Robust and Interpretable Multimodal Fusion in Med-VQA: A Perturbation Benchmark for Clinical Safety

Mansi Dhamne[1], **Vivek Gangwani**[1], **Swapnali Kurhade**[1]

[1]Sardar Patel Institute of Technology
Mumbai, India
{mansi.dhamne22, vivek.gangwani22, swapnali.kurhade}@spit.ac.in

Abstract

Ensuring safety and trust in Medical Visual Question Answering (Med-VQA) systems demands robustness to real-world noise, interpretability for clinical users, and computational efficiency for deployment in safety-critical settings. Current vision–language models rely on opaque, resource-intensive fusion schemes whose vulnerability to adversarial perturbations remains largely unexplored. We present a comprehensive study of adversarial robustness and explainable reasoning in Med-VQA through a systematic benchmark of lightweight, interpretable fusion strategies. Across three standard datasets—VQA-RAD, PATH-VQA, and SLAKE—we evaluate models under clinically plausible visual (blur, artifacts, compression) and linguistic (typos, paraphrasing, negation) perturbations. Results show that ResNet152+BERT achieves strong clean-set accuracy (88.4% on PATH-VQA) but loses up to 19% under impulse noise, while ResNet50+BiLSTM demonstrates higher resilience to text perturbations, outperforming on 20 of 26 linguistic corruption types. Interpretability analyses using Grad-CAM and Integrated Gradients reveal consistent attention realignment under corruption, enabling transparent error attribution. These findings establish one of the first empirical baselines for adversarial robustness in Med-VQA and highlight that efficient, shallow fusion can yield interpretable, resilient reasoning with significantly reduced compute and latency—paving the way for safe, deployable medical AI in real-world clinical workflows.

Introduction

Medical Visual Question Answering (Med-VQA) represents a specialized, high-impact extension of Visual Question Answering (VQA) designed for critical healthcare environments. Unlike general VQA, Med-VQA models must accurately answer natural language questions about complex medical images—such as X-rays, CT scans, or MRIs—by synthesizing perceptual recognition with expert-level clinical reasoning. These systems hold significant promise for augmenting clinical decision-making, accelerating patient triage, and relieving strained medical resources by providing rapid, context-aware diagnostic support.

The urgency of advancing Med-VQA is underscored by a widening global gap between diagnostic imaging demands and the availability of expert interpreters. Imaging volumes continue to rise sharply—projected to increase by over 26.9% in the coming decade—while the radiology workforce grows too slowly to keep pace. Med-VQA technologies are uniquely positioned to mitigate these challenges by augmenting human expertise, expanding diagnostic reach, and automating routine analysis. Yet, the clinical deployment of Med-VQA systems requires more than high accuracy: it demands *interpretable*, *robust models* capable of trustworthy results under real-world conditions.

Med-VQA tasks are substantially more demanding than general-domain VQA. Clinical questions require multi-level reasoning—combining perceptual tasks (e.g., organ localization, modality recognition) with higher-order inference based on domain knowledge (e.g., lesion assessment, comparison to prior studies). Despite this complexity, research progress is constrained by three interconnected barriers. First, the scarcity of large, high-quality, expert-annotated datasets creates a fundamental bottleneck; annotation remains resource-intensive and existing datasets often suffer from noisy labels (Bazi et al. 2023; Hu et al. 2024). Second, reliance on pretrained general-domain vision–language models (VLMs) like CLIP (Eslami, de Melo, and Meinel 2021) and R-LLaVA (Chen et al. 2024) provides transferable skills but lacks medical specificity, resulting in brittle, non-grounded responses (Kahl et al. 2024). Third, current Med-VQA models struggle with interpretability and robustness (Kahl et al. 2024; Gai et al. 2024; Hu et al. 2024; Ishmam et al. 2025): predictions often rest on superficial correlations rather than visual grounding, performance degrades under noisy inputs, and few models provide transparency into their decision processes,failing to provide the explainability required for clinicians to verify outputs and for regulatory bodies to certify them. Together, these issues undermine trust, a critical barrier to clinical adoption.

Deploying multimodal systems in clinical workflows requires robustness to real-world distribution shifts. Imaging data may be affected by artifacts, motion blur, poor lighting, or compression, while textual queries often contain misspellings, synonym substitutions, or ambiguity. These sources of noise, compounded by workflow inconsistencies such as low-quality scans, incomplete queries, and misaligned question–image pairs, can severely undermine model reliability in practice.

In this paper, we advance Med-VQA research by presenting a novel perspective that jointly considers robustness, fusion strategies, and interpretability. Specifically, we study how state-of-the-art heavy fusion models compare with shallow fusion counterparts under multimodal perturbations spanning visual, linguistic, and cross-modal modalities. Our analysis yields three key insights. First, while heavy fusion achieves superior clean-set accuracy, its performance degrades sharply under noise. Second, shallow fusion models, though slightly less accurate in clean conditions, degrade more gracefully and exhibit more interpretable behavior. Third, robustness can be substantially improved through targeted interventions such as denoising, perturbation-aware training, and lightweight cross-modal alignment. By coupling robustness analysis with interpretability considerations, our work highlights design trade-offs that are critical for the trustworthy adoption of Med-VQA in clinical practice.

Related Work

Med-VQA: Datasets and Systems

The Med-VQA task has evolved in parallel with the development of large-scale datasets and increasingly sophisticated systems. Early benchmarks such as VQA-RAD (Lau et al. 2018) and SLAKE (Liu et al. 2021) are limited in scale and annotation diversity, constraining model generalization (Lin et al. 2023). The release of PMC-VQA featuring 227,000 expertly curated image–question–answer tuples spanning multiple modalities, marked a turning point in covering real-world clinical heterogeneity and annotation complexity (Zhang et al. 2023). Recent studies highlight the persistent annotation bottleneck in assembling such datasets, with manual verification and de-noising remaining essential for high quality. The importance of benchmarking and open resources for reproducible research is repeatedly emphasized throughout the literature (Ray et al. 2024; Hong et al. 2024).

Alongside VLM adaptation, specialized medical architectures have shown promise in capturing domain-specific nuances. Bazi et al. (2023) proposed a Transformer encoder–decoder with ViT for image encoding and autoregressive answer generation, reaching 77.27% overall accuracy on VQA-RAD. Corresponding feature fusion (CFF) with semantic attention was introduced in (Zhu et al. 2022) to emphasize clinically meaningful keywords. Collectively, these works underscore two directions in Med-VQA: (i) repurposing large-scale VLMs, and (ii) designing medical-specific architectures.

Multimodal Fusion Strategies

Fusion mechanisms form the architectural backbone of VQA systems, shaping both predictive power and interpretability. Heavy fusion strategies leverage deep cross-modal interactions. In (Zhang et al. 2024), the authors proposed OMniBAN, combining orthogonality loss with bilinear attention. Transformer-based fusion networks (**?**Huang and Hu 2025; Zhu et al. 2022; **?**) employ deep stacked multi-head cross-modal attention layers. These approaches achieve strong performance but often at the expense of interpretability.

Aggregation-based multimodal fusion methods are common in med-VQA, with shallow operations such as averaging and concatenation supporting transparency and modularity. In (Ngiam et al. 2011) the authors show that shallow fusion architectures retain strong connections to their respective modalities, preserving interpretability and enabling attribution of model outputs to specific inputs—a property critical for clinical decision support. Transparent, shallow networks—such as the interpretable prototype approach in (Singh, Stefenon, and Yow 2025)—are now designed for medical imaging to guarantee traceable, trustworthy predictions at performance levels competitive with deep, entangled models. Collectively, these studies support the assertion that, despite small trade-offs in raw predictive power, shallow fusion strategies like concatenation, gated fusion, soft attention offer crucial interpretability advantages that are increasingly valued in medicine and high-stakes AI.

Robustness and Perturbation Analysis

Robustness in VQA, especially Med-VQA, remains a critical yet understudied challenge for real-world deployment. The Visual Robustness Benchmark (Ishmam et al. 2025) systematically exposed the vulnerability of VQA models to synthetic image corruptions, while (Ramshetty, Verma, and Kumar 2023) demonstrated steep accuracy drops from more realistic, semantically-driven cross-modal attribute insertions, revealing brittleness in model representations. Complementary work highlighted that even leading video-language models struggle with both visual and language-domain perturbations, confirming that multimodal robustness issues persist broadly across domains (Schiappa et al. 2022).

Robustness evaluation for Med-VQA is only beginning to take shape. SURE-VQA introduced the first benchmark tailored to medical settings, revealing that no fine-tuning strategy consistently delivers robustness, and even unimodal text baselines can rival multimodal models—exposing biases and highlighting the need for semantic rather than token-based evaluation (Kahl et al. 2024). However, comprehensive analysis of cross-modal perturbations in Med-VQA is still lacking, representing a significant gap in the literature (Mashrur et al. 2024).

Interpretability in Medical AI

Interpretability is central to Med-VQA's clinical viability. The field relies heavily on attention visualization, semantic attribution, and knowledge graphs to explain model decisions and avoid spurious statistical correlations (Hu et al. 2024; Sharma, Purushotham, and Reddy 2021). Recent meta-studies warn that attention explanations vary in reliability and user trust depending on visualization methods (Carvallo et al. 2025). There is a pressing need for systematic, clinically validated interpretability benchmarks, especially to compare white-box (shallow) and deep black-box fusion models.

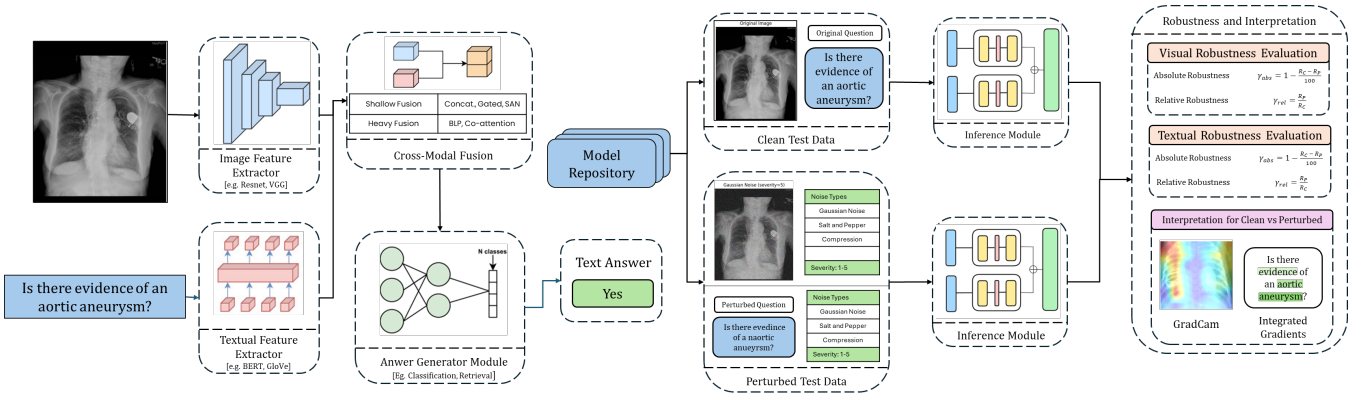

Figure 1: Overview of our Med-VQA robustness and interpretation framework. The pipeline evaluates both clean and perturbed test data under visual/textual corruptions, with robustness quantified and interpreted via GradCAM and Integrated Gradients.

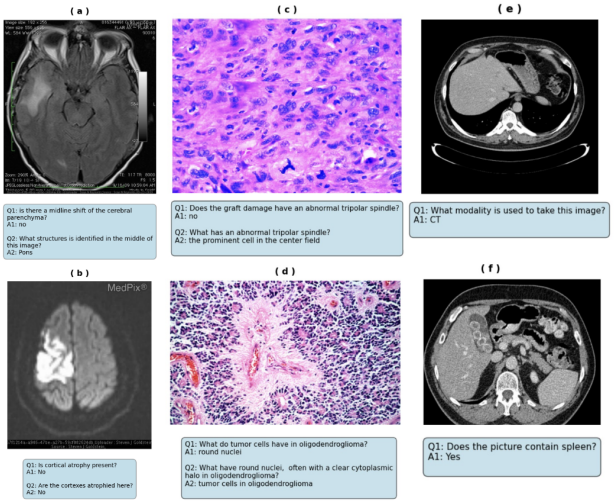

Figure 2: Sample images and paired QA examples from the VQA-RAD, PATH-VQA, and SLAKE datasets. Panels (a–b) correspond to VQA-RAD, (c–d) to PATH-VQA, and (e–f) to SLAKE.

Methodology

We present a rigorous benchmarking framework for evaluating the robustness of shallow-fusion architectures in Med-VQA. An overview of the framework is illustrated in Figure 1, highlighting the modular design spanning feature extraction, shallow-fusion strategies, perturbation pipelines, and interpretability modules. Our pipeline consists of four model variants, standardized datasets, realistic perturbation protocols, and reproducible training/evaluation implementation.

Model Variants

We formulate Med-VQA as either a discriminative classification task or a generative response task. Given a medical image v_i and a natural language question q_i, the objective is to predict the most appropriate answer. In the classification setting, the model selects the most probable answer $a \in A = \{a_1, a_2, \ldots, a_n\}$ from a predefined candidate set:

$$\hat{a} = \arg\max_{a \in A} P(a \mid v_i, q_i), \tag{1}$$

where $P(a \mid v_i, q_i)$ denotes the probability of answer a being correct given the image–question pair. In the generative setting, the task instead involves producing a free-form sequence $\hat{y} = (y_1, y_2, \ldots, y_T)$ that maximizes the conditional likelihood:

$$\hat{y} = \arg\max_{y} P(y \mid v_i, q_i). \tag{2}$$

To address both problem formulations, our approach adopts a **shallow-fusion architecture**, a widely used paradigm where visual and textual features are encoded independently and then concatenated into a joint representation. The fused embedding is passed to the answering head, which takes one of two forms: (i) a retrieval/classification head for discriminative answering, or (ii) a generative decoder for free-form synthesis. This design provides a flexible framework that allows systematic comparison of model components while ensuring robustness and interpretability. We specifically focus on shallow-fusion architectures not only for their computational efficiency but also because their modularity is a prerequisite for explainable AI (XAI), allowing for clearer attribution of predictions to either the visual or textual modality. This transparency is essential for debugging failure modes in safety-critical systems. We evaluate four representative model variants spanning different paradigms. A summary of these models, including their encoders, answering heads, and trainable parameters, is presented in Table 1.

Datasets

We evaluate our methods on three public medical VQA datasets spanning radiology, pathology, and knowledge-enhanced clinical imaging. Figure 2 illustrates a couple of sample image–QA pairs from each dataset and Table 2 summarizes dataset statistics.

VQA-RAD (Lau et al. 2018) is a foundational radiology VQA dataset featuring images from MedPix. It provides a

Model	Visual Encoder	Text Encoder	Answering Method	Trainable Params
ResNet152 + BERT	ResNet-152	BERT-base	Hybrid: Classification (closed), Retrieval (open)	172M
VGG19 + LSTM	VGG-19	LSTM	Generation (Seq2Seq decoder)	180M
ResNet50 + BiLSTM	ResNet-50	BiLSTM	Classification (MLP head)	28M
Faster-RCNN + GRU	Faster-RCNN	GRU	Generation (LSTM-based head)	158M

Table 1: Overview of shallow-fusion model variants, including visual and text encoders, answering paradigms, and the scale of trainable parameters.

Dataset	Images	Open Ended Qs	Close Ended Qs
VQA-RAD	315	941	1297
SLAKE	642	3226	1693
PATH-VQA	4998	16,465	16,334

Table 2: Summary of Med-VQA datasets showing the number of images and distribution of open and closed-ended questions.

benchmark for clinically-focused reasoning by pairing images with clinician-validated, manually-authored questions. The dataset includes both open-ended and closed (yes/no or short factual) questions that address critical clinical tasks such as lesion localization, anatomical identification, and diagnosis.

PATH-VQA (He et al. 2020) focuses on digital pathology, using images from textbooks and archives. It contains a large collection of verified question-answer pairs ($\sim$32K) that test histological and pathological reasoning. Queries often require interpreting stains and identifying cellular structures, tissue types, morphological abnormalities, and disease-related pathology.

SLAKE (Liu et al. 2021) is a unique radiology VQA dataset that goes beyond pure visual-textual understanding. It is notable for its bilingual support (English/Chinese) and for enriching VQA with semantic labels and ontology-based knowledge. This structure allows for an explicit evaluation of models' ability to perform knowledge-enhanced reasoning, making it a more complex challenge than purely vision-based tasks.

Fusion Strategies

One of the most critical design decisions in vision-language models for medical VQA is the fusion strategy that combines visual and textual representations. Heavy (deep) fusion architectures enable extensive multimodal interaction by incorporating cross-modal attention at multiple layers of the network. While this approach yields rich cross-modal representations, it comes with substantial computational demands and a large parameter footprint. In contrast, shallow fusion keeps the visual and textual encoders independent, merging their outputs only once at the representation level before passing them to task-specific heads. This paradigm strikes a balance by offering computational efficiency while still capturing complementary information across modalities.

To better understand the trade-offs between these strategies, we focus on shallow fusion and evaluate three representative mechanisms across five model variants:

Concatenation-Based Fusion The most straightforward fusion approach directly concatenates visual and textual embeddings, followed by learnable projection:

$$f_{\text{cat}}(v, q) = \text{MLP}(\sigma(W_{\text{cat}}[v; q] + b)) \tag{3}$$

where $v \in \mathbb{R}^{d_v}$ denotes the visual embedding extracted from the image encoder, $q \in \mathbb{R}^{d_q}$ represents the textual embedding produced by the question encoder, and $[v; q]$ indicates vector concatenation. The weight matrix W_{cat} and bias b parameterize a linear transformation, $\sigma(\cdot)$ is a non-linear activation function, and the MLP maps the fused representation into a joint feature space.

Cross-Modal Gated Fusion To enable adaptive weighting between modalities, cross-modal gated fusion introduces a gating mechanism that learns how much each modality should contribute to the final representation. Formally:

$$\tilde{v} = W_v v, \quad \tilde{q} = W_q q \tag{4}$$

$$g = \sigma(W_g[\tilde{v}; \tilde{q}] + b_g) \tag{5}$$

$$f_{\text{gate}}(v, q) = \text{MLP}\big(g \odot \tilde{v} + (1 - g) \odot \tilde{q}\big), \tag{6}$$

where $v \in \mathbb{R}^{d_v}$ and $q \in \mathbb{R}^{d_q}$ are the visual and textual embeddings, respectively. $W_v \in \mathbb{R}^{d \times d_v}$ and $W_q \in \mathbb{R}^{d \times d_q}$ project them into a shared d-dimensional space, while $[\tilde{v}; \tilde{q}]$ denotes concatenation. The gating vector $g \in \mathbb{R}^d$, computed via a sigmoid activation $\sigma(\cdot)$, adaptively balances the contribution of each modality. Finally, an MLP maps the gated combination into the joint feature space.

Additive Fusion (Element-wise Sum) Another simple yet effective approach is additive fusion, where visual and textual embeddings are projected into the same dimension and combined via element-wise summation:

$$f_{\text{add}}(v, q) = \text{MLP}(\sigma(W_v v + W_q q + b)) \tag{7}$$

where $W_v \in \mathbb{R}^{d \times d_v}$ and $W_q \in \mathbb{R}^{d \times d_q}$ are linear projection matrices that map the visual embedding $v \in \mathbb{R}^{d_v}$ and textual embedding $q \in \mathbb{R}^{d_q}$ into a shared d-dimensional space, $b \in \mathbb{R}^d$ is a bias term, and $\sigma(\cdot)$ is a non-linear activation.

Perturbation Protocol

To comprehensively assess robustness, we adapt perturbation frameworks from prior multimodal studies (Schiappa et al. 2022) to the medical VQA domain. Our design introduces both visual and textual perturbations that mimic realistic distribution shifts as well as adversarial noise sources. Models are evaluated under a zero-shot robustness setting, meaning they are tested directly on perturbed inputs without any fine-tuning or adaptation.

Visual Perturbations. The visual perturbations are organized into four categories: **Noise**, **Blur**, **Camera**, and **Digital**. Each perturbation type is applied at severity scales from 1 to 5, with higher values introducing progressively more difficult distortions.

- **Noise:** Impulse noise (salt-and-pepper artifacts), Gaussian noise, and speckle noise.
- **Blur:** Zoom blur, defocus blur, and motion blur.
- **Camera:** Static rotations, small random rotations, and pixel-level translations.
- **Digital:** JPEG compression artifacts, simulating lossy image storage and transmission.

Textual Perturbations. In parallel, we design a diverse set of textual perturbations to stress-test language robustness. These fall into three categories: **Medical**, **Natural**, and **Synthetic**.

- **Medical:** Includes medically-grounded modifications such as replacing organ/pathology/modality terms, introducing domain abbreviations, or injecting distractors and irrelevant findings.
- **Natural:** Covers everyday perturbations like abbreviation expansion/compression, negation toggles, paraphrasing, character-level edits, token swaps, random drops, and additive filler phrases.
- **Synthetic:** Heuristic structure-based perturbations, including noun/verb dropping, synonym/antonym replacement, and random substitutions of noun-like tokens.

Evaluation Metrics

We evaluate performance across three categories of questions. For closed-ended VQA, we report overall accuracy as well as accuracy on yes/no questions. For open-ended VQA, we use Exact Match (EM) along with standard generation metrics, including BLEU-4, METEOR, and ROUGE-L. Finally, we summarize overall performance using a weighted accuracy across all question types.

To quantify robustness, we adopt two complementary metrics: absolute robustness and relative robustness (Hendrycks and Dietterich 2019; Schiappa et al. 2022). Given a trained model f, let R_c^f denote the retrieval score on the clean test set, and R_p^f the retrieval score under a perturbation p. The absolute robustness for perturbation p is defined as:

$$\gamma_p^a = 1 - \frac{R_c^f - R_p^f}{100}. \tag{8}$$

For visual perturbations, we report aggregated robustness by averaging across severity levels. For textual perturbations, aggregation is performed across perturbation subtypes rather than severities. To account for variation in baseline model accuracy, we compute relative robustness as:

$$\gamma_p^r = 1 - \frac{R_c^f - R_p^f}{R_c^f}. \tag{9}$$

Both γ_p^a and γ_p^r typically range between 0 and 1, where 0 indicates no robustness (complete performance degradation) and 1 corresponds to perfect robustness (no drop under perturbation).

Experimental Setup

All models are trained exclusively on clean data and evaluated under both clean and perturbed test conditions. To optimize performance, we employ Optuna for hyperparameter tuning (learning rate, weight decay, dropout, and LoRA rank), with search spaces defined individually for each model variant. Training follows a unified protocol: AdamW optimizer with cosine learning rate decay, weight decay of 1×10^{-4}, gradient norm clipping at 1.0, and mixed-precision training (AMP/FP16). Encoder learning rates are fixed at 1×10^{-5}, while task-specific heads use 1×10^{-4}. Models are trained with a mini-batch size of 32 (accumulated to an effective 128), early stopping with patience 10, and a maximum of 50 epochs. For reproducibility, we fix random seeds (42) across all frameworks. All code, configuration files, and training logs (via W&B) will be released publicly.

Results and Analysis

Clean Performance

Our clean performance evaluation reveals a fundamental trade-off between models optimized for classification versus those for text generation. As shown in Table 3, the ResNet152+BERT model consistently dominates on factual, closed-ended queries, achieving the highest ‘Closed‘ and ‘Yes/No‘ accuracies across all three datasets (72.1–88.4%). This highlights the strength of combining deep, rich visual features from ResNet152 with the powerful contextual language understanding of BERT, making it ideal for answering direct questions.

In stark contrast, the ResNet50+BiLSTM model excels at open-ended, explanatory tasks. It consistently achieves the best ‘Open Exact‘ match scores and superior language-generation metrics like ‘BLEU-4‘ and ‘METEOR‘, indicating its proficiency in producing coherent, well-formed textual answers. The other models, VGG19+LSTM and Faster-RCNN+GRU, underperform significantly, with VGG19+LSTM showing particular weakness (11.5% ‘Open Exact‘).

Visual Robustness

When subjected to common visual corruptions, ResNet152+BERT emerges as the indisputable leader in robustness. As detailed in Table 4, it maintains high and stable relative scores (> 0.8) across all major perturbation categories—noise, blur, camera, and digital—indicating strong resilience to real-world artifacts.

Conversely, ResNet50+BiLSTM shows catastrophic failure under specific conditions. Its performance plummets to near-zero (‘0.06 ± 0.12‘) under ‘Impulse‘ noise, a direct reflection of its extreme vulnerability to even minor random visual perturbations.

Textual Robustness

In contrast to the visual domain, ResNet50+BiLSTM demonstrates superior resilience to text perturbations, outperforming ResNet152+BERT on 20 out of 26 perturbation types. Both models show greater robustness to textual changes than to visual ones, with near-perfect scores (1.000)

Table 3: Clean test performance on VQA-RAD, SLAKE, and PATH-VQA datasets for all model variants.

Dataset	Model	Closed	Yes/No	Open Exact	BLEU-4	METEOR	ROUGE-L	Overall Acc
VQA-RAD	ResNet152 + BERT	**72.1**	**76.1**	40.2	0.159	0.351	0.446	**59.8**
	ResNet50 + BiLSTM	67.1	70.9	**41.4**	**0.175**	**0.352**	**0.476**	57.1
	VGG19 + LSTM	59.8	64.4	11.5	0.042	0.113	0.203	41.1
	Faster-RCNN + GRU	66.7	72.1	23.2	0.102	0.207	0.278	49.8
SLAKE	ResNet152 + BERT	**82.0**	**83.5**	33.5	0.118	0.238	0.318	**52.0**
	ResNet50 + BiLSTM	77.0	78.4	**36.8**	**0.126**	**0.246**	0.314	49.5
	VGG19 + LSTM	70.2	72.5	12.3	0.048	0.119	0.218	38.7
	Faster-RCNN + GRU	75.1	77.3	23.6	0.092	0.183	**0.322**	45.2
PATH-VQA	ResNet152 + BERT	**88.4**	**88.4**	24.3	0.069	0.176	0.283	56.4
	ResNet50 + BiLSTM	84.5	86.1	**29.7**	**0.117**	**0.323**	**0.609**	**59.7**
	VGG19 + LSTM	52.3	55.1	14.6	0.058	0.141	0.227	36.5
	Faster-RCNN + GRU	82.6	82.6	16.2	0.039	0.108	0.191	49.4

Table 4: Relative robustness scores γ^r with standard deviations $\pm\sigma$ for each individual visual perturbation type across all severity levels. Model 1: ResNet152 + BERT, Model 2: ResNet50 + BiLSTM

Perturbation	Model 1	Model 2
Noise Perturbations		
Impulse	0.81 ± 0.04	0.06 ± 0.12
Gaussian	0.82 ± 0.05	0.39 ± 0.38
Speckle	0.90 ± 0.00	0.84 ± 0.01
Blur Perturbations		
Zoom	0.88 ± 0.02	0.60 ± 0.04
Defocus	0.87 ± 0.02	0.74 ± 0.10
Motion	0.88 ± 0.01	0.57 ± 0.05
Camera Perturbations		
Static	0.87 ± 0.02	0.53 ± 0.03
Rotation	0.92 ± 0.01	0.86 ± 0.01
Translation	0.84 ± 0.02	0.52 ± 0.04
Digital Perturbations		
JPEG	0.89 ± 0.01	0.52 ± 0.03

on common operations like abbreviation expansion and synonym replacement.

As seen in Table 5, ResNet50+BiLSTM consistently achieves higher composite scores across medical, natural, and synthetic perturbations.

Fusion Strategy Comparison

We evaluated three shallow fusion strategies—**gated**, **concatenated**, and **additive**—under various noise perturbations. Gated fusion consistently demonstrates the highest robustness, maintaining stable performance across impulse, Gaussian, and speckle noise, followed closely by additive fusion. Concatenated fusion, in contrast, is more sensitive, showing pronounced degradation under impulse and Gaussian noise.

Explainable Failure Analysis: Building Clinical Trust Through Transparency

To ensure safe deployment of Medical Visual Question Answering (Med-VQA) systems in clinical workflows, accuracy must be complemented by interpretability. Clinicians require insight into a model's decision rationale to assess reliability and identify failure modes. This subsection employs two complementary explainable AI (XAI) methods, Gradient-weighted Class Activation Mapping (Grad-CAM) for visual grounding and Integrated Gradients (IG) for linguistic attribution, to interpret model behavior under visual and textual perturbations. These analyses expose how multimodal components respond to noise, offering quantitative insights into robustness and transparency.

Table 5: Composite robustness scores (mean $\pm$ std of γ^r values) for text perturbations by category. Model 1: ResNet152 + BERT, Model 2: ResNet50 + BiLSTM

Perturbation	Model 1	Model 2
Medical Perturbations		
Medical Replace	0.96 ± 0.01	$\mathbf{0.99 \pm 0.01}$
Negation Flip	0.99 ± 0.05	$\mathbf{1.00 \pm 0.01}$
Adversarial Typo	0.95 ± 0.07	0.95 ± 0.04
Structural Reorder	$\mathbf{1.00 \pm 0.00}$	1.00 ± 0.00
Distractor Injection	0.88 ± 0.09	0.85 ± 0.07
Drop Critical NN	0.95 ± 0.05	$\mathbf{0.99 \pm 0.02}$
Abbreviation Compress Med	$\mathbf{0.99 \pm 0.01}$	0.99 ± 0.01
Natural Perturbations		
Abbreviation Expand	1.00 ± 0.00	1.00 ± 0.00
Abbreviation Compress	1.00 ± 0.00	$\mathbf{1.00 \pm 0.00}$
Negation Flip Natural	$\mathbf{1.01 \pm 0.05}$	1.00 ± 0.02
Paraphrase	0.96 ± 0.04	$\mathbf{1.00 \pm 0.01}$
Change Character	0.90 ± 0.10	$\mathbf{0.96 \pm 0.03}$
Swap Text	0.89 ± 0.06	$\mathbf{0.98 \pm 0.02}$
Add Text	$\mathbf{1.01 \pm 0.05}$	1.00 ± 0.01
Drop First	0.92 ± 0.08	$\mathbf{0.98 \pm 0.01}$
Drop Last	0.93 ± 0.05	$\mathbf{0.96 \pm 0.03}$
Drop First and Last	0.92 ± 0.08	$\mathbf{0.94 \pm 0.00}$
Random Drop	0.96 ± 0.02	$\mathbf{0.97 \pm 0.01}$
Synthetic Perturbations		
Drop NN	0.96 ± 0.05	$\mathbf{0.99 \pm 0.02}$
Drop VB	0.99 ± 0.04	$\mathbf{0.99 \pm 0.00}$
Drop VB+NN	0.96 ± 0.06	$\mathbf{0.98 \pm 0.02}$
Random NN	0.95 ± 0.04	$\mathbf{0.99 \pm 0.01}$
Synonym Replace	0.99 ± 0.01	$\mathbf{1.00 \pm 0.00}$
Antonym Replace	0.98 ± 0.01	$\mathbf{0.99 \pm 0.01}$

Visual Grounding with Grad-CAM: Diagnosing Attention Shifts

Grad-CAM generates a localization heatmap by weighting

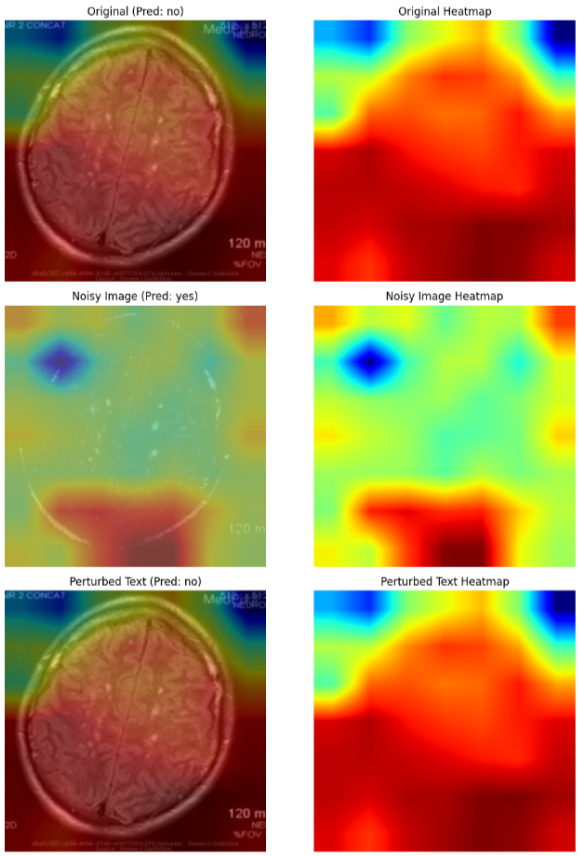

Figure 3: Grad-CAM overlays and corresponding heatmaps for clean (top), visually perturbed (middle), and textually perturbed (bottom) inputs. Visual perturbations cause diffuse and misplaced activations, whereas textual noise preserves anatomical focus.

the final convolutional feature maps with the gradients of the target prediction, thereby highlighting image regions most influential to the model's answer. This allows verification of whether attention aligns with relevant anatomical structures or drifts toward artifacts.

For the VQA-RAD query "Are the sulci abnormal?" (ground truth: *no*), Figure 3 compares Grad-CAM outputs across clean, visually perturbed, and textually perturbed inputs. On the clean image (top), attention is concentrated along the cerebral sulci, supporting a correct 'no' prediction. Under Gaussian noise (middle), the prediction flips to 'yes', and attention diffuses toward non-diagnostic regions—indicating a reliance on spurious features. By contrast, when only the question is perturbed (bottom; "are the uslci abnormal?"), the visual focus largely remains intact, and the prediction stays correct. The Intersection over Union (IoU) between clean and noisy attention maps averages **0.1934**, evidencing that visual grounding deteriorates sharply under image perturbations, while remaining relatively stable under textual noise.

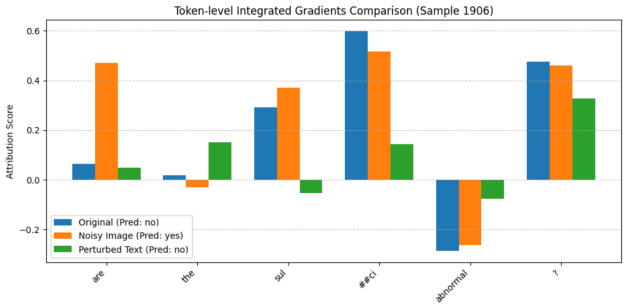

Figure 4: Integrated Gradients attribution scores for clean, visually perturbed, and textually perturbed questions. Textual misspellings retain key token importance, whereas visual corruption induces attribution dispersion.

Linguistic Reasoning with Integrated Gradients: Probing Semantic Robustness

Integrated Gradients (IG) attribute prediction influence to input tokens by integrating gradients from a baseline representation to the actual input, satisfying completeness and sensitivity. This enables examination of how the model's linguistic reasoning shifts under perturbations.

Figure 4 visualizes token-level attributions for the same question across three settings: clean, visually perturbed, and textually perturbed. In the clean case, key medical term 'sulci" receives the highest positive attributions, confirming that the model's answer is grounded in semantically relevant tokens. When the image is corrupted, attribution weights redistribute erratically, showing elevated influence for less informative tokens (e.g., "are"), consistent with visual confusion and misprediction ('yes'). However, when only the question is perturbed (misspelling "uslci"), the attribution pattern is slightly similar to the clean case, with minimal change in relative importance. Quantitatively, comparing attribution distributions yields an average Kullback–Leibler divergence of $\mathbf{D}_{KL}$ **= 9.2336** between clean and noisy textual queries, indicating moderate but controlled semantic drift—substantially lower than the attention divergence observed visually.

Quantitative Insights and Clinical Implications Cross-modal analysis reveals that visual explanations degrade more severely than textual ones under perturbations: attention overlap (IoU = 0.1934) indicates unstable visual grounding, whereas semantic divergence (D_{KL} = 9.2336) remains limited, demonstrating greater linguistic resilience. This asymmetry suggests that the model's language encoder maintains semantic integrity even when the vision branch fails to generalize under distribution shifts. Such explainable diagnostics enable fine-grained auditing of multimodal systems, helping identify which modality drives incorrect predictions and informing targeted retraining strategies. By quantifying reasoning drift, this XAI framework strengthens clinical interpretability and fosters confidence in Med-VQA as a decision-support tool.

Conclusion

This study establishes empirical baselines for adversarial robustness and interpretability in Medical Visual Question Answering (Med-VQA) through a systematic benchmark of lightweight, shallow-fusion architectures. Across VQA-RAD, PathVQA, and SLAKE datasets, four model variants—ResNet152+BERT, ResNet50+BiLSTM, VGG19+LSTM, and Faster-RCNN+GRU—were evaluated under 13 visual perturbations and 26 textual corruptions. Key findings reveal that while ResNet152+BERT delivers peak clean accuracy (88.4% on PathVQA closed-ended tasks), it suffers sharp degradation under visual noise (relative robustness 0.06 ± 0.12 for impulse noise). In contrast, ResNet50+BiLSTM exhibits superior resilience, outperforming on 20 of 26 perturbation types and excelling in open-ended generation (BLEU-4: 0.117). Fusion strategy comparisons highlight gated mechanisms as most robust, with Grad-CAM and Integrated Gradients analyses exposing attention realignments (IoU: 0.1934 under noise) that enable transparent error attribution.

These results underscore practical applications for deployable Med-VQA in resource-constrained clinical workflows: shallow fusion reduces computational demands while preserving interpretability, facilitating rapid triage in radiology and pathology settings. By quantifying multimodal vulnerabilities, the framework enhances clinician trust, enabling seamless integration into hospital systems for augmented diagnostics and reduced interpretive delays.

Future work should address identified gaps using state-of-the-art techniques. Perturbation-specific errors, such as impulse noise brittleness, can be mitigated through adversarial training with diffusion-based denoising or robust optimization via certified defenses like randomized smoothing. Extending interpretability to heavy-fusion models—such as Transformer-based VLMs—requires hybrid XAI pipelines that disentangle cross-modal attention, potentially via layer-wise relevance propagation or knowledge-graph alignment. Such advancements will yield certifiable Med-VQA systems resilient to real-world distribution shifts, advancing safe and trustworthy AI adoption in healthcare.

References

Bazi, Y.; Al Rahhal, M. M.; Bashmal, L.; and Zuair, M. 2023. Vision-Language Model for Visual Question Answering in Medical Imagery. *Bioengineering*, 10(3): 380.

Carvallo, A.; Parra, D.; Brusilovsky, P.; Valdivieso, H.; Rada, G.; Donoso, I.; and Araujo, V. 2025. User Perception of Attention Visualizations: Effects on Interpretability Across Evidence-Based Medical Documents. arXiv:2508.10004.

Chen, X.; Lai, Z.; Ruan, K.; Chen, S.; Liu, J.; and Liu, Z. 2024. R-LLaVA: Improving Med-VQA Understanding Through Visual Region of Interest. arXiv:2410.20327.

Eslami, S.; de Melo, G.; and Meinel, C. 2021. Does CLIP Benefit Visual Question Answering in the Medical Domain as Much as It Does in the General Domain? arXiv:2112.13906.

Gai, X.; Zhou, C.; Liu, J.; Feng, Y.; Wu, J.; and Liu, Z. 2024. MedThink: Explaining Medical Visual Question Answering via Multimodal Decision-Making Rationale. arXiv:2404.12372.

He, X.; Zhang, Y.; Mou, L.; Xing, E.; and Xie, P. 2020. PathVQA: 30,000+ Questions for Medical Visual Question Answering. arXiv:2003.10286.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. arXiv:1903.12261.

Hong, X.; Song, Z.; Li, L.; Wang, X.; and Liu, F. 2024. BESTMVQA: A Benchmark Evaluation System for Medical Visual Question Answering. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 435–451. Springer.

Hu, X.; Gu, L.; Kobayashi, K.; Liu, L.; Zhang, M.; Harada, T.; Summers, R. M.; and Zhu, Y. 2024. Interpretable Medical Image Visual Question Answering via Multi-Modal Relationship Graph Learning. *Medical Image Analysis*, 97: 103279.

Huang, C.; and Hu, Z. 2025. A Multimodal Transformer-Based Visual Question Answering Method Integrating Local and Global Information. *PLoS ONE*, 20(7): e0324757.

Ishmam, M. F.; Tashdeed, I.; Saadat, T. A.; Ashmafee, M. H.; Kamal, A. R. M.; and Hossain, M. A. 2025. Visual Robustness Benchmark for Visual Question Answering (VQA). In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 6623–6633.

Kahl, K.-C.; Erkan, S.; Traub, J.; Lüth, C. T.; Maier-Hein, K.; Maier-Hein, L.; and Jaeger, P. F. 2024. SURE-VQA: Systematic Understanding of Robustness Evaluation in Medical VQA Tasks. arXiv:2411.19688.

Lau, J. J.; Gayen, S.; Ben Abacha, A.; and Demner-Fushman, D. 2018. A Dataset of Clinically Generated Visual Questions and Answers about Radiology Images. *Scientific Data*, 5(1): 1–10.

Lin, Z.; Zhang, D.; Tao, Q.; Shi, D.; Haffari, G.; Wu, Q.; He, M.; and Ge, Z. 2023. Medical Visual Question Answering: A Survey. *Artificial Intelligence in Medicine*, 143: 102611.

Liu, B.; Zhan, L.-M.; Xu, L.; Ma, L.; Yang, Y.; and Wu, X.-M. 2021. SLAKE: A Semantically-Labeled Knowledge-Enhanced Dataset for Medical Visual Question Answering. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 1650–1654.

Mashrur, A.; Luo, W.; Zaidi, N. A.; and Robles-Kelly, A. 2024. Robust Visual Question Answering via Semantic Cross Modal Augmentation. *Computer Vision and Image Understanding*, 238: 103862.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal Deep Learning. In *International Conference on Machine Learning (ICML)*, 689–696.

Ramshetty, S.; Verma, G.; and Kumar, S. 2023. Cross-Modal Attribute Insertions for Assessing the Robustness of Vision-and-Language Learning. arXiv:2306.11065.

Ray, S.; Gupta, K.; Kundu, S.; Kasat, P. A.; Aditya, S.; and Goyal, P. 2024. ERVQA: A Dataset to Benchmark the

Readiness of Large Vision Language Models in Hospital Environments. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 15594–15608. Association for Computational Linguistics.

Schiappa, M.; Vyas, S.; Palangi, H.; Rawat, Y.; and Vineet, V. 2022. Robustness Analysis of Video-Language Models Against Visual and Language Perturbations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 34405–34420.

Sharma, D.; Purushotham, S.; and Reddy, C. K. 2021. MedFuseNet: An Attention-Based Multimodal Deep Learning Model for Visual Question Answering in the Medical Domain. *Scientific Reports*, 11(1): 19826.

Singh, G.; Stefenon, S. F.; and Yow, K.-C. 2025. The Shallowest Transparent and Interpretable Deep Neural Network for Image Recognition. *Scientific Reports*, 15(1): 13940.

Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering. arXiv:2305.10415.

Zhang, Z.; Wang, J.; Qin, Z.; Zhu, R.; and Gong, X. 2024. Efficient Bilinear Attention-Based Fusion for Medical Visual Question Answering. arXiv:2410.21000.

Zhu, H.; He, X.; Wang, M.; Zhang, M.; and Qing, L. 2022. Medical Visual Question Answering via Corresponding Feature Fusion Combined with Semantic Attention. *Mathematical Biosciences and Engineering*, 19(10): 10192–10212.