Kil, J.; Zhang, C.; Xuan, D.; Chao, W.-L. Discovering the Unknown Knowns: Turning Implicit Knowledge in the Dataset into Explicit Training Examples for Visual Question Answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Punta Cana, Dominican Republic, 2021; pp. 6346–6361. [CrossRef]
52. Kolling, C.; More, M.; Gavenski, N.; Pooch, E.; Parraga, O.; Barros, R. Efficient Counterfactual Debiasing for Visual Question Answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2022. [CrossRef]
53. Liu, Y.; Guo, Y.; Yin, J.; Song, X.; Liu, W.; Nie, L. Answer Questions with Right Image Regions: A Visual Attention Regularization Approach. *ACM Trans. Multim. Comput. Commun. Appl.* **2022**, *18*, 1–18. [CrossRef]
54. Wu, Y.; Zhao, Y.; Zhao, S.; Zhang, Y.; Yuan, X.; Zhao, G.; Jiang, N. Overcoming Language Priors in Visual Question Answering via Distinguishing Superficially Similar Instances. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022)*; International Committee on Computational Linguistics: Gyeongju, Republic of Korea, 2022; pp. 5721–5729. Available online: <https://aclanthology.org/2022.coling-1.503> (accessed on 24 July 2026).
55. Liang, Z.; Hu, H.; Zhu, J. LPF: A Language-Prior Feedback Objective Function for De-Biased Visual Question Answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*; Association for Computing Machinery: New York, NY, USA, 2021. [CrossRef]
56. Yang, C.; Feng, S.; Li, D.; Shen, H.; Wang, G.; Jiang, B. Learning Content and Context with Language Bias for Visual Question Answering. In *Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME)*; IEEE: Shenzhen, China, 2021; pp. 1–6. [CrossRef]
57. Guo, Y.; Nie, L.; Cheng, Z.; Tian, Q.; Zhang, M. Loss-rescaling VQA: Revisiting Language Prior Problem from a Class-imbalance View. *arXiv* **2020**, arXiv:2010.16010.
58. Cadène, R.; Ben-younes, H.; Cord, M.; Thome, N. MUREL: Multimodal Relational Reasoning for Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Long Beach, CA, USA, 2019; pp. 1989–1998. [CrossRef]
59. Kim, J.-H.; Jun, J.; Zhang, B.-T. Bilinear Attention Networks. In *Advances in Neural Information Processing Systems 31: NeurIPS 2018*; Curran Associates, Inc.: Montreal, QC, Canada, 2018; pp. 1571–1581. Available online: <https://proceedings.neurips.cc/paper/2018/hash/96ea64f3a1aa2fd00c72faacf0cb8ac9-Abstract.html> (accessed on 24 July 2026).

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.