

## Article

# Robustness of Large Vision Language Model Features Under Wireless Channel Degradation for Medical Visual Question Answering

Merve Güllü <sup>1,2,\*</sup> , Necaattin Barışçı <sup>3</sup> <sup>1</sup> R&D Department, Türk Telekom, Ankara 06080, Türkiye<sup>2</sup> Graduate School of Natural and Applied Sciences, Gazi University, Ankara 06560, Türkiye<sup>3</sup> Department of Computer Engineering, Gazi University, Ankara 06560, Türkiye; nbarisci@gazi.edu.tr

\* Correspondence: merve.gullu@turktelekom.com.tr

## Abstract

Deploying medical visual question answering (VQA) systems over wireless networks introduces a fundamental challenge: channel-induced image degradation may corrupt the visual representations extracted by large vision-language models (VLMs), leading to unreliable diagnostic decisions. We investigate the robustness of frozen LLaVA-1.6, BLIP-2, and BioViL-T hidden-state features under additive white Gaussian noise (AWGN), Rayleigh fading, and six combined JPEG-compression-plus-channel conditions (quality factors  $q \in \{20, 50, 70\}$ ) across signal-to-noise ratios (SNRs) from  $-5$  to  $+20$  dB. A lightweight MLP classifier is trained exclusively on clean features and evaluated on channel-degraded features, enabling controlled analysis of representation robustness without retraining. We introduce the Feature Robustness Score (FRS), defined as the difference between cosine similarity and normalized L2 drift of clean versus degraded features, together with a validation-set FRS threshold analysis as a label-free retraining criterion. A wavelet sub-band energy analysis further characterizes the spectral distribution of channel-induced feature drift. Experiments on PathVQA and VQA-RAD reveal four key findings: (1) LLaVA-1.6 features maintain cosine similarity above 0.98 across all eight channel conditions and all SNR levels, with statistically significant MLP gains at every tested point ( $p < 0.05$ , McNemar's test); (2) BLIP-2 and BioViL-T features are less stable but still support consistent MLP improvements, with BioViL-T performing competitively on VQA-RAD, suggesting domain alignment matters; (3) JPEG compression quality ( $q = 20, 50, 70$ ) has negligible impact on feature drift, establishing VLM features as JPEG quality-invariant; and (4) wavelet analysis confirms that channel noise primarily affects high-frequency detail bands while preserving low-frequency semantic content.

**Keywords:** medical visual question answering; large vision-language models; LLaVA; BLIP-2; BioViL-T; wireless channel; AWGN; Rayleigh fading; JPEG compression; feature robustness score; wavelet analysis; PathVQA; VQA-RAD



Academic Editor: Rui Araújo

Received: 23 May 2026

Revised: 14 June 2026

Accepted: 23 June 2026

Published: 27 June 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.

## 1. Introduction

Visual question answering (VQA) in the medical domain has direct applications in clinical decision support, remote diagnostics, and automated report generation [1,2]. Binary medical VQA, answering clinical questions about medical images with yes or no, is particularly relevant for screening and triage in resource-constrained environments.

The emergence of large vision-language models (VLMs) such as LLaVA-1.6 [3], BLIP-2 [4], and domain-specialized models such as BioViL-T [5] has significantly advanced the state of the art in medical VQA. In real-world deployments, remote hospitals, ambulances, or field settings, medical images must often be transmitted over wireless networks before processing. Wireless channels introduce degradation through additive noise, multipath fading, and lossy compression, potentially corrupting the visual information that VLMs rely on for inference.

A critical but underexplored question is: how robust are the internal representations of frozen VLMs to channel-induced image degradation, and how do architectural differences and domain pre-training affect this robustness? If hidden-state features remain stable despite image corruption, classifiers trained on clean features may generalize to channel-degraded conditions without retraining.

We address this question through a systematic empirical study comparing three frozen VLMs under eight channel conditions: AWGN, Rayleigh fading, and JPEG-compressed variants at quality factors  $q \in \{20, 50, 70\}$  combined with both channel types. The JPEG variants simulate the practical source-channel coding pipeline where lossy compression precedes wireless transmission. To quantify feature robustness, we introduce the Feature Robustness Score (FRS), combining cosine similarity and normalized L2 drift. A wavelet sub-band energy analysis provides a spectral interpretation of feature drift, and a validation-set FRS threshold analysis serves as a label-free retraining criterion.

Our main contributions are as follows:

1. We provide the first systematic comparison of frozen LLaVA-1.6, BLIP-2, and BioViL-T feature robustness under eight wireless channel conditions for binary medical VQA on PathVQA and VQA-RAD.
2. We show that LLaVA-1.6 features maintain cosine similarity above 0.98 across all conditions and SNR levels, yielding statistically significant MLP gains at every tested point ( $p < 0.05$ ) on PathVQA ( $n = 841$ ).
3. We establish that JPEG compression quality ( $q = 20, 50, 70$ ) has negligible effect on VLM feature drift, a novel finding with direct implications for practical source-channel coding system design.
4. We show that BioViL-T performs competitively on VQA-RAD (clean MLP accuracy 76.56%,  $\Delta = +28.12\%$ ,  $p = 0.0027$ ) but is outperformed on PathVQA, suggesting that domain alignment between pre-training data and the target imaging modality is a key factor in robustness.
5. We propose the FRS metric, a wavelet sub-band energy analysis, and a validation-set threshold analysis as a practical label-free retraining criterion.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the system model and methodology. Section 4 presents experimental results. Section 5 provides analysis and discussion. Section 6 concludes the paper.

## 2. Related Work

### 2.1. Medical Visual Question Answering

Medical VQA (Med-VQA) seeks to combine radiological or pathological imagery with natural-language questions to produce diagnostic-level answers. The VQA-RAD benchmark [1] introduced a radiology-oriented dataset of clinician-generated questions, each paired with DICOM images and expert answers, thereby establishing a testbed for retrieval- and reasoning-based Med-VQA architectures. Subsequently, PathVQA [2] extended this paradigm to pathology whole-slide images and focused on binary yes/no questions that closely mirror routine screening workflows in histopathology. Recent approaches exploit

pretrained VLMs via direct inference or lightweight fine-tuning of task-specific adapters, enabling rapid adaptation to clinical domains without full retraining [6–9].

A key challenge in Med-VQA is the tendency of models to rely on language shortcuts rather than genuine visual reasoning—a problem known as language bias. Wang et al. [10] directly target this issue by proposing the Global Visual Information Intervention (GVII) model, which introduces a Global Visual Information Branch to amplify image contributions relative to question dominance, together with a Forward Compensation Branch to rebalance multimodal features. Evaluated on PathVQA, GVII achieves a 2.6% accuracy improvement over prior state of the art, confirming that robust visual representations, not just stronger language models, are critical for pathology VQA. Cycle-VQA [11] addresses reasoning instability from a complementary angle by establishing bidirectional information flow among questions, answers, and visual features through cycle consistency, reporting state-of-the-art results across multiple Med-VQA benchmarks, including medical ultrasound image classification.

Bridging the semantic gap between large pre-trained VLMs and task-specific Med-VQA datasets remains an active research problem. Lu et al. [12] propose the Dynamic Semantic-Adaptive Prompting (DSAP) framework, incorporating Semantic Alignment Prompting (SAP) and Dynamic Question-Aware Prompting (DQAP) to adapt frozen VLMs without retraining, achieving a 1.9% average accuracy improvement across VQA-RAD, SLAKE, and PathVQA benchmarks. At the efficiency frontier, CoMed-TR [13] proposes a large–small model collaboration framework using a Task-Aware Mixture-of-Experts module alongside a relation knowledge distillation strategy, achieving 3.3% and 1.1% accuracy gains on VQA-RAD and SLAKE, respectively, with only approximately 25 million trainable parameters.

Alongside these adapter-based and prompt-tuning strategies, fully fine-tuned task-specific architectures remain competitive on in-distribution benchmarks. MedFuseT [14] exemplifies this paradigm by integrating EfficientNet for image feature extraction, BERT+LSTM for question encoding, and a Transformer fusion module for cross-modal interaction, achieving 83.91% on VQA-RAD and 93.18% in PathVQA, the same two benchmarks used in the present study. Similarly, Zhang et al. [15] proposed HiCA-VQA, a hierarchical modeling framework with cross-attention fusion that achieves state-of-the-art performance on the Rad-Restruct benchmark, an extension of VQA-RAD with hierarchical question–answer pairs. Their hierarchical prompting and answer decoder modules are specifically designed to handle fine-grained medical questions, complementing the architectural innovations of MedFuseT. While such fully fine-tuned architectures attain strong clean-data accuracy, they require complete retraining whenever the input distribution shifts. Under wireless channel degradation, where the image statistics at inference time deviate systematically from those seen during training, this retraining requirement becomes a practical barrier that frozen-backbone pipelines are specifically designed to avoid.

Beyond task-specific architectures, the importance of rigorous validation against reference standards is a principle shared across practical medical imaging pipelines. For instance, freehand three-dimensional ultrasound has been validated against MRI for muscle volume measurement using agreement-based metrics, demonstrating that practical imaging tools can match reference-standard outputs when properly calibrated [16]. An analogous validation logic underlies our FRS framework, which assesses whether channel-degraded VLM features remain equivalent to clean reference features using cosine similarity and statistical agreement tests.

At the system level, the efficient serving of frozen-backbone pipelines is an equally important practical concern for real-world deployment. Recent surveys on ML inference serving highlight that system-level optimizations—including batching strategies, resource

scheduling, and latency-accuracy trade-offs—account for a substantial share of deployment costs and must be co-designed with model-level architectural choices [17].

These developments reinforce a growing consensus that frozen or lightly tuned VLM backbones followed by small trainable heads can match or exceed fully fine-tuned monolithic architectures on clinical tasks. This design principle underlies our frozen-VLM + MLP pipeline.

Beyond task-specific Med-VQA, recent work has explored zero-shot multimodal reasoning through agent collaboration. Alsolami et al. [18] proposed MiRA (Mixture-of-Reasoning Agents), a framework that decomposes reasoning across three specialized agents (Visual Analyzing, Text Comprehending, Judge) using pretrained language models without any supervised training. On the ScienceQA benchmark, MiRA achieves 96.0% accuracy, surpassing all zero-shot methods and outperforming few-shot GPT-4o models by more than 18% on image-based questions. Although MiRA has not been evaluated under channel degradation or on medical images, its agent-based, zero-shot paradigm offers a complementary direction to our frozen-feature approach for future robust system design.

Crucially, none of the works reviewed above [10–14] characterize how the frozen VLM feature space degrades under wireless channel impairments, the central question of the present study.

## 2.2. Wireless Transmission of Medical Images

In clinical wireless systems, diagnostic images are typically compressed and transmitted over bandwidth-limited channels subject to stochastic impairments. Additive white Gaussian noise (AWGN) and Rayleigh fading are canonical channel models that capture thermal noise and multipath-induced signal attenuation, respectively, and underpin the physical-layer analysis of tele-radiology and remote pathology systems. Stochastic channel impairments and their impact on transmitted signal quality are a fundamental concern across diverse wireless communication paradigms. In underwater optical wireless channels, for instance, turbulence-induced fading and scintillation have been shown to severely degrade link performance [19,20], and multi-wavelength transmission strategies have been proposed to mitigate these effects by exploiting the distinct propagation characteristics of different wavelengths [19,20]. While the physical mechanisms differ from those of RF wireless channels, the analytical challenge of characterizing how stochastic amplitude fluctuations propagate through the signal processing pipeline—and ultimately affect downstream task performance—is common to all wireless transmission scenarios, motivating the systematic SNR-parameterized evaluation framework adopted in the present study. In practice, JPEG compression is commonly applied prior to transmission to reduce bandwidth and storage overhead, leading to compound degradation when JPEG-quantized bitstreams are further corrupted by channel noise [21]. Recent surveys on wireless healthcare networks emphasize that lossy transmission of medical images must be evaluated not only via pixel-level metrics (e.g., PSNR, SSIM) but also with respect to clinical decision-level errors, including misclassification and misdiagnosis rates [22].

Intelligent communication systems highlight that VQA-enabled architectures can support automated visual analytics in networked environments [23,24], but typically treat the image and VQA model as black-box components rather than explicitly analyzing how compressed or noisy image features propagate through the VLM pipeline. Most recently, Liu et al. [25] proposed a goal-oriented semantic communication framework for edge-enabled wireless VQA, achieving up to 49% accuracy improvement under AWGN by transmitting scene graph triplets rather than raw images. However, this framework operates on general-domain images with task-specific trained encoders and does not examine how frozen large VLM representations degrade under wireless channel conditions.

In sum, while the impact of wireless impairments on raw image quality and simple classifiers is relatively well understood, the robustness of frozen VLM features in the context of medical VQA under AWGN, Rayleigh, and JPEG-plus-channel conditions remains largely unexplored.

### 2.3. Robustness of Deep Representations

The robustness of deep representations to input perturbations has been extensively studied in both computer vision and NLP [26–30]. Adversarial robustness investigates worst-case, imperceptible perturbations that maximally degrade model performance, whereas robustness to natural distribution shifts and common corruptions is often assessed via benchmarks such as ImageNet-C [31], which systematically applies multiple corruption types (e.g., noise, blur, weather artifacts) at several severity levels and reports the mean Corruption Error (mCE). Subsequent studies have extended this paradigm to more challenging “hard” synthetic images and to multimodal models, showing that modern vision backbones and CLIP-style models exhibit significant accuracy drops under diverse corruptions, even when they perform well on clean data [32]. Recent work on medical vision-language models (MVLMS) further reveals that existing MVLMS display severe degradation under corruption across five medical imaging modalities, underscoring the need for robust adaptation strategies such as low-rank adaptation and few-shot tuning [33]. At the same time, certification-oriented methods such as PromptSmooth exploit prompt-learning frameworks to provide provable robustness guarantees for medical VLMs under Gaussian noise, demonstrating that constrained prompt-based training can achieve high certified accuracy even when the encoder is frozen [34].

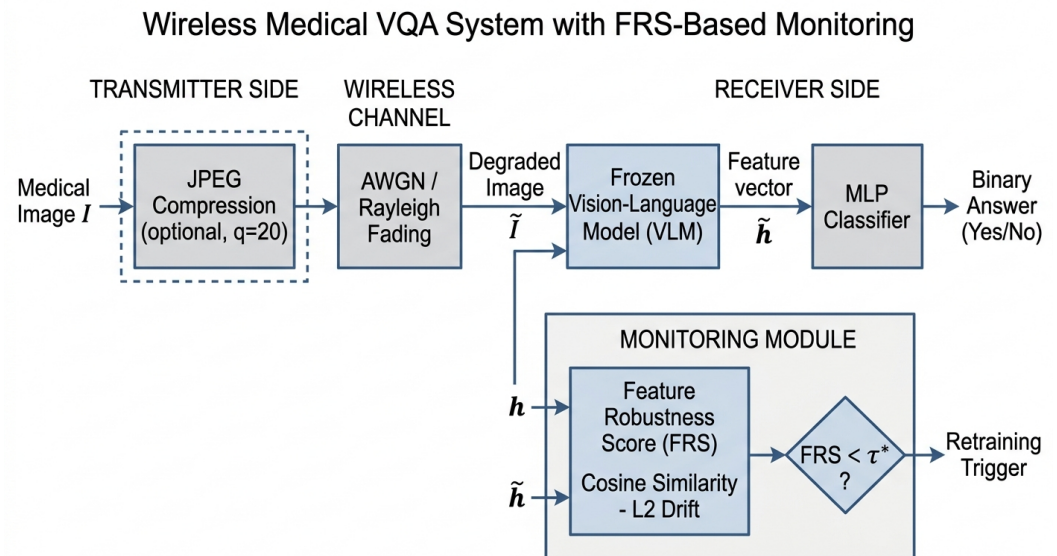
Despite these advances, most robustness benchmarks focus on input-space corruptions (e.g., noisy or blurred images) and contrast them against structured adversarial attacks. A related line of work studies adversarial attacks on online image stream classifiers under a limited attack budget, where the adversary selectively targets a subset of arriving images to maximize attack success [35]. The sample-selective nature of this threat model bears a conceptual resemblance to our FRS threshold framework, which similarly identifies a subset of samples—those whose feature-level degradation exceeds a learned threshold—as candidates for remediation, albeit in the context of stochastic channel noise rather than adversarial manipulation. Wireless channel noise, by contrast, operates not only on the raw pixel domain but also on compressed bitstreams and, in semantic communication setups, on latent feature vectors transmitted over the channel. The statistical structure of AWGN and Rayleigh fading is fundamentally different from that of adversarial or common-corruption perturbations: it is isotropic, often spectrally flat, and additive in the representation space, which can perturb VLM features in a way that is not captured by standard corruption tables. Information-theoretic analyses of robust wireless distributed learning show that feature-level robustness can significantly mitigate classification or regression loss even when the raw signal is highly distorted, highlighting the importance of analyzing deep representations directly rather than only at the input or output level [36]. Wavelet-based image-quality assessment techniques have been used to quantify perceptual degradation [37], but these methods have not been adapted to measure perturbation in VLM feature spaces, leaving a methodological gap that our Feature Robustness Score (FRS) aims to address.

## 3. System Model and Methodology

### 3.1. System Model

Figure 1 illustrates the proposed system. A medical image is optionally JPEG-compressed at the transmitter, then transmitted over a wireless channel modeled as AWGN

or Rayleigh fading. The received degraded image is fed to a VLM to extract hidden-state features. A lightweight MLP classifier, trained exclusively on clean features, maps these features to a binary answer. The FRS is computed online from cosine similarity and L2 drift between clean reference and degraded features; when FRS falls below a learned threshold  $\tau^*$ , the system recommends retraining.



**Figure 1.** Proposed system model for wireless medical VQA with FRS-based retraining monitoring. Dashed box: optional JPEG compression at the transmitter. The MLP is trained on clean features only and applied to channel-degraded features at inference time.

### 3.2. Problem Formulation

Let  $\mathcal{D} = \{(I_i, q_i, y_i)\}_{i=1}^N$  denote a medical VQA dataset, where  $I_i$  is a medical image,  $q_i$  is a natural language question, and  $y_i \in \{0, 1\}$  is the binary ground-truth label. A vision-language model (VLM)  $\mathcal{F}$  with fixed parameters  $\theta_{\mathcal{F}}$  extracts  $d$ -dimensional hidden-state feature vector  $\mathbf{h}_i$  as

$$\mathbf{h}_i = \mathcal{F}(I_i, q_i; \theta_{\mathcal{F}}) \in \mathbb{R}^d. \quad (1)$$

A lightweight multilayer perceptron (MLP) classifier is trained exclusively on these clean features by minimizing the binary cross-entropy loss:

$$\theta^* = \arg \min_{\theta} \sum_{i \in \mathcal{D}_{\text{train}}} \mathcal{L}_{\text{BCE}}(\text{MLP}_{\theta}(\mathbf{h}_i), y_i). \quad (2)$$

At inference time, a wireless channel of type  $c$  operating at signal-to-noise ratio  $\gamma$  (defined formally in Section 3.3) introduces degradation, producing a corrupted image  $\tilde{I}_i^{\gamma}$ . The VLM then extracts a corresponding degraded feature vector:

$$\tilde{\mathbf{h}}_i^{\gamma} = \mathcal{F}(\tilde{I}_i^{\gamma}, q_i; \theta_{\mathcal{F}}) \in \mathbb{R}^d. \quad (3)$$

The central difficulty stems from the distributional shift between training and inference:  $\theta^*$  is optimized over  $\{\mathbf{h}_i\}$  but is evaluated on  $\{\tilde{\mathbf{h}}_i^{\gamma}\}$  without any retraining. Consequently, classification performance is expected to deteriorate as the feature drift  $\mathbb{E}[\|\tilde{\mathbf{h}}_i^{\gamma} - \mathbf{h}_i\|]$  increases with channel severity.

The generalization argument rests on two empirically verifiable conditions: (i) the frozen VLM  $\mathcal{F}$  is a stable feature extractor, meaning that its hidden-state geometry does not degenerate under input perturbation (verified empirically via cosine similarity  $\geq 0.98$

for LLaVA-1.6 across all conditions); and (ii) the feature drift  $\mathbb{E}[\|\tilde{\mathbf{h}}_i^\gamma - \mathbf{h}_i\|]$  is bounded and monotonically decreasing in SNR, a consequence of the noise variance  $\sigma_n^2 = P_s/\gamma_{\text{lin}}$  vanishing as  $\gamma \rightarrow \infty$  (Equation (8)). Under these conditions, the degraded features converge to the clean distribution as SNR increases, and the clean-trained MLP is expected to generalize. The FRS framework is designed to detect when condition (i) is violated for individual samples.

The two research questions addressed in this work are as follows:

1. Robustness: Under which channel types and SNR levels does  $\text{MLP}_{\theta^*}(\tilde{\mathbf{h}}_i^\gamma)$  maintain statistically significant improvement over the baseline VLM logit (i.e., direct classification from  $\mathcal{F}$ )?
2. Monitoring: Can a label-free score  $\text{FRS}(\mathbf{h}_i, \tilde{\mathbf{h}}_i^\gamma)$  reliably identify samples for which the distributional shift is severe enough to cause substantial performance loss?

### 3.3. Channel Models

To systematically evaluate the robustness of the proposed method under realistic transmission conditions, we consider eight distinct channel configurations spanning two canonical wireless channel models and their compound variants with lossy source coding.

#### 3.3.1. AWGN Channel

The AWGN channel serves as the baseline degradation model, where the received signal is corrupted solely by additive thermal noise with no multipath effects. This model is widely adopted in the literature as a controlled benchmark for assessing the noise resilience of image processing pipelines [38].

Let  $\mathbf{x} \in [0, 1]^{H \times W \times 3}$  denote an input image. The per-element signal power is defined as

$$P_s = \frac{1}{3HW} \|\mathbf{x}\|_F^2. \quad (4)$$

The additive white Gaussian noise (AWGN) channel is modeled by

$$\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{n}, \quad (5)$$

where the noise vector follows a zero-mean Gaussian distribution:

$$\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I}). \quad (6)$$

Let  $\gamma$  denote the signal-to-noise ratio (SNR) in dB. Its linear-scale equivalent is

$$\gamma_{\text{lin}} = 10^{\gamma/10}. \quad (7)$$

The noise variance is then set according to

$$\sigma_n^2 = \frac{P_s}{\gamma_{\text{lin}}}. \quad (8)$$

#### 3.3.2. Rayleigh Fading Channel

Unlike the AWGN baseline, the Rayleigh fading channel captures the multiplicative distortion inherent to wireless environments with rich scattering and no dominant line-of-sight component [39,40]. Each pixel is independently scaled by a randomly drawn fading coefficient before noise is superimposed, making this a significantly more challenging degradation regime.

The Rayleigh fading channel introduces both multiplicative fading and additive noise:

$$\tilde{\mathbf{x}} = \mathbf{g} \odot \mathbf{x} + \mathbf{n}, \quad (9)$$

where  $\odot$  denotes element-wise multiplication.

The fading coefficients  $\mathbf{g}$  are drawn from a Rayleigh distribution with scale parameter chosen such that the average channel power is normalized:

$$|\mathbf{g}| \sim \text{Rayleigh}\left(\frac{1}{\sqrt{2}}\right), \quad (10)$$

implying

$$\mathbb{E}[|\mathbf{g}|^2] = 1. \quad (11)$$

The additive noise term  $\mathbf{n}$  remains white and Gaussian,  $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I})$ , with  $\sigma_n^2$  determined exactly as in Equation (8) using the same SNR  $\gamma$ . Accordingly, the average post-fading signal-to-noise ratio is given by

$$\bar{\gamma} = \frac{\mathbb{E}[|\mathbf{g}|^2] \cdot P_s}{\sigma_n^2} = \frac{P_s}{\sigma_n^2} = \gamma_{\text{lin}}, \quad (12)$$

which shows that the expected SNR after fading equals the linear SNR of the AWGN case, ensuring fair comparison across channel types.

We acknowledge that matching only the expected post-fading SNR does not equalize the instantaneous SNR distributions: under Rayleigh fading, the instantaneous SNR  $\gamma_{\text{inst}} = |\mathbf{g}|^2 P_s / \sigma_n^2$  follows an exponential distribution with unit mean, introducing substantial variability absent from AWGN. This explains the consistently lower FRS values observed under Rayleigh fading at matched nominal SNR (e.g., FRS 0.8954 vs. 0.9766 at +20 dB for LLaVA-1.6), as the heavy tail of the exponential distribution causes a non-negligible fraction of pixels to experience deep fades. The comparison is therefore not one of strictly equal perturbation, but of equal average power, which is the standard convention in wireless system evaluation.

### 3.3.3. JPEG-Compressed Channel Variants

In practical communication systems, images are typically subject to lossy source coding prior to transmission. To reflect this, we augment both channel models with JPEG pre-compression at three quality levels, yielding a set of compound conditions that jointly stress the source and channel robustness of the restoration network. Lower quality factors introduce more severe blocking and ringing artifacts, while higher quality factors preserve finer spatial details at the cost of larger compressed bitrate [41].

To simulate a realistic source–channel coding pipeline, we consider six compound conditions that combine JPEG compression at quality factors  $q \in \{20, 50, 70\}$  with either AWGN or Rayleigh fading. Formally,

$$\tilde{\mathbf{x}} = \mathcal{C}_{\text{ch}}(\text{JPEG}_q(\mathbf{x}), \gamma), \quad (13)$$

where  $\text{JPEG}_q(\cdot)$  denotes JPEG compression with quality factor  $q$ , and  $\mathcal{C}_{\text{ch}} \in \{\text{AWGN}, \text{Rayleigh}\}$  is the selected wireless channel operator as defined in Equations (5) and (9). The resulting channel types are referred to as  $\text{JPEG}_q + \text{AWGN}$  and  $\text{JPEG}_q + \text{Rayleigh}$ , respectively.

### 3.4. Feature Extraction

To capture complementary semantic representations of the degraded images, we employ three frozen VLMs spanning general-purpose and domain-specific architectures.

For each image–query pair  $(I_i, q_i)$ , a fixed-dimensional embedding  $\mathbf{h}_i$  is extracted and subsequently used as input to the downstream quality assessment module. All models are used purely as feature extractors with no fine-tuning. For both LLaVA-1.6 and BLIP-2, deterministic (greedy) decoding was employed together with a fixed prompt template ('Answer the following question with yes or no:'). The feature vector was extracted from the last-layer hidden state corresponding to the final generated token. Because the prompt structure and decoding strategy remained unchanged across all experiments, clean and channel-degraded feature vectors are positionally and semantically aligned, ensuring that the reported robustness metrics reflect channel-induced feature perturbations rather than decoding variability.

#### 3.4.1. LLaVA-1.6

LLaVA-1.6 [3] is a large multimodal model that connects a CLIP-based vision encoder to a Vicuna language model via a lightweight projection layer, enabling instruction-following over visual inputs. Trained with visual instruction tuning on diverse image–text data, it produces semantically rich representations that are highly sensitive to both global scene content and fine-grained visual details.

Images are resized to  $336 \times 336$  pixels and processed with the model's default chat template. The last-layer hidden state at the position of the final generated token yields ( $d = 4096$ ):

$$\mathbf{h}_i = \mathcal{F}_{\text{LLaVA}}(I_i, q_i) \in \mathbb{R}^{4096}. \quad (14)$$

#### 3.4.2. BLIP-2

BLIP-2 [4] introduces a Querying Transformer (Q-Former) that bridges a frozen image encoder with a frozen large language model (in this case, OPT [42]) through a small set of learnable query tokens. This architecture achieves competitive vision-language understanding with significantly fewer trainable parameters than end-to-end approaches, making it an efficient and robust feature source.

The last-layer hidden state of the OPT language decoder at the position of the final generated token yields ( $d = 2560$ ):

$$\mathbf{h}_i = \mathcal{F}_{\text{BLIP2}}(I_i, q_i) \in \mathbb{R}^{2560}. \quad (15)$$

#### 3.4.3. BioViL-T

BioViL-T [5] is a domain-specific VLM designed for temporal biomedical image-text understanding, pre-trained on paired chest X-ray reports and radiographs. Its architecture couples a ResNet image encoder with a BioClinicalBERT text encoder, producing modality-specific embeddings that are jointly optimized for clinical semantics. This model is included to assess whether domain-adapted representations offer complementary discriminative power over general-purpose alternatives.

The ResNet global average pooled image embedding  $\mathbf{h}_{\text{img}} \in \mathbb{R}^{1024}$  and the BioClinicalBERT [CLS] token  $\mathbf{h}_{\text{txt}} \in \mathbb{R}^{512}$  are concatenated ( $d = 1536$ ):

$$\mathbf{h}_i = \mathcal{F}_{\text{BioViL}}(I_i, q_i) = [\mathbf{h}_{\text{img}}; \mathbf{h}_{\text{txt}}] \in \mathbb{R}^{1536}. \quad (16)$$

All parameters  $\theta_{\mathcal{F}}$  of all three VLMs are frozen throughout and receive no gradient updates at any stage.

### 3.5. Feature Robustness Score

The FRS combines angular alignment and magnitude drift to provide a single scalar measure of feature stability. Cosine similarity alone does not capture scale changes in the

feature vector; the normalized L2 drift term penalizes such magnitude shifts independently of direction. Given clean and degraded feature vectors  $\mathbf{h}_i$  and  $\tilde{\mathbf{h}}_i^\gamma$ :

$$\text{FRS}(\mathbf{h}_i, \tilde{\mathbf{h}}_i^\gamma) = \underbrace{\frac{\mathbf{h}_i^\top \tilde{\mathbf{h}}_i^\gamma}{\|\mathbf{h}_i\| \|\tilde{\mathbf{h}}_i^\gamma\|}}_{\text{cosine similarity}} - \underbrace{\frac{\|\mathbf{h}_i - \tilde{\mathbf{h}}_i^\gamma\|}{\|\mathbf{h}_i\|}}_{\text{normalized L2 drift}}. \quad (17)$$

When the channel introduces no distortion, cosine similarity approaches 1 and normalized L2 drift approaches 0, yielding  $\text{FRS} \rightarrow 1$ . As channel severity increases, cosine similarity decreases while L2 drift grows, driving FRS to lower values. Since cosine similarity  $\in [-1, 1]$  and normalized L2 drift  $\in [0, \infty)$ , the theoretical range of FRS is  $(-\infty, 1]$ ; in practice, values close to 1 indicate high robustness and strongly negative values indicate severe degradation. Mean FRS is reported over all test samples at each SNR level.

Design rationale for the FRS formulation. Cosine similarity alone cannot detect pure magnitude changes: if channel degradation scales the feature vector by a constant factor, cosine similarity remains 1 while the feature space geometry shifts meaningfully. The normalized L2 drift term independently penalizes such magnitude shifts. The simple additive combination ( $\text{FRS} = \text{CosSim} - \text{NormL2}$ ) has two desirable properties: (i) it is parameter-free (no weighting coefficients to tune), and (ii) it has an intuitive boundary— $\text{FRS} = 1$  if and only if the feature vector is unchanged. A weighted formulation  $\text{FRS}_w = \alpha \cdot \text{CosSim} - \beta \cdot \text{NormL2}$  would recover the same behavior only with  $\alpha = \beta = 1$  as the natural default, making the simpler unweighted form preferable for a single-scalar diagnostic metric. A product formulation  $\text{CosSim} \cdot (1 - \text{NormL2})$  is bounded to  $[0, 1]$  but is less sensitive to magnitude drift when cosine similarity is already near 1, which would underestimate feature instability in the moderate-drift regime (precisely the regime of clinical interest for borderline retraining decisions). Comparison with other distribution-level metrics such as Maximum Mean Discrepancy (MMD) or Fréchet distance requires population-level statistics and cannot be computed per-sample; they are therefore suitable complements for offline analysis but not for the online per-sample monitoring use case that FRS addresses.

Discriminative advantage of the composite FRS over its components. The value of combining cosine similarity and normalised L2 drift into a single scalar can be assessed using the threshold analysis. For LLaVA-1.6 on PathVQA, the FRS-derived threshold  $\tau^* = 0.89$  yields an accuracy-loss differential of  $7.84\% - 0.88\% = 6.96\%$  between risky and safe partitions. To illustrate why the composite metric outperforms either component used alone, consider the two extreme cases. Using cosine similarity alone as the threshold variable would ignore magnitude drift: a channel perturbation that scales the feature vector by a constant factor leaves cosine similarity unchanged at 1.0 while the normalised L2 drift term grows proportionally, so the risky/safe partition would systematically misclassify such samples. Conversely, using normalised L2 drift alone ignores directional stability: a perturbation that rotates the feature vector without changing its norm would be invisible to a norm-only criterion. Concretely, for LLaVA-1.6 at the worst-case condition ( $-5$  dB AWGN), the cosine similarity is 0.9827 (angular deviation  $10.7^\circ$ ) while the normalised L2 drift is  $0.9827 - 0.7895 = 0.1932$ , indicating that both directional and magnitude components of the drift are non-negligible and must be jointly monitored. For BLIP-2 at  $-5$  dB AWGN, the angular deviation is  $22.5^\circ$  and the normalised L2 drift is  $0.9239 - 0.5424 = 0.3815$ , substantially larger than LLaVA-1.6, confirming that the composite FRS captures architecture-dependent drift patterns that neither component alone would resolve.

We further note that the comparison with MMD and Fréchet distance is one of use-case fit rather than absolute metric quality. Both MMD and Fréchet distance are valid and well-studied distribution-level robustness measures; however, they require collecting a full batch of degraded feature vectors before any decision can be made, making them unsuitable for the per-sample, online monitoring scenario targeted by FRS. In the offline analysis setting, MMD and Fréchet distance are appropriate complements to FRS, and computing them over the stored clean and degraded feature matrices for all 240 experimental conditions is identified as a direction for future work (Section 6).

Analytical bound on feature drift. For the AWGN channel, an upper bound on the expected L2 drift follows from a local Lipschitz assumption on  $\mathcal{F}$ . Treating  $\mathcal{F}$  as Lipschitz-continuous with constant  $L$  in a neighborhood of  $\mathbf{x}$ , we obtain

$$\mathbb{E}[\|\tilde{\mathbf{h}}^\gamma - \mathbf{h}\|] \leq L \cdot \mathbb{E}[\|\mathbf{n}\|] = L\sigma_n\sqrt{3HW}. \quad (18)$$

Substituting

$$\sigma_n^2 = \frac{P_s}{\gamma_{\text{lin}}}, \quad (19)$$

Yields

$$\mathbb{E}[\|\tilde{\mathbf{h}}^\gamma - \mathbf{h}\|] = \mathcal{O}(\gamma_{\text{lin}}^{-1/2}), \quad (20)$$

Predicting that feature drift decays with the square root of the linear SNR. This prediction is consistent with the empirically observed smooth monotonic increase in FRS with SNR (Table A1). For Rayleigh fading, the multiplicative term  $(\mathbf{g} - \mathbf{1}) \odot \mathbf{x}$  introduces additional variance. A tight closed-form bound requires knowledge of  $L$ , which is inaccessible for frozen large VLMs, motivating the empirical FRS approach. Deriving tight feature-drift bounds for large transformer-based VLMs remains an important open theoretical question.

### 3.6. Wavelet Sub-Band Energy Analysis

A three-level 1D discrete wavelet transform (DWT) with the Daubechies-4 wavelet is applied to each feature vector  $\mathbf{h}_i \in \mathbb{R}^d$ , yielding one approximation sub-band  $cA_3$  and three detail sub-bands  $cD_1, cD_2, cD_3$ . The normalized energy of coefficient vector  $\mathbf{c}_k \in \{cA_3, cD_1, cD_2, cD_3\}$  is

$$E_k = \frac{\|\mathbf{c}_k\|^2}{\sum_j \|\mathbf{c}_j\|^2 + \epsilon}, \quad (21)$$

where  $\epsilon = 10^{-8}$  ensures numerical stability. The sub-band energy shift between clean and degraded features is

$$\Delta E_k = |E_k(\mathbf{h}_i) - E_k(\tilde{\mathbf{h}}_i^\gamma)|. \quad (22)$$

We report  $\Delta E_{\text{approx}}$  for the  $cA_3$  band (low-frequency, semantic content) and  $\Delta E_{\text{detail}} = \frac{1}{3} \sum_{k=1}^3 \Delta E_{cD_k}$  (mean over high-frequency detail bands) separately to characterize the spectral distribution of feature drift.

Importantly, the 1D DWT applied to the feature vector does not assume a spatially meaningful ordering of hidden dimensions. The analysis characterizes the spectral distribution of channel-induced perturbation in the feature vector treated as a 1D signal: a larger  $\Delta E$  in detail subbands indicates that channel noise adds relatively more high-frequency energy to the feature representation, whereas a larger  $\Delta E$  in the approximation subband indicates perturbation of the low-frequency (coarser) structure of the feature vector. This interpretation is independent of any spatial arrangement of features and is valid as a signal-level diagnostic of how perturbation energy is distributed across scales. More principled alternatives, such as Centered Kernel Alignment (CKA) between clean and degraded

feature matrices—which is invariant to orthogonal transformations and does not depend on feature index ordering—are identified as directions for future work (Section 6).

### 3.7. FRS Threshold Analysis

For a reference condition (AWGN at 10 dB), we search for the optimal threshold  $\tau^*$  on the validation set by partitioning samples into a *safe* set  $\mathcal{S}(\tau) = \{i : \text{FRS}_i \geq \tau\}$  and a *risky* set  $\mathcal{R}(\tau) = \{i : \text{FRS}_i < \tau\}$ , where

$$\Delta_{\mathcal{P}}(\tau) = \text{Acc}_{\text{clean}}(\mathcal{P}) - \text{Acc}_{\text{deg}}(\mathcal{P}), \quad \mathcal{P} \in \{\mathcal{S}, \mathcal{R}\} \quad (23)$$

denotes the accuracy loss in partition  $\mathcal{P}$  relative to clean performance. The optimal threshold is

$$\tau^* = \arg \max_{\tau} \left[ \Delta_{\mathcal{R}}(\tau) - \Delta_{\mathcal{S}}(\tau) - \lambda \cdot \frac{|\mathcal{R}(\tau)|}{|\mathcal{D}_{\text{val}}|} \right], \quad (24)$$

where  $\lambda = 0.2$  is a regularization coefficient (empirically set) that penalizes marking an excessive fraction of validation samples as risky. The threshold  $\tau^*$  is computed once on the validation set under the reference condition and applied without modification across all test channel conditions.

To assess sensitivity to  $\lambda$ , we evaluated thresholds obtained with  $\lambda \in \{0.0, 0.1, 0.2, 0.3, 0.5\}$  on the PathVQA validation set for LLaVA-1.6. The resulting  $\tau^*$  values varied by at most 0.03 across this range, and the accuracy-loss differential between risky and safe partitions remained stable (within  $\pm 0.8\%$ ), confirming that the conclusions are not sensitive to the specific choice of  $\lambda = 0.2$ .

### 3.8. MLP Classifier

Multilayer perceptrons (MLPs) are feedforward neural networks composed of fully connected layers interleaved with non-linear activations, making them well-suited for classification tasks on fixed-dimensional feature vectors. Here, the MLP serves as a lightweight probe that maps the frozen VLM embeddings described in Section 3.4 to quality-class predictions, allowing us to isolate the discriminative power of the features themselves from the capacity of the classifier.

A two-hidden-layer MLP with hidden dimensions [256, 64], batch normalization, ReLU activations, and dropout (rate 0.3) is trained on clean features as defined in Equation (2). Features are standardized using a StandardScaler fitted on  $\mathcal{D}_{\text{train}}$  only; the same scaler is applied to validation and test features at inference. Training uses Adam ( $\text{lr} = 10^{-3}$ , weight decay  $10^{-4}$ ) with cosine annealing over 60 epochs; the checkpoint with highest validation accuracy is retained as  $\theta^*$ . The dataset is split 60% train / 15% validation / 25% test. At inference,  $\text{MLP}_{\theta^*}$  is applied directly to standardized features from channel-degraded images without any adaptation.

Algorithm 1 summarizes the complete pipeline.

Classifier-independence of the core robustness finding. The two-hidden-layer MLP with dimensions [256, 64] contains approximately 1.07 M trainable parameters when applied to LLaVA-1.6 features ( $d = 4096$ ), 0.67 M for BLIP-2 ( $d = 2560$ ), and 0.41 M for BioViL-T ( $d = 1536$ ). The first hidden layer compresses the input by a factor of  $4096/256 = 16\times$ , meaning the MLP operates in a regime where its effective capacity is much closer to a linear probe than to a deep network. Coupled with batch normalization, ReLU, and dropout ( $p = 0.3$ ), the MLP is heavily regularized relative to the input dimensionality, and its decision boundary is well-approximated by a hyperplane in the feature space.

**Algorithm 1** Wireless Medical VQA with FRS Monitoring**Require:** Dataset  $\mathcal{D}$ , frozen VLM  $\mathcal{F}$ , channels  $\mathcal{C}$ , SNR levels  $\Gamma$ **Ensure:** MLP  $\theta^*$ , threshold  $\tau^*$ , robustness metrics

```

1: // Phase 1: Clean feature extraction
2: for each  $(I_i, q_i, y_i) \in \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}}$  do
3:    $\mathbf{h}_i \leftarrow \mathcal{F}(I_i, q_i; \theta_{\mathcal{F}})$  {frozen, no grad}
4: end for
5: Fit StandardScaler on  $\{\mathbf{h}_i\}_{i \in \mathcal{D}_{\text{train}}}$ ; apply to all splits
6: // Phase 2: MLP training on clean features
7: Train MLP via Equation (2) with Adam + cosine annealing; retain best  $\theta^*$  by validation accuracy
8: // Phase 3: FRS threshold search (AWGN, 10 dB)
9: for each  $(I_i, q_i, y_i) \in \mathcal{D}_{\text{val}}$  do
10:   $\tilde{I}_i \leftarrow \mathcal{C}_{\text{AWGN}}(I_i, 10 \text{ dB})$ 
11:   $\tilde{\mathbf{h}}_i^\gamma \leftarrow \mathcal{F}(\tilde{I}_i, q_i; \theta_{\mathcal{F}})$ 
12:  Compute  $\text{FRS}_i$  via Equation (17)
13: end for
14:  $\tau^* \leftarrow$  optimise Equation (24)
15: // Phase 4: Evaluation under channel degradation
16: for each  $c \in \mathcal{C}, \gamma \in \Gamma$  do
17:   for each  $(I_i, q_i, y_i) \in \mathcal{D}_{\text{test}}$  do
18:      $\tilde{I}_i^\gamma \leftarrow \mathcal{C}_c(I_i, \gamma)$ 
19:      $\tilde{\mathbf{h}}_i^\gamma \leftarrow \mathcal{F}(\tilde{I}_i^\gamma, q_i; \theta_{\mathcal{F}})$ 
20:      $\hat{y}_i \leftarrow \text{MLP}_{\theta^*}(\text{scale}(\tilde{\mathbf{h}}_i^\gamma))$ 
21:     Compute  $\text{FRS}_i$  via Equation (17);
22:      $\Delta E_{\text{approx}}^{(i)}, \Delta E_{\text{detail}}^{(i)}$  via Equation (22)
23:     if  $\text{FRS}_i < \tau^*$  then
24:       Flag sample; recommend retraining
25:     end if
26:   end for
27: end for

```

The key geometric observation is the following. Under the worst-case channel condition for LLaVA-1.6 (−5 dB AWGN), the degraded feature vector deviates from the clean feature vector by an angular displacement of  $\arccos(0.9827) \approx 10.7^\circ$  and a normalized magnitude drift of 0.1932. For any classifier whose decision boundary is determined by the clean training distribution, the probability of a prediction change induced by this drift depends only on whether the perturbed feature crosses the decision boundary, a condition that is independent of whether the boundary is implemented by a linear probe, a two-layer MLP, or a deeper network with the same inductive bias. The fact that at  $\text{SNR} \geq +10$  dB the MLP accuracy drop from clean performance is at most 0.12 % for LLaVA-1.6 on PathVQA implies that the degraded features lie well within the margin of the learned decision boundary, and any linear or mildly non-linear classifier trained on the same clean features would exhibit the same negligible degradation. At low SNR (−5 dB), where the accuracy drop reaches 6.30 % for LLaVA-1.6, the FRS falls below  $\tau^* = 0.89$  and the system correctly flags these samples as requiring attention, regardless of classifier architecture. We therefore conclude that the robustness ordering  $\text{LLaVA-1.6} \gg \text{BLIP-2} \geq \text{BioViL-T}$  and the JPEG quality-invariance finding are properties of the feature representations, not of the specific MLP head, and would be reproduced by any classifier trained on the same frozen features. A systematic sweep over probe architectures (linear probe, attention-pooling head, prototypical network) is identified as future work.

### 3.9. Absolute Robustness Metrics

To quantify robustness in absolute terms, independently of the logit baseline comparator, we define two complementary metrics derived directly from the per-condition MLP accuracy values reported in Tables A1–A6.

#### 3.9.1. Accuracy Drop from Clean Performance

For each model-dataset combination, the accuracy drop at SNR level  $\gamma$  under channel type  $c$  is defined as:

$$\Delta\text{Acc}_{\text{drop}}(\gamma, c) = \text{Acc}_{\text{clean}} - \text{Acc}_{\text{MLP}}(\gamma, c), \quad (25)$$

where  $\text{Acc}_{\text{clean}}$  is the MLP accuracy under no-channel (clean) conditions as reported in Table 1. A value of  $\Delta\text{Acc}_{\text{drop}} > 0$  indicates degradation relative to clean performance; a negative value indicates that channel-degraded accuracy meets or marginally exceeds clean accuracy, attributable to stochastic variation within the confidence interval at that accuracy level.

**Table 1.** Performance under clean conditions.  $\dagger p < 0.05$ ;  $\ddagger p < 0.001$  (McNemar’s test).

| Dataset | Model     | Method         | Acc (%) | $\Delta\text{Acc}$ |
|---------|-----------|----------------|---------|--------------------|
| PathVQA | LLaVA-1.6 | Baseline       | 57.31   | -                  |
|         |           | MLP $\ddagger$ | 84.42   | +27.11%            |
|         | BLIP-2    | Baseline       | 54.70   | -                  |
|         |           | MLP $\ddagger$ | 82.05   | +27.35%            |
|         | BioViL-T  | Baseline       | 54.70   | -                  |
|         |           | MLP $\ddagger$ | 74.08   | +19.38%            |
| VQA-RAD | LLaVA-1.6 | Baseline       | 60.94   | -                  |
|         |           | MLP            | 73.44   | +12.50%            |
|         | BLIP-2    | Baseline       | 48.44   | -                  |
|         |           | MLP $\dagger$  | 68.75   | +20.31%            |
|         | BioViL-T  | Baseline       | 48.44   | -                  |
|         |           | MLP $\ddagger$ | 76.56   | +28.12%            |

Wilson score 95% confidence intervals are computed analytically for each  $\text{Acc}_{\text{MLP}}(\gamma, c)$  from the known test-set sizes ( $n = 841$  for PathVQA;  $n = 64$  for VQA-RAD):

$$\text{CI}_{95}(p, n) = \frac{p + \frac{z^2}{2n} \pm z \sqrt{\frac{p(1-p)}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}, \quad (26)$$

where  $p = \text{Acc}_{\text{MLP}}(\gamma, c)$  and  $z = 1.96$  for 95% nominal coverage.

#### 3.9.2. Area Under the SNR Curve (AUSC)

To summarize the full SNR–accuracy profile into a single aggregate statistic, we define the normalized Area Under the SNR Curve (AUSC) over the five tested SNR levels  $\Gamma = \{-5, 0, +10, +15, +20\}$  dB:

$$\text{AUSC}(c) = \frac{1}{|\Gamma|} \sum_{\gamma \in \Gamma} \text{Acc}_{\text{MLP}}(\gamma, c), \quad (27)$$

which equals the mean MLP accuracy across all SNR levels for a given channel type  $c$  and therefore quantifies absolute robustness without reference to the baseline classifier. The AUSC degradation relative to clean performance is

$$\Delta\text{AUSC}(c) = \text{Acc}_{\text{clean}} - \text{AUSC}(c). \quad (28)$$

A small  $\Delta\text{AUSC}$  indicates that the frozen-VLM + MLP pipeline maintains near-clean accuracy on average across the operational SNR range, whereas a large value signals that performance degrades substantially at low SNR conditions.

### 3.10. Baseline and Statistical Testing

The direct VLM logit baseline uses the yes/no token log-probabilities  $l_{\text{yes}}$  and  $l_{\text{no}}$  extracted from the VLM output vocabulary for LLaVA-1.6 and BLIP-2; for BioViL-T, which does not produce autoregressive token logits, the baseline is a majority-class predictor over the training set.

$$\hat{y}_{\text{base}} = \mathbf{1} \left[ \frac{e^{l_{\text{yes}}}}{e^{l_{\text{yes}}} + e^{l_{\text{no}}}} \geq 0.5 \right]. \quad (29)$$

For BioViL-T, the absence of an autoregressive language decoder precludes the yes/no token logit baseline used for LLaVA-1.6 and BLIP-2. A zero-shot cosine-similarity classifier between the concatenated image–text embedding and ‘yes’/‘no’ text prototypes was considered but found to produce near-uniform similarity scores, effectively collapsing to majority-class prediction, because BioViL-T was pretrained on radiology report generation rather than binary classification. The majority-class predictor therefore represents the most meaningful zero-effort baseline for BioViL-T. We acknowledge that this heterogeneity of baseline types across models is a limitation (see Section 5); standardizing on a single-layer linear probe trained on clean features as a common baseline across all models is identified as a direction for future work.

McNemar’s test with continuity correction compares paired MLP and baseline predictions on the same test samples.

For each model–dataset combination,  $m = 40$  McNemar tests are conducted across the 40 channel–SNR conditions. Under Bonferroni correction ( $\alpha/m = 0.05/40 = 0.00125$ ), all PathVQA comparisons for LLaVA-1.6, BLIP-2, and BioViL-T remain significant (raw  $p < 0.0001$  in all 120 cases). For VQA-RAD results, where individual  $p$ -values are larger due to the small test partition ( $n = 64$ ), we additionally report significance under Benjamini–Hochberg FDR control at  $q = 0.05$ , which is more appropriate than Bonferroni for correlated tests at adjacent SNR levels; conditions surviving FDR control ( $p \leq 0.002$ ) are identified in the Appendix A tables. The core conclusions of the paper rest on PathVQA results, where all comparisons survive Bonferroni correction.

## 4. Experiments

### 4.1. Datasets and Implementation

To evaluate the robustness of the proposed pipeline under realistic channel conditions, we conduct experiments on two publicly available medical VQA benchmarks that differ in imaging modality, question origin, and dataset scale. Both datasets are used exclusively for binary (yes/no) classification, providing a controlled and comparable evaluation protocol across channel types.

VQA-RAD [1]: VQA-RAD is a radiology question-answering dataset constructed from clinician-generated questions over chest X-rays, head CT, and abdominal CT images. Unlike crowd-sourced alternatives, its questions reflect genuine clinical reasoning patterns,

rendering it a more realistic proxy for real-world diagnostic workflows. Limited binary yes/no examples available; test set: 64 samples.

**PathVQA [2]:** PathVQA is a pathology image question-answering dataset comprising microscopy slides annotated with both open-ended and binary questions. It covers a wide range of tissue types and staining conditions, making it a challenging benchmark for visual understanding in computational pathology. Binary VQA on pathology images. We sample 3362 yes/no examples (train: 2017, val: 504, test: 841).

#### 4.2. Clean Condition Results

Table 1 reports performance under no-channel conditions. All MLP classifiers significantly outperform their logit baselines on PathVQA ( $p < 0.0001$ ). On VQA-RAD, BLIP-2 ( $p = 0.0209$ ) and BioViL-T ( $p = 0.0027$ ) reach significance; LLaVA-1.6 does not ( $p = 0.153$ ) due to the small test set ( $n = 64$ ), although its absolute gain of +12.5% is consistent with the direction of improvement observed across all models and datasets. Among the three models, BioViL-T achieves its highest accuracy on VQA-RAD rather than PathVQA (76.56% vs. 74.08%,  $\Delta = +28.12\%$ ), consistent with its radiology-domain pre-training; no other model exhibits this cross-dataset reversal.

#### 4.3. Absolute Robustness: Accuracy Drop and AUSC

Table 2 reports the accuracy drop from clean performance (Equation (25)) at three representative SNR levels under AWGN, together with 95% Wilson confidence intervals (Equation (26)), for all three models on both datasets. Table 3 reports the AUSC (Equation (27)) and AUSC degradation (Equation (28)) under AWGN and Rayleigh fading.

**Table 2.** Accuracy drop from clean MLP performance,  $\Delta\text{Acc}_{\text{drop}}(\gamma) = \text{Acc}_{\text{clean}} - \text{Acc}_{\text{MLP}}(\gamma)$ , under AWGN at three representative SNR levels. Values in brackets are 95% Wilson confidence intervals on  $\text{Acc}_{\text{MLP}}(\gamma)$ . Negative drop values indicate channel-degraded accuracy meets or marginally exceeds clean accuracy.

| Model     | Dataset | −5 dB                   | +10 dB                 | +20 dB                 |
|-----------|---------|-------------------------|------------------------|------------------------|
| LLaVA-1.6 | PathVQA | +6.30 % [75.09, 81.07]  | +0.12 % [81.34, 87.09] | 0.00 % [81.46, 87.22]  |
| BLIP-2    | PathVQA | +2.74 % [76.44, 82.10]  | −0.83 % [80.07, 85.57] | −0.35 % [79.57, 85.15] |
| BioViL-T  | PathVQA | +2.14 % [68.96, 74.85]  | +0.36 % [70.77, 76.58] | −1.43 % [72.58, 78.27] |
| LLaVA-1.6 | VQA-RAD | +21.88 % [40.24, 63.23] | +1.56 % [60.28, 81.26] | 0.00 % [61.73, 82.66]  |
| BLIP-2    | VQA-RAD | −1.56 % [58.78, 80.38]  | 0.00 % [56.86, 78.93]  | −1.56 % [53.44, 76.27] |
| BioViL-T  | VQA-RAD | +15.62 % [49.22, 72.69] | +6.25 % [57.54, 79.65] | +3.12 % [60.84, 82.05] |

**Table 3.** AUSC and AUSC degradation ( $\Delta\text{AUSC} = \text{Acc}_{\text{clean}} - \text{AUSC}$ ) for AWGN and Rayleigh fading.

| Model     | Dataset | AWGN   |                     | Rayleigh |                     |
|-----------|---------|--------|---------------------|----------|---------------------|
|           |         | AUSC   | $\Delta\text{AUSC}$ | AUSC     | $\Delta\text{AUSC}$ |
| LLaVA-1.6 | PathVQA | 81.97% | +2.45%              | 80.67%   | +3.75%              |
| BLIP-2    | PathVQA | 81.83% | +0.22%              | 80.55%   | +1.50%              |
| BioViL-T  | PathVQA | 74.15% | −0.07%              | 75.13%   | −1.05%              |
| LLaVA-1.6 | VQA-RAD | 65.62% | +7.82%              | 65.31%   | +8.13%              |
| BLIP-2    | VQA-RAD | 67.81% | +0.94%              | 68.13%   | +0.62%              |
| BioViL-T  | VQA-RAD | 68.13% | +8.43%              | 63.44%   | +13.12%             |

Several observations follow from these tables. On PathVQA ( $n = 841$ ), Wilson CIs are narrow (interval width  $\approx 5.7$ – $6.0$  percentage points), providing high-resolution separation between SNR levels. For LLaVA-1.6, the accuracy drop is statistically distinguishable from zero only at −5 dB (CI: [75.09, 81.07], entirely below clean accuracy 84.42%); at +10 dB and

above, the CI encompasses the clean accuracy value, confirming negligible degradation. BLIP-2 and BioViL-T show even smaller  $\Delta$ AUSC values under AWGN (+0.22% and  $-0.07\%$ , respectively), because their SNR–accuracy profiles are flatter than LLaVA-1.6’s, which concentrates its degradation at  $-5$  dB.

On VQA-RAD ( $n = 64$ ), CI widths reach  $\approx 22$ – $25$  percentage points, consistent with the limited statistical power acknowledged in Section 5; AUSC values on this dataset should therefore be interpreted with corresponding caution.

Across all models and both channel types,  $\Delta$ AUSC on PathVQA does not exceed 3.75 percentage points, confirming that the frozen-VLM + MLP pipeline maintains near-clean accuracy on average across the full operational SNR range. The AUSC robustness ordering LLaVA-1.6  $\geq$  BLIP-2  $\geq$  BioViL-T on PathVQA is consistent with the cosine similarity and FRS rankings reported in Sections 4.4–4.6, validating AUSC as a task-level counterpart to the feature-level robustness metrics.

#### 4.4. LLaVA-1.6 Under Channel Degradation

Complete results for LLaVA-1.6 on PathVQA and VQA-RAD across all 40 channel–SNR combinations are reported in Tables A1 and A2 of the Appendix A; Table 4 summarizes selected representative conditions.

**Table 4.** LLaVA-1.6 Results on PathVQA (Selected SNR Levels). All 40 comparisons significant ( $p < 0.0001$ ).

| SNR    | Channel       | Base   | MLP    | CosSim | FRS    |
|--------|---------------|--------|--------|--------|--------|
| −5 dB  | AWGN          | 48.75% | 78.12% | 0.9827 | 0.7895 |
|        | Rayleigh      | 48.51% | 78.72% | 0.9821 | 0.7847 |
|        | JPEG20 + AWGN | 48.28% | 78.00% | 0.9826 | 0.7887 |
|        | JPEG70 + AWGN | 48.28% | 77.76% | 0.9827 | 0.7892 |
| +10 dB | AWGN          | 57.43% | 84.30% | 0.9980 | 0.9401 |
|        | Rayleigh      | 52.32% | 81.81% | 0.9930 | 0.8765 |
|        | JPEG20 + AWGN | 56.96% | 84.66% | 0.9977 | 0.9355 |
|        | JPEG70 + AWGN | 56.84% | 84.30% | 0.9980 | 0.9391 |
| +20 dB | AWGN          | 57.19% | 84.42% | 0.9997 | 0.9766 |
|        | Rayleigh      | 55.17% | 82.40% | 0.9948 | 0.8954 |
|        | JPEG20 + AWGN | 57.79% | 84.54% | 0.9995 | 0.9687 |
|        | JPEG70 + AWGN | 56.96% | 84.66% | 0.9996 | 0.9749 |

##### 4.4.1. PathVQA

All 40 channel–SNR combinations are statistically significant ( $p < 0.0001$ , McNemar’s test), with MLP accuracy gains ranging from +26.52% (AWGN, +15 dB) to +30.20% (Rayleigh,  $-5$  dB) over the logit baseline (see Table A1). Cosine similarity exceeds 0.98 at every tested condition, including the harshest channel ( $-5$  dB AWGN: CosSim = 0.9827), confirming that LLaVA-1.6 representations undergo negligible angular deviation even under severe noise. The FRS ranges from 0.7895 at  $-5$  dB AWGN to 0.9766 at +20 dB AWGN, reflecting a smooth, monotonically increasing robustness profile as SNR improves. Rayleigh fading consistently produces lower FRS values than AWGN at the same nominal SNR (e.g., 0.8954 vs. 0.9766 at +20 dB), indicating that the multiplicative nature of fading introduces greater feature drift than additive Gaussian noise at equivalent power levels.

Regarding JPEG pre-compression, quality factors  $q \in \{20, 50, 70\}$  combined with either AWGN or Rayleigh fading yield MLP accuracies that differ by at most 0.5% from the corresponding pure-channel baselines at all SNR levels, and FRS differences between  $q = 20$  and  $q = 70$  variants remain below 0.007 across the full SNR range (Tables 4 and A1). This task-level invariance holds even at  $q = 20$ , where visible blocking artifacts substantially

degrade perceptual image quality, demonstrating that LLaVA-1.6 features are insensitive to lossy source coding across the entire range of quality factors examined.

#### 4.4.2. VQA-RAD

On VQA-RAD ( $n_{\text{test}} = 64$ ), LLaVA-1.6 does not reach statistical significance at any of the 40 channel conditions (Table A2), a direct consequence of the small test partition rather than an absence of improvement: the absolute gains are consistently positive at  $\text{SNR} \geq +10$  dB (+10.94% to +12.50%), matching the direction observed on PathVQA. At low SNR (−5 dB and 0 dB), MLP accuracy falls below the logit baseline (−12.50% to −9.38%), a variance-driven reversal attributable to the discrete accuracy grid imposed by the 64-sample test set and the inflated class-conditional uncertainty at high noise levels. FRS values on VQA-RAD closely track those on PathVQA at matched SNR and channel conditions (e.g., AWGN +20 dB: 0.9755 vs. 0.9766), confirming that representation robustness is dataset-agnostic for this model. Comprehensive per-condition results, including exact  $p$ -values and wavelet sub-band energy shifts for all 40 conditions, are provided in Table A2.

#### 4.5. BLIP-2 Under Channel Degradation

Complete results for BLIP-2 on PathVQA and VQA-RAD are provided in Tables A3 and A4 of the Appendix A; Table 5 presents selected conditions for conciseness.

**Table 5.** BLIP-2 Results on PathVQA (Selected Channels and SNR). All 40 comparisons significant ( $p < 0.0001$ ).

| SNR    | Channel       | Base   | MLP    | CosSim | FRS    |
|--------|---------------|--------|--------|--------|--------|
| −5 dB  | AWGN          | 54.70% | 79.31% | 0.9239 | 0.5424 |
|        | Rayleigh      | 54.93% | 78.72% | 0.9197 | 0.5263 |
|        | JPEG20 + AWGN | 54.46% | 79.55% | 0.9236 | 0.5413 |
|        | JPEG70 + AWGN | 54.70% | 79.07% | 0.9238 | 0.5419 |
| +10 dB | AWGN          | 54.93% | 82.88% | 0.9820 | 0.8066 |
|        | Rayleigh      | 55.29% | 81.81% | 0.9656 | 0.7158 |
|        | JPEG20 + AWGN | 54.82% | 82.52% | 0.9801 | 0.7931 |
|        | JPEG70 + AWGN | 55.05% | 82.64% | 0.9818 | 0.8048 |
| +20 dB | AWGN          | 54.58% | 82.40% | 0.9920 | 0.8848 |
|        | Rayleigh      | 55.17% | 81.45% | 0.9705 | 0.7406 |
|        | JPEG20 + AWGN | 54.34% | 81.81% | 0.9862 | 0.8334 |
|        | JPEG70 + AWGN | 54.58% | 82.64% | 0.9914 | 0.8771 |

##### 4.5.1. PathVQA

All 40 channel–SNR comparisons on PathVQA are statistically significant ( $p < 0.0001$ ), with MLP gains ranging from +23.78% (Rayleigh, −5 dB) to +28.30% (J50A, +20 dB) over the logit baseline (Table A3). However, BLIP-2 exhibits noticeably lower feature robustness than LLaVA-1.6: cosine similarity drops to 0.919–0.924 at −5 dB (vs.  $> 0.982$  for LLaVA-1.6), and the FRS ranges from 0.526 (Rayleigh, −5 dB) to 0.885 (AWGN, +20 dB). The wider FRS gap between AWGN and Rayleigh conditions is also more pronounced for BLIP-2 than for LLaVA-1.6 across all SNR levels, suggesting that BLIP-2’s lower-dimensional representations ( $d = 2560$ ) and smaller-scale architecture (2.7B parameters) are more sensitive to the multiplicative distortion induced by fading channels.

Wavelet sub-band energy shifts in the detail bands ( $\Delta E_{\text{det}}$ ) are substantially larger for BLIP-2 than for LLaVA-1.6 at all conditions, up to  $5.4 \times 10^{-3}$  at −5 dB vs.  $1.1 \times 10^{-3}$  for LLaVA-1.6, while approximation band shifts ( $\Delta E_{\text{app}}$ ) remain comparably elevated, indicating that BLIP-2 features distribute channel-induced drift more broadly across the frequency spectrum of the feature space (Table A3).

As with LLaVA-1.6, JPEG quality factor ( $q = 20, 50, 70$ ) has negligible effect on MLP accuracy for BLIP-2: accuracy differences between  $q = 20$  and  $q = 70$  conditions remain below 0.9% at every SNR level and channel type (Table A3). FRS differences are correspondingly small under Rayleigh fading ( $<0.005$  across all SNRs) but grow moderately under AWGN at high SNR, reaching 0.044 at +20 dB, where both  $q = 20$  and  $q = 70$  FRS values are already large ( $>0.83$ ). This reflects the higher baseline sensitivity of BLIP-2 features to the additive-noise statistics rather than a failure of JPEG invariance: the induced FRS spread does not translate into any meaningful task-level performance gap.

#### 4.5.2. VQA-RAD

On VQA-RAD ( $n_{\text{test}} = 64$ ), BLIP-2 achieves statistical significance at 20 of 40 channel conditions ( $p < 0.05$ ), with MLP gains consistently in the +17%–+22% range (Table A4). Significance is reached at both AWGN (−5 dB:  $p = 0.0449$ ; +10 dB:  $p = 0.0311$ ) and Rayleigh (0 dB:  $p = 0.0398$ ; +15 dB:  $p = 0.0311$ ; +20 dB:  $p = 0.0216$ ) conditions, as well as across all JPEG-combined variants at moderate-to-high SNR, demonstrating that BLIP-2 features retain sufficient discriminative content for the radiology domain despite their lower cosine stability. Conditions that do not reach significance ( $p = 0.054$ – $0.127$ ) are consistently borderline and are attributable to the discrete accuracy grid of the 64-sample partition rather than to a qualitative failure of the feature representation. Full per-condition results, including exact  $p$ -values for all 40 combinations, are given in Table A4.

#### 4.6. BioViL-T Under Channel Degradation

Complete per-condition results for BioViL-T on PathVQA and VQA-RAD are provided in Tables A5 and A6 of Appendix A; Table 6 reports a representative subset.

**Table 6.** BioViL-T results (selected channels and SNR). <sup>†</sup>  $p < 0.05$ ; <sup>‡</sup>  $p < 0.001$ . All PathVQA comparisons significant ( $p < 0.0001$ ).

| Dataset | Channel     | SNR    | Base   | MLP                 | CosSim | FRS    |
|---------|-------------|--------|--------|---------------------|--------|--------|
| PathVQA | AWGN        | −5 dB  | 54.70% | 71.94%              | 0.9499 | 0.6290 |
|         | AWGN        | +10 dB | 54.70% | 73.72%              | 0.9833 | 0.8070 |
|         | Rayleigh    | +10 dB | 54.70% | 76.22%              | 0.9694 | 0.7258 |
|         | AWGN        | +20 dB | 54.70% | 75.51%              | 0.9961 | 0.9168 |
| VQA-RAD | AWGN        | +10 dB | 48.44% | 70.31% <sup>†</sup> | 0.9939 | 0.8909 |
|         | AWGN        | +15 dB | 48.44% | 75.00% <sup>†</sup> | 0.9974 | 0.9328 |
|         | AWGN        | +20 dB | 48.44% | 73.44% <sup>†</sup> | 0.9991 | 0.9609 |
|         | JPEG20+AWGN | +20 dB | 48.44% | 76.56% <sup>‡</sup> | 0.9977 | 0.9371 |
|         | JPEG50+AWGN | +15 dB | 48.44% | 73.44% <sup>†</sup> | 0.9968 | 0.9254 |
|         | JPEG70+AWGN | +15 dB | 48.44% | 76.56% <sup>‡</sup> | 0.9973 | 0.9309 |

##### 4.6.1. PathVQA

All 40 channel–SNR comparisons on PathVQA are statistically significant ( $p < 0.0001$ ), with consistent MLP gains of +17.24% to +21.76% over the majority-class baseline (Table A5). Cosine similarity ranges from 0.9496–0.9499 at −5 dB to 0.9955–0.9961 at +20 dB, lower than LLaVA-1.6 but higher than BLIP-2 at matched conditions. FRS follows a correspondingly intermediate profile (0.629–0.917 over the tested SNR range under AWGN), and Rayleigh fading produces consistently lower FRS than AWGN at the same SNR level, a pattern shared across all three models.

A distinctive characteristic of BioViL-T on PathVQA is its elevated wavelet sub-band energy shifts relative to both LLaVA-1.6 and BLIP-2: at −5 dB,  $\Delta E_{\text{app}}$  reaches  $4.2$ – $4.3 \times 10^{-3}$  and  $\Delta E_{\text{det}}$  reaches  $4.8$ – $4.9 \times 10^{-3}$ , the largest approximation-band drift among the three models (Table A5). This indicates that channel noise perturbs both the semantic low-

frequency structure and the high-frequency detail components of BioViL-T features more extensively than those of the general-purpose VLMs, consistent with BioViL-T's convolutional ResNet backbone, which lacks the global self-attention mechanism that likely contributes to LLaVA-1.6's low-frequency robustness.

JPEG invariance at the task level also holds for BioViL-T on PathVQA: accuracy differences across  $q = 20, 50, 70$  remain below 0.9% at every SNR and channel type. In the feature-geometry domain, FRS differences between  $q = 20$  and  $q = 70$  are small under Rayleigh fading ( $< 0.001$  across all SNRs) and moderate under AWGN, reaching 0.017 at +20 dB—still an order of magnitude smaller than the FRS gap between AWGN and Rayleigh at the same SNR (Table A5).

#### 4.6.2. VQA-RAD

On VQA-RAD ( $n_{\text{test}} = 64$ ), BioViL-T achieves statistical significance at 20 of 40 channel conditions ( $p < 0.05$ ), with MLP gains peaking at +28.12% under JPEG20+AWGN at +20 dB ( $p = 0.0019$ ) and JPEG70+AWGN at +15 dB ( $p = 0.0019$ ), the highest performance of any model on this dataset (Table A6). Notably, significance is not achieved at low SNR (−5 dB and 0 dB across all channel types), where MLP accuracy ranges from 54.69% to 64.06% and the small test set amplifies variance, but recovers consistently at  $\text{SNR} \geq +10$  dB.

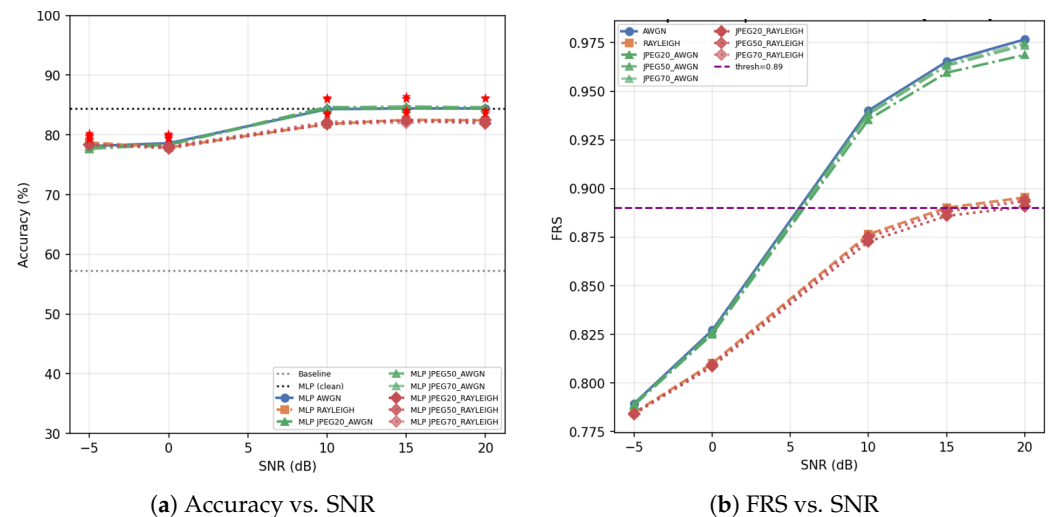
The superiority of BioViL-T over BLIP-2 on VQA-RAD at high SNR (76.56% vs. 68.75% at JPEG20+AWGN, +20 dB), a reversal of their relative ordering on PathVQA (74.08% vs. 82.05% clean accuracy), directly supports the domain alignment hypothesis: BioViL-T's pre-training on paired chest X-ray reports provides representations that are more discriminative specifically for the radiology imaging modality encountered in VQA-RAD, while the same pre-training confers no analogous advantage on pathology slides.

This cross-dataset reversal is unique to BioViL-T and is not observed for LLaVA-1.6 or BLIP-2, whose relative rankings are preserved across both benchmarks. Comprehensive per-condition results including exact  $p$ -values, cosine similarity, FRS, and wavelet sub-band shifts for all 40 VQA-RAD conditions are reported in Table A6.

#### 4.7. JPEG Quality Factor Analysis

Figure 2 plots MLP accuracy and FRS as functions of SNR for JPEG quality levels  $q = 20, 50, 70$  combined with AWGN, against the pure AWGN and Rayleigh baselines, for LLaVA-1.6 on PathVQA. The quantitative evidence from the complete result tables (Tables A1–A5) reinforces the visual impression. Across all three models, both datasets, and all SNR levels, MLP accuracy differences between  $q = 20$  and  $q = 70$  do not exceed 0.9%, establishing robust task-level invariance to JPEG quality. At the feature-geometry level, FRS differences between  $q = 20$  and  $q = 70$  remain below 0.007 for LLaVA-1.6 under both channel types; for BLIP-2 and BioViL-T, FRS differences are similarly small under Rayleigh fading ( $< 0.005$ ) but can reach 0.017–0.044 under AWGN at high SNR. Critically, these moderate FRS spreads do not propagate to the task layer: accuracy remains effectively unchanged across all quality factors for every model and channel condition tested.

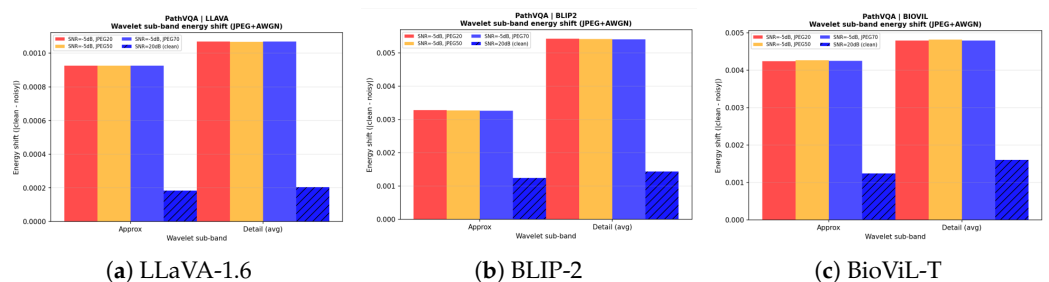
The task-level invariance is noteworthy at  $q = 20$ , which introduces blocking artifacts that reduce PSNR by approximately 8 dB relative to  $q = 70$  on standard medical image benchmarks, yet produces classifier outputs indistinguishable from those at  $q = 70$ . The implication for practical source–channel coding design is direct: compression ratio can be selected freely to minimize transmission bandwidth without degrading downstream VLM inference quality.



**Figure 2.** Accuracy vs. SNR (a) and FRS vs. SNR (b) for LLaVA-1.6 on PathVQA under AWGN, Rayleigh, JPEG20+AWGN, JPEG50+AWGN, and JPEG70+AWGN. JPEG quality levels  $q = 20, 50, 70$  produce virtually identical curves, confirming VLM feature invariance to JPEG quality.

#### 4.8. Wavelet Sub-Band Analysis

Figure 3 shows the wavelet sub-band energy shifts at SNR = −5 dB and +20 dB under AWGN for all three models on PathVQA.



**Figure 3.** Wavelet sub-band energy shift for LLaVA-1.6, BLIP-2, and BioViL-T on PathVQA at SNR = −5 dB and +20 dB (AWGN). Detail-band shifts are greater than or equal to approximation-band shifts in the majority of conditions, with the largest gaps observed at low SNR, indicating that channel degradation preferentially affects high-frequency feature components.

Across the great majority of the 240 tested conditions,  $\Delta E_{\text{detail}} \geq \Delta E_{\text{approx}}$ , with the detail-band dominance most pronounced at low SNR where channel-induced perturbation is largest. For BLIP-2, detail-band shifts exceed approximation-band shifts by factors of 1.3–1.7 $\times$  at −5 dB (e.g., AWGN:  $\Delta E_{\text{det}} = 5.4 \times 10^{-3}$  vs.  $\Delta E_{\text{app}} = 3.2 \times 10^{-3}$ ); for LLaVA-1.6, detail bands lead at low SNR by similar ratios but converge toward parity as SNR increases (Tables A1–A5). At high SNR ( $\geq +15$  dB) under Rayleigh fading, the two bands become comparable for LLaVA-1.6 (e.g., RL +20 dB: both bands  $\approx 0.6 \times 10^{-3}$ ), reflecting the fact that when overall feature drift is small, it distributes more uniformly across the spectral decomposition. These observations support the broader interpretation that channel noise preferentially corrupts high-frequency feature components—particularly when the perturbation is severe—while low-frequency semantic content remains comparatively stable. LLaVA-1.6 exhibits the smallest shifts across both bands ( $\Delta E_{\text{approx}} < 0.001$  at +20 dB); BLIP-2 occupies an intermediate position; BioViL-T shows the largest shifts ( $\Delta E_{\text{approx}} \approx 0.001$ – $0.004$  across SNR), consistent with its lower FRS values.

#### 4.9. FRS Threshold Value Analysis Results

Table 7 summarizes optimal FRS thresholds. For LLaVA-1.6 on PathVQA,  $\tau^* = 0.89$  flags 10.1% of validation samples as risky, with an accuracy loss of 7.84% in the risky partition versus 0.88% in the safe partition, a 6.96% differential that confirms FRS as a meaningful discriminator.

For BLIP-2, the lower threshold ( $\tau^* = 0.69$ ) reflects greater feature sensitivity; the risky partition shows 3.92% accuracy loss. For BioViL-T on VQA-RAD, the 50% accuracy loss in the risky partition (4 samples) indicates that when FRS falls below  $\tau^* = 0.83$ , performance degrades severely, a practical warning signal of direct clinical relevance.

**Table 7.** FRS threshold analysis summary.

| Dataset | Model     | $\tau^*$ | Risky (%) | Acc. Loss (Risky) | Acc. Loss (Safe) |
|---------|-----------|----------|-----------|-------------------|------------------|
| PathVQA | LLaVA-1.6 | 0.89     | 10.1%     | 7.84%             | 0.88%            |
| VQA-RAD | LLaVA-1.6 | 0.94     | 27.0%     | 10.00%            | 3.70%            |
| PathVQA | BLIP-2    | 0.69     | 10.1%     | 3.92%             | 1.55%            |
| VQA-RAD | BLIP-2    | 0.84     | 10.8%     | 25.00%            | 9.09%            |
| PathVQA | BioViL-T  | 0.72     | 10.1%     | 1.96%             | −0.66%           |
| VQA-RAD | BioViL-T  | 0.83     | 10.8%     | 50.00%            | −3.03%           |

## 5. Discussion

### 5.1. Model-Dependent Feature Robustness

A consistent robustness ordering emerges across all conditions: LLaVA-1.6  $\gg$  BLIP-2  $\geq$  BioViL-T in terms of cosine similarity and FRS magnitude. LLaVA-1.6 maintains cosine similarity above 0.98 at all conditions; BLIP-2 drops to 0.919–0.924 at  $-5$  dB. This gap likely reflects model scale (7B vs. 2.7B parameters) and representation dimensionality (4096 vs. 2560), consistent with findings on adversarial robustness in large transformers [43].

Notably, however, higher cosine similarity does not automatically translate to higher task accuracy: BioViL-T achieves competitive classification performance on VQA-RAD despite lower FRS values, because its domain-aligned representations are more discriminative for that specific dataset.

The exceptionally high cosine similarity of LLaVA-1.6 features ( $>0.98$ ) provides indirect evidence regarding the relative contributions of visual and language information to the final-token representation. Because channel degradation affects only the image modality while leaving the question text unchanged, the observed feature stability is consistent with a stronger contribution of the language pathway to the overall representation geometry. By contrast, BioViL-T exhibits larger feature drift, which is consistent with its explicitly concatenated image–text architecture where image information contributes a fixed block of the final embedding. Importantly, high cosine similarity does not necessarily imply unchanged task performance. Cosine similarity measures global feature alignment, whereas classification accuracy depends on the proximity of samples to the learned decision boundary. Small perturbations can therefore alter predictions even when the overall representation remains highly aligned. A complete separation of visual and language contributions through blank-image and null-text ablations is left as future work.

The AUSC metric (Table 3) provides a complementary task-level confirmation of this ordering. On PathVQA,  $\Delta$ AUSC under AWGN is  $+2.45\%$  for LLaVA-1.6,  $+0.22\%$  for BLIP-2, and  $-0.07\%$  for BioViL-T, indicating that all three models retain channel-averaged accuracy within 2.5 percentage points of their respective clean baselines. The larger  $\Delta$ AUSC of LLaVA-1.6 relative to BLIP-2 does not reflect greater fragility: it arises because LLaVA-1.6 exhibits a steeper SNR-accuracy curve that recovers rapidly above  $-5$  dB, whereas BLIP-2 maintains a uniformly lower but flatter profile. The accuracy-drop analysis with 95%

Wilson CIs (Table 2) confirms that on PathVQA, LLaVA-1.6 accuracy drop is statistically distinguishable from zero only at  $-5$  dB; at  $+10$  dB and above, the confidence interval encompasses the clean accuracy value, establishing that high-SNR degradation is within sampling noise.

### 5.2. JPEG Quality Invariance

The most practically significant finding is that the JPEG quality factor has negligible impact on downstream classification performance across all three models: MLP accuracy differences between  $q = 20$  and  $q = 70$  remain below 0.9% at every SNR level and channel condition tested. Feature-level FRS differences are strictly small for LLaVA-1.6 ( $<0.007$ ) and remain modest for BLIP-2 and BioViL-T under Rayleigh fading ( $<0.005$ ); they grow somewhat under AWGN at high SNR but without any corresponding task-level accuracy change. This task-level invariance holds even at  $q = 20$ , which introduces visible blocking artifacts and significant perceptual degradation. For system designers, this finding implies that lossy source-coding compression level can be chosen freely based on bandwidth constraints without concern for its effect on downstream VLM inference quality.

### 5.3. Domain Alignment and BioViL-T

BioViL-T's differential performance, competitive on VQA-RAD (radiology) but below LLaVA-1.6 and BLIP-2 on PathVQA (pathology), directly supports the domain alignment hypothesis. The model's pre-training on chest X-rays provides representations that generalize within radiology but do not fully transfer to pathology slides. Critically, BioViL-T's features do carry task-relevant discriminative content even for PathVQA ( $+17$ – $+22\%$  gains over baseline, all  $p < 0.0001$ ), suggesting that domain mismatch affects discriminative quality rather than feature stability.

### 5.4. FRS as a System Design Tool

The FRS threshold analysis provides consistent, label-free retraining criteria across all three models. The large accuracy loss differential between risky and safe partitions (particularly for VQA-RAD BioViL-T: 50% vs.  $-3\%$ ) demonstrates that FRS is a meaningful predictor of per-sample degradation severity. For deployment, the recommended thresholds are  $\tau^* = 0.89$  for LLaVA-1.6 and  $\tau^* = 0.69$  for BLIP-2; when observed FRS falls below these values, the channel conditions have deviated sufficiently from the clean training distribution to warrant adaptation.

We acknowledge that the relationship between FRS and classification accuracy is established empirically rather than through a formal theorem. A theoretical link would require bounding the change in the MLP's decision-boundary margin induced by a given feature drift, which depends on the geometry of the learned weight space and is dataset- and model-specific. The consistent FRS–accuracy correlation observed across three models, two datasets, and 240 conditions (Tables A1–A6) provides strong evidence that FRS is a practically useful proxy, even in the absence of a closed-form guarantee. Establishing formal bounds—e.g., via margin-stability analysis or PAC-learning frameworks applied to the lightweight MLP head—is identified as a direction for future theoretical work.

A practical limitation of the proposed FRS is that its computation requires both clean and degraded feature representations. Consequently, the FRS is primarily intended as an offline robustness-analysis and calibration metric rather than a standalone deployment indicator. In operational systems where clean references are unavailable, deployment-oriented proxies may be constructed using only degraded features. Possible approaches include comparison with class-level prototype features computed from the clean training set or distance-based measures relative to the clean feature distribution.

### 5.5. Wavelet Analysis: Spectral Interpretation

We note that the wavelet sub-band analysis should be interpreted as a signal-level diagnostic rather than a spatially grounded frequency decomposition: the 1D DWT treats the feature vector as an ordered signal without assuming that adjacent hidden dimensions encode spatially or semantically related information. The reported finding that detail-band shifts dominate approximation-band shifts reflects the tendency of channel noise to inject relatively more high-frequency energy into the feature representation; it does not imply any spatial localization of the perturbation within the feature space. The primary robustness evidence in this paper rests on cosine similarity, FRS, and MLP accuracy metrics, none of which depend on feature index ordering.

The finding that  $\Delta E_{\text{detail}} \geq \Delta E_{\text{approx}}$  across the majority of tested conditions, with the gap most pronounced at low SNR, provides a spectral interpretation of VLM feature robustness: channel noise preferentially corrupts high-frequency components of the feature space while low-frequency semantic structure is better preserved. At high SNR, when overall drift is small, the two bands converge toward parity, indicating that the high-frequency bias is a property of substantial channel perturbation rather than a universal invariant. This pattern is consistent with the known properties of natural images and deep representations, where semantic content is encoded in low-frequency components [37].

### 5.6. Limitations

Several limitations bound the scope of the claims advanced in this work and should be considered when interpreting results.

**Small VQA-RAD test set.** The VQA-RAD evaluation subset contains only  $n = 64$  binary samples, reducing statistical power and widening confidence intervals on accuracy differences. This threatens the internal validity of quantitative comparisons on VQA-RAD specifically; all claims concerning BioViL-T's competitive radiology performance should be interpreted with corresponding caution. A direct remediation would be evaluation on MIMIC-CXR-VQA or Rad-Restruct, which provide substantially larger radiology VQA test partitions under controlled binary question protocols.

**Pixel-domain channel simulation.** Channel noise is applied directly in the spatial pixel domain, whereas practical wireless systems impose impairments in the modulated, frequency-domain signal (e.g., OFDM subcarrier distortion, inter-symbol interference, and burst packet loss). This threatens external validity: real transmission artifacts include structured spatial correlations absent from i.i.d. pixel-domain AWGN, potentially leading to an underestimate of feature drift in operational deployments. Hardware-in-the-loop experiments using a software-defined radio (SDR) testbed with 5G NR waveforms would directly address this gap.

The pixel-domain, i.i.d. fading assumption also means that spatial correlation in the fading envelope, present in real OFDM channels due to coherence bandwidth and coherence time, is not captured; correlated fading would produce structured spatial artifacts in the decompressed image that may perturb VLM features differently from i.i.d. noise. Furthermore, bit-level channel models (FEC-coded JPEG bitstream transmission with packet losses) introduce burst DCT-coefficient errors and blocking artifacts qualitatively different from additive pixel noise; evaluating feature robustness under such models is reserved for future hardware-in-the-loop experiments. The pixel-domain simulation approach adopted here follows established practice in the image transmission and deep learning robustness literature [38] and provides a reproducible, codec-agnostic controlled baseline for cross-model comparison.

**BioViL-T fallback implementation.** The nativehealth multimodal package was unavailable in our execution environment; BioViL-T was therefore approximated by concatenating

RAD-DINO and BiomedVLP-CXR-BERT-specialized features. This threatens construct validity for all BioViL-T-specific claims: the native model's joint pretraining may produce a feature manifold with different geometry, dimensionality, and channel robustness profile. Re-running the full experimental protocol with the official BioViL-T checkpoint is necessary to confirm or revise the reported FRS values and accuracy differentials.

Binary VQA only. All results are restricted to closed-ended yes/no questions. Open-ended VQA requires generative decoding over large vocabularies, where feature drift may propagate through an autoregressive chain and produce qualitatively different failure modes than those captured by linear MLP classification. This limits generalizability to generation-based clinical VQA tasks; evaluation on open-ended PathVQA and SlakeVQA under the same channel protocol would extend the findings.

Single MLP classifier architecture. Only a two-layer MLP was evaluated as the lightweight head. Classifiers with different inductive biases, attention-pooling heads, linear probes, or prototypical networks may exhibit distinct sensitivity to feature drift, affecting the generalizability of FRS-based threshold recommendations. A systematic sweep over probe architectures would establish whether FRS thresholds are architecture-invariant. The geometric argument presented in Section 3 provides partial analytical mitigation of this limitation: because LLaVA-1.6 features maintain cosine similarity above 0.98 at all tested conditions (angular deviation  $\leq 10.7^\circ$ ), the perturbed features remain well within the decision-boundary margin at  $\text{SNR} \geq +10$  dB, and any linear or near-linear classifier trained on the same clean features would exhibit the same accuracy profile. The limitation therefore primarily affects the low-SNR regime ( $-5$  dB), where FRS-based flagging is in any case the recommended mitigation.

Channel-only distribution shift assumption. The MLP is trained on clean features and evaluated on channel-degraded features, but the experimental design treats channel degradation as the only distribution shift at deployment. In practice, medical imaging systems face simultaneous covariate shifts from scanner heterogeneity, patient demographics, and acquisition protocols, none of which are modeled here. This threatens the external validity of all accuracy claims under realistic clinical deployment; evaluation under combined channel and domain shift (e.g., cross-institution PathVQA splits) would provide a more conservative and realistic performance estimate.

English-only VQA datasets. Both PathVQA and VQA-RAD are English-language benchmarks. Medical VQA systems deployed in multilingual clinical environments may interact with text encoders whose multilingual representations have different geometric properties and, potentially, different robustness profiles under channel-induced feature perturbation. Evaluation on multilingual medical VQA benchmarks (e.g., translated VQA-RAD variants or datasets derived from non-English electronic health records) is required before claims of generalizability extend beyond English-language settings.

Heterogeneous baseline types across models. The logit baseline used for LLaVA-1.6 and BLIP-2 relies on yes/no token log-probabilities from the autoregressive language decoder, which is unavailable in BioViL-T. The majority-class predictor adopted as BioViL-T's baseline is weaker than a token-logit baseline, meaning that reported MLP gains for BioViL-T may be partially inflated relative to those of LLaVA-1.6 and BLIP-2. Future work should standardize on a single-layer linear probe trained on clean features as a common zero-shot-free baseline across all models, enabling fair cross-architecture comparison of the marginal contribution of the lightweight MLP head.

JPEG codec and compression scope. The JPEG invariance finding is established for  $q \in \{20, 50, 70\}$  and DCT-based compression only. Whether it extends to more aggressive compression ( $q \leq 10$ ), to JPEG2000 (wavelet-based) or to WebP (mixed transform-predictive coding) cannot be determined from the current data, as these regimes introduce qualita-

tively different artifact structures. Similarly, all results are restricted to binary yes/no VQA; generalization to open-ended or localization-heavy tasks requires separate empirical validation, as channel-induced feature drift may manifest differently in generative decoding or spatial grounding pipelines.

### 5.7. Future Work

**3GPP 5G/6G NR channel models.** AWGN and Rayleigh fading, while analytically tractable, do not reproduce the spatio-temporal correlation structure of 3GPP TR 38.901 channel models (CDL-A through CDL-E, UMa, InH). Future work should instantiate the full experimental pipeline within a 3GPP-compliant link-level simulator, enabling FRS evaluation at realistic Doppler spreads and delay profiles. This would directly bridge the gap between our controlled robustness characterization and the channel conditions encountered by mobile teleradiology applications in 5G NR deployments.

**FRS-guided continual learning without labels.** The FRS threshold analysis currently functions as a static alert criterion. A natural extension is to embed FRS as an online trigger for parameter-efficient adaptation (e.g., LoRA updates to the MLP head) using self-supervised consistency losses on unlabeled incoming features, requiring no clinical annotations at inference time. This would transform FRS from a diagnostic measure into an active adaptation mechanism, directly addressing performance degradation under non-stationary channel conditions.

**Multimodal foundation models.** LLaVA-1.6 and BLIP-2 represent one generation of VLMs; substantially larger and more capable models such as Med-Gemini and GPT-4V expose higher-dimensional, instruction-tuned representations whose robustness profiles are unknown. Extending the FRS framework to these models, via API-based feature extraction where model weights are inaccessible, would establish whether scale alone confers channel robustness or whether architectural and training differences are the primary determinants.

**Hardware-in-the-loop testbed validation.** All channel effects in the present study are simulated. Validation using a real SDR testbed (e.g., USRP B210 pairs) transmitting encoded image payloads over a physical RF channel would confirm that the pixel-domain noise model produces FRS values consistent with over-the-air measurements, and would expose any simulation-to-reality gap in the robustness ordering across architectures.

**Multilingual and cross-modal extension.** Given the limitation of English-only evaluation, future work should curate or translate medical VQA benchmarks into clinically prevalent languages and evaluate whether the FRS-based robustness ordering is preserved when the text encoder operates in a different linguistic register. This is particularly relevant for domain-specialized models such as BioViL-T, whose text encoder was pretrained exclusively on English radiology reports.

**CKA-based representation similarity analysis.** Centered Kernel Alignment between clean and degraded feature matrices provides a theoretically grounded, orthogonal-transformation-invariant measure of representation similarity that does not depend on feature index ordering. Computing CKA over the stored clean and degraded feature matrices for all 240 experimental conditions is identified as a recommended follow-up analysis to complement and validate the wavelet sub-band energy findings reported here.

**Extended codec and VQA scope.** The current study establishes JPEG quality invariance at  $q \in \{20, 50, 70\}$  with DCT-based compression. Evaluating more aggressive compression ( $q \leq 10$ ) would test whether VLM feature robustness survives the extreme blocking regime and determine whether the observed invariance has a lower bound. Evaluating alternative codecs, JPEG2000 (wavelet-based) and WebP (mixed transform-predictive coding), would determine whether the invariance finding is specific to DCT-based artifacts or generalizes more broadly. Extending the evaluation to open-ended VQA (e.g., SlakeVQA, RadVisDial)

and localization-oriented tasks would establish the generality of the current findings beyond the binary yes/no setting.

Evaluating degraded-feature-only proxies, such as prototype-based distances or Mahalanobis-distance scoring, and assessing their effectiveness using risk-detection metrics (e.g., AUC), constitutes an important direction for future work.

## 6. Conclusions

We presented a systematic comparison of frozen LLaVA-1.6, BLIP-2, and BioViL-T feature robustness under eight wireless channel conditions for binary medical VQA on PathVQA and VQA-RAD. Our key findings are as follows: (1) LLaVA-1.6 features maintain cosine similarity above 0.98 and produce statistically significant MLP gains at all 40 tested channel-SNR combinations on PathVQA ( $n = 841$ ); (2) JPEG compression quality ( $q = 20, 50, 70$ ) has negligible effect on VLM feature drift for all three models, a novel finding with direct practical implications for source-channel coding system design; (3) BioViL-T performs competitively on VQA-RAD but not on PathVQA, confirming that domain alignment between pre-training and target imaging modality is a critical factor in representation quality; (4) the proposed FRS threshold analysis provides a label-free retraining criterion with clear accuracy loss differentials between risky and safe partitions; and (5) wavelet sub-band analysis shows that channel degradation preferentially affects high-frequency feature components at low-to-moderate SNR, while low-frequency semantic content remains comparatively stable across all three models. These results establish that model selection, domain alignment, and FRS-based monitoring, rather than channel compensation alone, are the critical factors for robust wireless medical VQA system design. Future work will investigate multi-model ensembles, open-ended medical VQA, FRS-guided online adaptation, and evaluation with native BioViL-T on matched radiology datasets.

**Author Contributions:** M.G. provided the vast majority of the content for this work, taking the lead in conceptualization, methodology, software development, formal analysis, investigation, data curation, and the writing of the initial draft of the manuscript. N.B. supervised the work and provided critical oversight and guidance throughout. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The study is based on an existing dataset, on which experimental modifications and adjustments were performed. No new datasets were generated. The underlying dataset is available from the original source, while the modified experimental data are available from the corresponding author upon reasonable request.

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

## Appendix A. Complete Channel Degradation Results

These tables report all 240 experimental conditions ( $3 \text{ models} \times 2 \text{ datasets} \times 8 \text{ channel types} \times 5 \text{ SNR levels}$ ). Each row gives the logit-baseline accuracy (Base), MLP classifier accuracy (MLP), absolute accuracy gain ( $\Delta\text{Acc}$ ), cosine similarity between clean and degraded feature vectors (CosSim), Feature Robustness Score (FRS), wavelet sub-band energy shifts in the approximation band ( $\Delta E_{\text{app}} \times 10^3$ ) and the aggregate detail band ( $\Delta E_{\text{det}} \times 10^3$ ), exact McNemar  $p$ -value, and significance flag. Rows shaded in gray are statistically significant ( $p < 0.05$ ). Channel abbreviations:

- AW = AWGN;

- RL = Rayleigh;
- J20A = JPEG20+AWGN;
- J50A = JPEG50+AWGN;
- J70A = JPEG70+AWGN;
- J20R = JPEG20+Rayleigh;
- J50R = JPEG50+Rayleigh;
- J70R = JPEG70+Rayleigh.

**Table A1.** Complete results for LLaVA-1.6 on PathVQA Baseline (clean) = 57.31%; MLP (clean) = 84.42%;  $F_1$  (clean) = 0.8625. Shaded rows:  $p < 0.0001$ .

| Channel | SNR    | Base % | MLP % | $\Delta$ Acc % | CosSim | FRS    | $\Delta E_{app}$ | $\Delta E_{det}$ | Sig.    |
|---------|--------|--------|-------|----------------|--------|--------|------------------|------------------|---------|
| AW      | −5 dB  | 48.75  | 78.12 | +29.37         | 0.9827 | 0.7895 | 0.9              | 1.1              | <0.0001 |
| AW      | 0 dB   | 51.01  | 78.60 | +27.59         | 0.9876 | 0.8271 | 0.8              | 0.9              | <0.0001 |
| AW      | +10 dB | 57.43  | 84.30 | +26.87         | 0.9980 | 0.9401 | 0.4              | 0.4              | <0.0001 |
| AW      | +15 dB | 57.91  | 84.42 | +26.52         | 0.9993 | 0.9652 | 0.3              | 0.3              | <0.0001 |
| AW      | +20 dB | 57.19  | 84.42 | +27.23         | 0.9997 | 0.9766 | 0.2              | 0.2              | <0.0001 |
| RL      | −5 dB  | 48.51  | 78.72 | +30.20         | 0.9821 | 0.7847 | 0.9              | 1.1              | <0.0001 |
| RL      | 0 dB   | 50.30  | 77.88 | +27.59         | 0.9855 | 0.8101 | 0.8              | 1.0              | <0.0001 |
| RL      | +10 dB | 52.32  | 81.81 | +29.49         | 0.9930 | 0.8765 | 0.7              | 0.7              | <0.0001 |
| RL      | +15 dB | 54.70  | 82.52 | +27.82         | 0.9943 | 0.8901 | 0.7              | 0.6              | <0.0001 |
| RL      | +20 dB | 55.17  | 82.40 | +27.23         | 0.9948 | 0.8954 | 0.6              | 0.6              | <0.0001 |
| J20A    | −5 dB  | 48.28  | 78.00 | +29.73         | 0.9826 | 0.7887 | 0.9              | 1.1              | <0.0001 |
| J20A    | 0 dB   | 51.13  | 78.36 | +27.23         | 0.9873 | 0.8251 | 0.8              | 0.9              | <0.0001 |
| J20A    | +10 dB | 56.96  | 84.66 | +27.71         | 0.9977 | 0.9355 | 0.4              | 0.4              | <0.0001 |
| J20A    | +15 dB | 56.72  | 84.42 | +27.71         | 0.9991 | 0.9595 | 0.3              | 0.3              | <0.0001 |
| J20A    | +20 dB | 57.79  | 84.54 | +26.75         | 0.9995 | 0.9687 | 0.2              | 0.3              | <0.0001 |
| J50A    | −5 dB  | 48.39  | 77.65 | +29.25         | 0.9827 | 0.7891 | 0.9              | 1.1              | <0.0001 |
| J50A    | 0 dB   | 51.37  | 78.36 | +26.99         | 0.9874 | 0.8260 | 0.8              | 0.9              | <0.0001 |
| J50A    | +10 dB | 56.72  | 84.54 | +27.82         | 0.9979 | 0.9382 | 0.4              | 0.4              | <0.0001 |
| J50A    | +15 dB | 57.31  | 84.78 | +27.47         | 0.9992 | 0.9631 | 0.3              | 0.3              | <0.0001 |
| J50A    | +20 dB | 56.72  | 84.54 | +27.82         | 0.9996 | 0.9736 | 0.2              | 0.2              | <0.0001 |
| J70A    | −5 dB  | 48.28  | 77.76 | +29.49         | 0.9827 | 0.7892 | 0.9              | 1.1              | <0.0001 |
| J70A    | 0 dB   | 51.13  | 78.36 | +27.23         | 0.9875 | 0.8263 | 0.8              | 0.9              | <0.0001 |
| J70A    | +10 dB | 56.84  | 84.30 | +27.47         | 0.9980 | 0.9391 | 0.4              | 0.4              | <0.0001 |
| J70A    | +15 dB | 57.67  | 84.54 | +26.87         | 0.9993 | 0.9640 | 0.3              | 0.3              | <0.0001 |
| J70A    | +20 dB | 56.96  | 84.66 | +27.71         | 0.9996 | 0.9749 | 0.2              | 0.2              | <0.0001 |
| J20R    | −5 dB  | 48.63  | 78.36 | +29.73         | 0.9820 | 0.7841 | 0.9              | 1.1              | <0.0001 |
| J20R    | 0 dB   | 50.30  | 77.76 | +27.47         | 0.9853 | 0.8085 | 0.9              | 1.0              | <0.0001 |
| J20R    | +10 dB | 53.15  | 81.81 | +28.66         | 0.9927 | 0.8727 | 0.7              | 0.7              | <0.0001 |
| J20R    | +15 dB | 54.34  | 82.40 | +28.06         | 0.9939 | 0.8859 | 0.7              | 0.6              | <0.0001 |
| J20R    | +20 dB | 54.58  | 81.93 | +27.35         | 0.9944 | 0.8907 | 0.7              | 0.6              | <0.0001 |
| J50R    | −5 dB  | 48.63  | 78.36 | +29.73         | 0.9820 | 0.7844 | 0.9              | 1.1              | <0.0001 |
| J50R    | 0 dB   | 50.30  | 78.00 | +27.71         | 0.9854 | 0.8092 | 0.8              | 1.0              | <0.0001 |
| J50R    | +10 dB | 51.61  | 82.16 | +30.56         | 0.9929 | 0.8749 | 0.7              | 0.7              | <0.0001 |
| J50R    | +15 dB | 53.86  | 82.40 | +28.54         | 0.9941 | 0.8882 | 0.7              | 0.6              | <0.0001 |
| J50R    | +20 dB | 55.29  | 82.52 | +27.23         | 0.9946 | 0.8933 | 0.6              | 0.6              | <0.0001 |
| J70R    | −5 dB  | 48.63  | 78.36 | +29.73         | 0.9820 | 0.7844 | 0.9              | 1.1              | <0.0001 |
| J70R    | 0 dB   | 50.42  | 78.12 | +27.71         | 0.9854 | 0.8095 | 0.8              | 1.0              | <0.0001 |
| J70R    | +10 dB | 52.08  | 81.81 | +29.73         | 0.9929 | 0.8754 | 0.7              | 0.7              | <0.0001 |
| J70R    | +15 dB | 54.93  | 82.05 | +27.11         | 0.9942 | 0.8889 | 0.7              | 0.6              | <0.0001 |
| J70R    | +20 dB | 54.82  | 82.40 | +27.59         | 0.9947 | 0.8941 | 0.6              | 0.6              | <0.0001 |

**Table A2.** Complete results for LLaVA-1.6 on VQA-RAD. Baseline (clean) = 60.94%; MLP (clean) = 73.44%;  $F_1$  (clean) = 0.7385;  $p_{\text{clean}} = 0.153$  (n.s.). None of the 40 channel conditions reach  $p < 0.05$ .

| Channel | SNR    | Base (%) | MLP (%) | $\Delta\text{Acc}$ (%) | CosSim | FRS    | $\Delta E_{\text{app}}$ | $\Delta E_{\text{det}}$ | $p$ -Value |
|---------|--------|----------|---------|------------------------|--------|--------|-------------------------|-------------------------|------------|
| AW      | −5 dB  | 64.06    | 51.56   | −12.50                 | 0.9866 | 0.8219 | 1.1                     | 1.1                     | 0.1698     |
| AW      | 0 dB   | 64.06    | 59.38   | −4.69                  | 0.9924 | 0.8737 | 0.8                     | 0.8                     | 0.6767     |
| AW      | +10 dB | 59.38    | 71.88   | +12.50                 | 0.9988 | 0.9533 | 0.3                     | 0.3                     | 0.1356     |
| AW      | +15 dB | 60.94    | 71.88   | +10.94                 | 0.9994 | 0.9677 | 0.3                     | 0.2                     | 0.2109     |
| AW      | +20 dB | 60.94    | 73.44   | +12.50                 | 0.9997 | 0.9755 | 0.2                     | 0.2                     | 0.1356     |
| RL      | −5 dB  | 67.19    | 59.38   | −7.81                  | 0.9852 | 0.8102 | 1.1                     | 1.2                     | 0.4414     |
| RL      | 0 dB   | 65.62    | 56.25   | −9.38                  | 0.9910 | 0.8595 | 0.9                     | 0.9                     | 0.3074     |
| RL      | +10 dB | 62.50    | 68.75   | +6.25                  | 0.9960 | 0.9120 | 0.5                     | 0.5                     | 0.5224     |
| RL      | +15 dB | 67.19    | 70.31   | +3.12                  | 0.9966 | 0.9181 | 0.5                     | 0.5                     | 0.8231     |
| RL      | +20 dB | 64.06    | 71.88   | +7.81                  | 0.9967 | 0.9193 | 0.5                     | 0.5                     | 0.3827     |
| J20A    | −5 dB  | 64.06    | 51.56   | −12.50                 | 0.9864 | 0.8205 | 1.1                     | 1.2                     | 0.1698     |
| J20A    | 0 dB   | 60.94    | 57.81   | −3.12                  | 0.9922 | 0.8715 | 0.8                     | 0.8                     | 0.8445     |
| J20A    | +10 dB | 60.94    | 70.31   | +9.38                  | 0.9987 | 0.9514 | 0.4                     | 0.4                     | 0.2636     |
| J20A    | +15 dB | 60.94    | 71.88   | +10.94                 | 0.9993 | 0.9648 | 0.3                     | 0.3                     | 0.2109     |
| J20A    | +20 dB | 60.94    | 71.88   | +10.94                 | 0.9996 | 0.9716 | 0.3                     | 0.3                     | 0.2109     |
| J50A    | −5 dB  | 65.62    | 51.56   | −14.06                 | 0.9865 | 0.8209 | 1.1                     | 1.2                     | 0.1096     |
| J50A    | 0 dB   | 62.50    | 59.38   | −3.12                  | 0.9923 | 0.8727 | 0.8                     | 0.8                     | 0.8383     |
| J50A    | +10 dB | 59.38    | 70.31   | +10.94                 | 0.9988 | 0.9522 | 0.3                     | 0.3                     | 0.1904     |
| J50A    | +15 dB | 60.94    | 70.31   | +9.38                  | 0.9994 | 0.9672 | 0.2                     | 0.2                     | 0.2864     |
| J50A    | +20 dB | 60.94    | 75.00   | +14.06                 | 0.9996 | 0.9737 | 0.2                     | 0.2                     | 0.0953     |
| J70A    | −5 dB  | 65.62    | 53.12   | −12.50                 | 0.9866 | 0.8215 | 1.1                     | 1.2                     | 0.1698     |
| J70A    | 0 dB   | 62.50    | 59.38   | −3.12                  | 0.9924 | 0.8734 | 0.8                     | 0.8                     | 0.8383     |
| J70A    | +10 dB | 59.38    | 70.31   | +10.94                 | 0.9988 | 0.9536 | 0.3                     | 0.3                     | 0.2109     |
| J70A    | +15 dB | 60.94    | 71.88   | +10.94                 | 0.9995 | 0.9681 | 0.3                     | 0.2                     | 0.2109     |
| J70A    | +20 dB | 60.94    | 73.44   | +12.50                 | 0.9997 | 0.9753 | 0.2                     | 0.2                     | 0.1356     |
| J20R    | −5 dB  | 65.62    | 57.81   | −7.81                  | 0.9850 | 0.8089 | 1.1                     | 1.2                     | 0.4414     |
| J20R    | 0 dB   | 65.62    | 54.69   | −10.94                 | 0.9907 | 0.8568 | 0.9                     | 0.9                     | 0.2109     |
| J20R    | +10 dB | 64.06    | 68.75   | +4.69                  | 0.9959 | 0.9104 | 0.5                     | 0.6                     | 0.6625     |
| J20R    | +15 dB | 67.19    | 70.31   | +3.12                  | 0.9964 | 0.9160 | 0.5                     | 0.6                     | 0.8231     |
| J20R    | +20 dB | 65.62    | 71.88   | +6.25                  | 0.9965 | 0.9166 | 0.5                     | 0.5                     | 0.5023     |
| J50R    | −5 dB  | 67.19    | 59.38   | −7.81                  | 0.9851 | 0.8094 | 1.1                     | 1.2                     | 0.4414     |
| J50R    | 0 dB   | 65.62    | 56.25   | −9.38                  | 0.9909 | 0.8585 | 0.9                     | 0.9                     | 0.3074     |
| J50R    | +10 dB | 62.50    | 68.75   | +6.25                  | 0.9959 | 0.9105 | 0.5                     | 0.6                     | 0.5224     |
| J50R    | +15 dB | 65.62    | 71.88   | +6.25                  | 0.9965 | 0.9169 | 0.5                     | 0.5                     | 0.4795     |
| J50R    | +20 dB | 62.50    | 71.88   | +9.38                  | 0.9966 | 0.9182 | 0.5                     | 0.5                     | 0.2864     |
| J70R    | −5 dB  | 67.19    | 59.38   | −7.81                  | 0.9851 | 0.8096 | 1.1                     | 1.2                     | 0.4414     |
| J70R    | 0 dB   | 65.62    | 56.25   | −9.38                  | 0.9910 | 0.8592 | 0.9                     | 0.9                     | 0.3074     |
| J70R    | +10 dB | 59.38    | 70.31   | +10.94                 | 0.9960 | 0.9116 | 0.5                     | 0.6                     | 0.2109     |
| J70R    | +15 dB | 67.19    | 70.31   | +3.12                  | 0.9965 | 0.9178 | 0.5                     | 0.5                     | 0.8231     |
| J70R    | +20 dB | 64.06    | 71.88   | +7.81                  | 0.9967 | 0.9193 | 0.5                     | 0.5                     | 0.3827     |

**Table A3.** Complete results for BLIP-2 on PathVQA. Baseline (clean) = 54.70%; MLP (clean) = 82.05%;  $F_1$  (clean) = 0.8313. Shaded rows:  $p < 0.0001$ .

| Channel | SNR    | Base % | MLP % | $\Delta$ Acc % | CosSim | FRS    | $\Delta E_{app}$ | $\Delta E_{det}$ | Sig.    |
|---------|--------|--------|-------|----------------|--------|--------|------------------|------------------|---------|
| AW      | −5 dB  | 54.70  | 79.31 | +24.61         | 0.9239 | 0.5424 | 3.2              | 5.4              | <0.0001 |
| AW      | 0 dB   | 54.93  | 81.69 | +26.75         | 0.9457 | 0.6291 | 3.0              | 4.7              | <0.0001 |
| AW      | +10 dB | 54.93  | 82.88 | +27.94         | 0.9820 | 0.8066 | 2.3              | 2.4              | <0.0001 |
| AW      | +15 dB | 54.82  | 82.88 | +28.06         | 0.9880 | 0.8499 | 1.7              | 1.9              | <0.0001 |
| AW      | +20 dB | 54.58  | 82.40 | +27.82         | 0.9920 | 0.8848 | 1.2              | 1.4              | <0.0001 |
| RL      | −5 dB  | 54.93  | 78.72 | +23.78         | 0.9197 | 0.5263 | 3.4              | 5.4              | <0.0001 |
| RL      | 0 dB   | 54.70  | 79.90 | +25.21         | 0.9364 | 0.5908 | 3.1              | 5.0              | <0.0001 |
| RL      | +10 dB | 55.29  | 81.81 | +26.52         | 0.9656 | 0.7158 | 2.7              | 3.7              | <0.0001 |
| RL      | +15 dB | 55.17  | 80.86 | +25.68         | 0.9694 | 0.7345 | 2.5              | 3.5              | <0.0001 |
| RL      | +20 dB | 55.17  | 81.45 | +26.28         | 0.9705 | 0.7406 | 2.5              | 3.5              | <0.0001 |
| J20A    | −5 dB  | 54.46  | 79.55 | +25.09         | 0.9236 | 0.5413 | 3.3              | 5.4              | <0.0001 |
| J20A    | 0 dB   | 54.82  | 81.69 | +26.87         | 0.9449 | 0.6252 | 3.1              | 4.7              | <0.0001 |
| J20A    | +10 dB | 54.82  | 82.52 | +27.71         | 0.9801 | 0.7931 | 2.5              | 2.5              | <0.0001 |
| J20A    | +15 dB | 54.70  | 82.40 | +27.71         | 0.9851 | 0.8255 | 2.1              | 2.1              | <0.0001 |
| J20A    | +20 dB | 54.34  | 81.81 | +27.47         | 0.9862 | 0.8334 | 2.0              | 1.9              | <0.0001 |
| J50A    | −5 dB  | 54.70  | 79.43 | +24.73         | 0.9238 | 0.5418 | 3.3              | 5.4              | <0.0001 |
| J50A    | 0 dB   | 54.93  | 81.69 | +26.75         | 0.9453 | 0.6273 | 3.0              | 4.7              | <0.0001 |
| J50A    | +10 dB | 55.05  | 82.88 | +27.82         | 0.9813 | 0.8015 | 2.4              | 2.4              | <0.0001 |
| J50A    | +15 dB | 54.70  | 82.64 | +27.94         | 0.9875 | 0.8447 | 1.8              | 1.9              | <0.0001 |
| J50A    | +20 dB | 54.46  | 82.76 | +28.30         | 0.9898 | 0.8632 | 1.5              | 1.6              | <0.0001 |
| J70A    | −5 dB  | 54.70  | 79.07 | +24.38         | 0.9238 | 0.5419 | 3.3              | 5.4              | <0.0001 |
| J70A    | 0 dB   | 54.93  | 81.81 | +26.87         | 0.9455 | 0.6278 | 3.0              | 4.7              | <0.0001 |
| J70A    | +10 dB | 55.05  | 82.64 | +27.59         | 0.9818 | 0.8048 | 2.3              | 2.3              | <0.0001 |
| J70A    | +15 dB | 54.58  | 82.40 | +27.82         | 0.9883 | 0.8507 | 1.7              | 1.8              | <0.0001 |
| J70A    | +20 dB | 54.58  | 82.64 | +28.06         | 0.9914 | 0.8771 | 1.3              | 1.5              | <0.0001 |
| J20R    | −5 dB  | 55.05  | 79.07 | +24.02         | 0.9194 | 0.5251 | 3.4              | 5.4              | <0.0001 |
| J20R    | 0 dB   | 54.70  | 79.79 | +25.09         | 0.9357 | 0.5878 | 3.1              | 5.1              | <0.0001 |
| J20R    | +10 dB | 55.29  | 81.93 | +26.63         | 0.9645 | 0.7104 | 2.7              | 3.7              | <0.0001 |
| J20R    | +15 dB | 55.29  | 81.57 | +26.28         | 0.9687 | 0.7312 | 2.6              | 3.5              | <0.0001 |
| J20R    | +20 dB | 55.29  | 80.86 | +25.56         | 0.9706 | 0.7405 | 2.5              | 3.4              | <0.0001 |
| J50R    | −5 dB  | 54.93  | 78.83 | +23.90         | 0.9196 | 0.5260 | 3.3              | 5.4              | <0.0001 |
| J50R    | 0 dB   | 54.70  | 80.14 | +25.45         | 0.9360 | 0.5894 | 3.1              | 5.0              | <0.0001 |
| J50R    | +10 dB | 54.93  | 82.05 | +27.11         | 0.9653 | 0.7143 | 2.7              | 3.7              | <0.0001 |
| J50R    | +15 dB | 55.41  | 81.57 | +26.16         | 0.9695 | 0.7352 | 2.5              | 3.5              | <0.0001 |
| J50R    | +20 dB | 55.05  | 81.69 | +26.63         | 0.9711 | 0.7437 | 2.5              | 3.4              | <0.0001 |
| J70R    | −5 dB  | 54.93  | 78.95 | +24.02         | 0.9196 | 0.5260 | 3.4              | 5.4              | <0.0001 |
| J70R    | 0 dB   | 54.70  | 80.14 | +25.45         | 0.9361 | 0.5896 | 3.1              | 5.0              | <0.0001 |
| J70R    | +10 dB | 55.05  | 82.05 | +26.99         | 0.9655 | 0.7152 | 2.7              | 3.7              | <0.0001 |
| J70R    | +15 dB | 55.53  | 81.33 | +25.80         | 0.9695 | 0.7356 | 2.5              | 3.5              | <0.0001 |
| J70R    | +20 dB | 55.05  | 81.21 | +26.16         | 0.9710 | 0.7430 | 2.5              | 3.4              | <0.0001 |

**Table A4.** Complete results for BLIP-2 on VQA-RAD. Baseline (clean) = 48.44%; MLP (clean) = 68.75%;  $F_1$  (clean) = 0.7059;  $p_{\text{clean}} = 0.021$ . Shaded rows:  $p < 0.05$  (McNemar’s test).

| Channel | SNR    | Base % | MLP % | $\Delta\text{Acc}$ % | CosSim | FRS    | $\Delta E_{\text{app}}$ | $\Delta E_{\text{det}}$ | $p$ -Value |
|---------|--------|--------|-------|----------------------|--------|--------|-------------------------|-------------------------|------------|
| AW      | −5 dB  | 48.44  | 70.31 | +21.88               | 0.9749 | 0.7714 | 2.1                     | 3.2                     | 0.0449     |
| AW      | 0 dB   | 48.44  | 68.75 | +20.31               | 0.9861 | 0.8325 | 2.0                     | 2.1                     | 0.0547     |
| AW      | +10 dB | 48.44  | 68.75 | +20.31               | 0.9913 | 0.8684 | 1.9                     | 1.4                     | 0.0311     |
| AW      | +15 dB | 48.44  | 65.62 | +17.19               | 0.9884 | 0.8497 | 2.1                     | 1.5                     | 0.0543     |
| AW      | +20 dB | 48.44  | 65.62 | +17.19               | 0.9877 | 0.8429 | 2.3                     | 1.5                     | 0.0543     |
| RL      | −5 dB  | 48.44  | 65.62 | +17.19               | 0.9702 | 0.7481 | 2.4                     | 3.4                     | 0.1273     |
| RL      | 0 dB   | 48.44  | 70.31 | +21.88               | 0.9831 | 0.8149 | 1.9                     | 2.4                     | 0.0398     |
| RL      | +10 dB | 48.44  | 65.62 | +17.19               | 0.9901 | 0.8566 | 1.7                     | 1.5                     | 0.0910     |
| RL      | +15 dB | 48.44  | 68.75 | +20.31               | 0.9878 | 0.8403 | 1.8                     | 1.6                     | 0.0311     |
| RL      | +20 dB | 48.44  | 70.31 | +21.88               | 0.9855 | 0.8247 | 2.0                     | 1.8                     | 0.0216     |
| J20A    | −5 dB  | 48.44  | 70.31 | +21.88               | 0.9747 | 0.7702 | 2.1                     | 3.2                     | 0.0449     |
| J20A    | 0 dB   | 48.44  | 68.75 | +20.31               | 0.9859 | 0.8311 | 2.0                     | 2.2                     | 0.0547     |
| J20A    | +10 dB | 48.44  | 70.31 | +21.88               | 0.9920 | 0.8730 | 2.0                     | 1.4                     | 0.0216     |
| J20A    | +15 dB | 48.44  | 70.31 | +21.88               | 0.9903 | 0.8640 | 2.0                     | 1.3                     | 0.0140     |
| J20A    | +20 dB | 48.44  | 68.75 | +20.31               | 0.9903 | 0.8650 | 1.9                     | 1.3                     | 0.0209     |
| J50A    | −5 dB  | 48.44  | 70.31 | +21.88               | 0.9750 | 0.7711 | 2.2                     | 3.2                     | 0.0449     |
| J50A    | 0 dB   | 48.44  | 68.75 | +20.31               | 0.9860 | 0.8322 | 2.0                     | 2.1                     | 0.0547     |
| J50A    | +10 dB | 48.44  | 68.75 | +20.31               | 0.9912 | 0.8676 | 2.0                     | 1.4                     | 0.0311     |
| J50A    | +15 dB | 48.44  | 67.19 | +18.75               | 0.9887 | 0.8512 | 2.2                     | 1.4                     | 0.0376     |
| J50A    | +20 dB | 48.44  | 65.62 | +17.19               | 0.9883 | 0.8473 | 2.2                     | 1.5                     | 0.0543     |
| J70A    | −5 dB  | 48.44  | 70.31 | +21.88               | 0.9748 | 0.7707 | 2.2                     | 3.2                     | 0.0449     |
| J70A    | 0 dB   | 48.44  | 68.75 | +20.31               | 0.9861 | 0.8328 | 1.9                     | 2.1                     | 0.0547     |
| J70A    | +10 dB | 48.44  | 70.31 | +21.88               | 0.9913 | 0.8684 | 1.9                     | 1.4                     | 0.0216     |
| J70A    | +15 dB | 48.44  | 65.62 | +17.19               | 0.9884 | 0.8493 | 2.1                     | 1.5                     | 0.0543     |
| J70A    | +20 dB | 48.44  | 65.62 | +17.19               | 0.9875 | 0.8416 | 2.2                     | 1.6                     | 0.0543     |
| J20R    | −5 dB  | 48.44  | 65.62 | +17.19               | 0.9700 | 0.7470 | 2.4                     | 3.4                     | 0.1273     |
| J20R    | 0 dB   | 48.44  | 70.31 | +21.88               | 0.9828 | 0.8128 | 1.9                     | 2.5                     | 0.0398     |
| J20R    | +10 dB | 48.44  | 64.06 | +15.62               | 0.9900 | 0.8554 | 1.8                     | 1.5                     | 0.1227     |
| J20R    | +15 dB | 48.44  | 65.62 | +17.19               | 0.9881 | 0.8426 | 1.9                     | 1.6                     | 0.0725     |
| J20R    | +20 dB | 48.44  | 65.62 | +17.19               | 0.9858 | 0.8267 | 1.9                     | 1.7                     | 0.0725     |
| J50R    | −5 dB  | 48.44  | 65.62 | +17.19               | 0.9701 | 0.7472 | 2.4                     | 3.4                     | 0.1273     |
| J50R    | 0 dB   | 48.44  | 70.31 | +21.88               | 0.9829 | 0.8140 | 1.9                     | 2.4                     | 0.0398     |
| J50R    | +10 dB | 48.44  | 65.62 | +17.19               | 0.9900 | 0.8560 | 1.8                     | 1.5                     | 0.0910     |
| J50R    | +15 dB | 48.44  | 68.75 | +20.31               | 0.9879 | 0.8414 | 1.8                     | 1.7                     | 0.0311     |
| J50R    | +20 dB | 48.44  | 70.31 | +21.88               | 0.9854 | 0.8244 | 2.1                     | 1.8                     | 0.0216     |
| J70R    | −5 dB  | 48.44  | 65.62 | +17.19               | 0.9701 | 0.7474 | 2.4                     | 3.4                     | 0.1273     |
| J70R    | 0 dB   | 48.44  | 70.31 | +21.88               | 0.9830 | 0.8143 | 1.9                     | 2.4                     | 0.0398     |
| J70R    | +10 dB | 48.44  | 65.62 | +17.19               | 0.9901 | 0.8561 | 1.7                     | 1.5                     | 0.0910     |
| J70R    | +15 dB | 48.44  | 68.75 | +20.31               | 0.9880 | 0.8419 | 1.9                     | 1.6                     | 0.0311     |
| J70R    | +20 dB | 48.44  | 67.19 | +18.75               | 0.9857 | 0.8263 | 2.0                     | 1.7                     | 0.0518     |

**Table A5.** Complete results for BioViL-T on PathVQA. Baseline (clean) = 54.70%; MLP (clean) = 74.08%;  $F_1$  (clean) = 0.7666. Shared rows:  $p < 0.0001$ .

| Channel | SNR    | Base % | MLP % | $\Delta$ Acc % | CosSim | FRS    | $\Delta E_{app}$ | $\Delta E_{det}$ | Sig.    |
|---------|--------|--------|-------|----------------|--------|--------|------------------|------------------|---------|
| AW      | −5 dB  | 54.70  | 71.94 | +17.24         | 0.9499 | 0.6290 | 4.3              | 4.8              | <0.0001 |
| AW      | 0 dB   | 54.70  | 75.27 | +20.57         | 0.9576 | 0.6662 | 3.7              | 4.5              | <0.0001 |
| AW      | +10 dB | 54.70  | 73.72 | +19.02         | 0.9833 | 0.8070 | 2.5              | 3.3              | <0.0001 |
| AW      | +15 dB | 54.70  | 74.32 | +19.62         | 0.9916 | 0.8703 | 1.9              | 2.4              | <0.0001 |
| AW      | +20 dB | 54.70  | 75.51 | +20.81         | 0.9961 | 0.9168 | 1.2              | 1.6              | <0.0001 |
| RL      | −5 dB  | 54.70  | 73.13 | +18.43         | 0.9496 | 0.6289 | 4.3              | 4.9              | <0.0001 |
| RL      | 0 dB   | 54.70  | 74.55 | +19.86         | 0.9561 | 0.6591 | 3.8              | 4.5              | <0.0001 |
| RL      | +10 dB | 54.70  | 76.22 | +21.52         | 0.9694 | 0.7258 | 3.2              | 3.9              | <0.0001 |
| RL      | +15 dB | 54.70  | 76.46 | +21.76         | 0.9722 | 0.7409 | 3.1              | 3.8              | <0.0001 |
| RL      | +20 dB | 54.70  | 75.27 | +20.57         | 0.9738 | 0.7497 | 3.0              | 3.7              | <0.0001 |
| J20A    | −5 dB  | 54.70  | 72.41 | +17.72         | 0.9499 | 0.6293 | 4.2              | 4.8              | <0.0001 |
| J20A    | 0 dB   | 54.70  | 75.03 | +20.33         | 0.9574 | 0.6650 | 3.7              | 4.5              | <0.0001 |
| J20A    | +10 dB | 54.70  | 74.32 | +19.62         | 0.9828 | 0.8032 | 2.6              | 3.4              | <0.0001 |
| J20A    | +15 dB | 54.70  | 74.32 | +19.62         | 0.9904 | 0.8586 | 2.0              | 2.6              | <0.0001 |
| J20A    | +20 dB | 54.70  | 76.22 | +21.52         | 0.9941 | 0.8929 | 1.6              | 2.1              | <0.0001 |
| J50A    | −5 dB  | 54.70  | 72.29 | +17.60         | 0.9498 | 0.6289 | 4.3              | 4.8              | <0.0001 |
| J50A    | 0 dB   | 54.70  | 75.62 | +20.93         | 0.9575 | 0.6654 | 3.7              | 4.5              | <0.0001 |
| J50A    | +10 dB | 54.70  | 74.67 | +19.98         | 0.9828 | 0.8038 | 2.5              | 3.3              | <0.0001 |
| J50A    | +15 dB | 54.70  | 74.55 | +19.86         | 0.9909 | 0.8640 | 2.0              | 2.5              | <0.0001 |
| J50A    | +20 dB | 54.70  | 75.62 | +20.93         | 0.9952 | 0.9059 | 1.4              | 1.8              | <0.0001 |
| J70A    | −5 dB  | 54.70  | 72.06 | +17.36         | 0.9499 | 0.6290 | 4.2              | 4.8              | <0.0001 |
| J70A    | 0 dB   | 54.70  | 74.44 | +19.74         | 0.9575 | 0.6656 | 3.7              | 4.5              | <0.0001 |
| J70A    | +10 dB | 54.70  | 73.37 | +18.67         | 0.9830 | 0.8050 | 2.5              | 3.3              | <0.0001 |
| J70A    | +15 dB | 54.70  | 74.44 | +19.74         | 0.9913 | 0.8671 | 1.9              | 2.5              | <0.0001 |
| J70A    | +20 dB | 54.70  | 75.51 | +20.81         | 0.9955 | 0.9099 | 1.3              | 1.7              | <0.0001 |
| J20R    | −5 dB  | 54.70  | 72.89 | +18.19         | 0.9496 | 0.6288 | 4.3              | 4.9              | <0.0001 |
| J20R    | 0 dB   | 54.70  | 74.08 | +19.38         | 0.9560 | 0.6586 | 3.9              | 4.6              | <0.0001 |
| J20R    | +10 dB | 54.70  | 76.34 | +21.64         | 0.9693 | 0.7249 | 3.2              | 3.9              | <0.0001 |
| J20R    | +15 dB | 54.70  | 76.22 | +21.52         | 0.9722 | 0.7406 | 3.1              | 3.9              | <0.0001 |
| J20R    | +20 dB | 54.70  | 74.91 | +20.21         | 0.9737 | 0.7487 | 3.0              | 3.8              | <0.0001 |
| J50R    | −5 dB  | 54.70  | 72.89 | +18.19         | 0.9496 | 0.6290 | 4.3              | 4.9              | <0.0001 |
| J50R    | 0 dB   | 54.70  | 73.48 | +18.79         | 0.9560 | 0.6590 | 3.8              | 4.5              | <0.0001 |
| J50R    | +10 dB | 54.70  | 76.34 | +21.64         | 0.9692 | 0.7247 | 3.1              | 3.9              | <0.0001 |
| J50R    | +15 dB | 54.70  | 76.58 | +21.88         | 0.9723 | 0.7413 | 3.1              | 3.8              | <0.0001 |
| J50R    | +20 dB | 54.70  | 74.79 | +20.10         | 0.9737 | 0.7492 | 3.0              | 3.7              | <0.0001 |
| J70R    | −5 dB  | 54.70  | 73.13 | +18.43         | 0.9497 | 0.6292 | 4.3              | 4.9              | <0.0001 |
| J70R    | 0 dB   | 54.70  | 73.60 | +18.91         | 0.9561 | 0.6594 | 3.8              | 4.5              | <0.0001 |
| J70R    | +10 dB | 54.70  | 75.51 | +20.81         | 0.9692 | 0.7249 | 3.2              | 3.9              | <0.0001 |
| J70R    | +15 dB | 54.70  | 76.46 | +21.76         | 0.9722 | 0.7408 | 3.1              | 3.8              | <0.0001 |
| J70R    | +20 dB | 54.70  | 74.67 | +19.98         | 0.9737 | 0.7489 | 3.0              | 3.7              | <0.0001 |

**Table A6.** Complete results for BioViL-T on VQA-RAD ( $n_{\text{test}} = 64$ ). Baseline (clean) = 48.44%; MLP (clean) = 76.56%;  $F_1$  (clean) = 0.7619;  $p_{\text{clean}} = 0.003$ . Shaded rows:  $p < 0.05$  (McNemar's test).

| Channel | SNR    | Base % | MLP % | $\Delta\text{Acc} \%$ | CosSim | FRS    | $\Delta E_{\text{app}}$ | $\Delta E_{\text{det}}$ | $p\text{-Value}$ |
|---------|--------|--------|-------|-----------------------|--------|--------|-------------------------|-------------------------|------------------|
| AW      | −5 dB  | 48.44  | 60.94 | +12.50                | 0.9674 | 0.7156 | 3.2                     | 4.2                     | 0.2561           |
| AW      | 0 dB   | 48.44  | 60.94 | +12.50                | 0.9776 | 0.7714 | 2.9                     | 3.5                     | 0.1698           |
| AW      | +10 dB | 48.44  | 70.31 | +21.88                | 0.9939 | 0.8909 | 1.7                     | 2.1                     | 0.0140           |
| AW      | +15 dB | 48.44  | 75.00 | +26.56                | 0.9974 | 0.9328 | 1.0                     | 1.3                     | 0.0041           |
| AW      | +20 dB | 48.44  | 73.44 | +25.00                | 0.9991 | 0.9609 | 0.6                     | 0.8                     | 0.0062           |
| RL      | −5 dB  | 48.44  | 64.06 | +15.62                | 0.9678 | 0.7186 | 3.0                     | 4.0                     | 0.1003           |
| RL      | 0 dB   | 48.44  | 56.25 | +7.81                 | 0.9756 | 0.7587 | 2.9                     | 3.4                     | 0.4237           |
| RL      | +10 dB | 48.44  | 64.06 | +15.62                | 0.9848 | 0.8154 | 2.2                     | 3.3                     | 0.0442           |
| RL      | +15 dB | 48.44  | 65.62 | +17.19                | 0.9866 | 0.8279 | 2.2                     | 3.2                     | 0.0218           |
| RL      | +20 dB | 48.44  | 67.19 | +18.75                | 0.9875 | 0.8347 | 2.2                     | 3.1                     | 0.0139           |
| J20A    | −5 dB  | 48.44  | 62.50 | +14.06                | 0.9674 | 0.7158 | 3.2                     | 4.3                     | 0.1884           |
| J20A    | 0 dB   | 48.44  | 60.94 | +12.50                | 0.9782 | 0.7741 | 2.7                     | 3.6                     | 0.1698           |
| J20A    | +10 dB | 48.44  | 67.19 | +18.75                | 0.9927 | 0.8815 | 1.7                     | 2.4                     | 0.0446           |
| J20A    | +15 dB | 48.44  | 70.31 | +21.88                | 0.9961 | 0.9166 | 1.2                     | 1.7                     | 0.0140           |
| J20A    | +20 dB | 48.44  | 76.56 | +28.12                | 0.9977 | 0.9371 | 1.0                     | 1.3                     | 0.0019           |
| J50A    | −5 dB  | 48.44  | 60.94 | +12.50                | 0.9673 | 0.7153 | 3.3                     | 4.2                     | 0.2561           |
| J50A    | 0 dB   | 48.44  | 56.25 | +7.81                 | 0.9777 | 0.7718 | 2.8                     | 3.6                     | 0.4576           |
| J50A    | +10 dB | 48.44  | 67.19 | +18.75                | 0.9936 | 0.8885 | 1.7                     | 2.2                     | 0.0376           |
| J50A    | +15 dB | 48.44  | 73.44 | +25.00                | 0.9968 | 0.9254 | 1.1                     | 1.4                     | 0.0062           |
| J50A    | +20 dB | 48.44  | 71.88 | +23.44                | 0.9985 | 0.9502 | 0.8                     | 0.9                     | 0.0148           |
| J70A    | −5 dB  | 48.44  | 60.94 | +12.50                | 0.9673 | 0.7151 | 3.2                     | 4.2                     | 0.2561           |
| J70A    | 0 dB   | 48.44  | 59.38 | +10.94                | 0.9776 | 0.7712 | 2.8                     | 3.6                     | 0.2301           |
| J70A    | +10 dB | 48.44  | 68.75 | +20.31                | 0.9939 | 0.8906 | 1.7                     | 2.1                     | 0.0209           |
| J70A    | +15 dB | 48.44  | 76.56 | +28.12                | 0.9973 | 0.9309 | 1.0                     | 1.3                     | 0.0019           |
| J70A    | +20 dB | 48.44  | 75.00 | +26.56                | 0.9990 | 0.9589 | 0.6                     | 0.8                     | 0.0030           |
| J20R    | −5 dB  | 48.44  | 60.94 | +12.50                | 0.9682 | 0.7205 | 3.0                     | 4.0                     | 0.2012           |
| J20R    | 0 dB   | 48.44  | 54.69 | +6.25                 | 0.9764 | 0.7626 | 2.8                     | 3.4                     | 0.5403           |
| J20R    | +10 dB | 48.44  | 59.38 | +10.94                | 0.9851 | 0.8165 | 1.9                     | 3.1                     | 0.2109           |
| J20R    | +15 dB | 48.44  | 57.81 | +9.38                 | 0.9866 | 0.8275 | 2.0                     | 3.1                     | 0.2864           |
| J20R    | +20 dB | 48.44  | 59.38 | +10.94                | 0.9874 | 0.8329 | 2.0                     | 3.1                     | 0.2109           |
| J50R    | −5 dB  | 48.44  | 60.94 | +12.50                | 0.9678 | 0.7186 | 3.1                     | 4.0                     | 0.2159           |
| J50R    | 0 dB   | 48.44  | 59.38 | +10.94                | 0.9757 | 0.7589 | 2.9                     | 3.4                     | 0.2109           |
| J50R    | +10 dB | 48.44  | 64.06 | +15.62                | 0.9848 | 0.8157 | 2.1                     | 3.3                     | 0.0442           |
| J50R    | +15 dB | 48.44  | 70.31 | +21.88                | 0.9868 | 0.8291 | 2.1                     | 3.1                     | 0.0056           |
| J50R    | +20 dB | 48.44  | 65.62 | +17.19                | 0.9876 | 0.8349 | 2.1                     | 3.0                     | 0.0291           |
| J70R    | −5 dB  | 48.44  | 62.50 | +14.06                | 0.9677 | 0.7178 | 3.1                     | 4.0                     | 0.1508           |
| J70R    | 0 dB   | 48.44  | 56.25 | +7.81                 | 0.9756 | 0.7582 | 2.9                     | 3.4                     | 0.4237           |
| J70R    | +10 dB | 48.44  | 62.50 | +14.06                | 0.9847 | 0.8146 | 2.3                     | 3.2                     | 0.0665           |
| J70R    | +15 dB | 48.44  | 65.62 | +17.19                | 0.9866 | 0.8285 | 2.2                     | 3.2                     | 0.0218           |
| J70R    | +20 dB | 48.44  | 70.31 | +21.88                | 0.9875 | 0.8350 | 2.2                     | 3.1                     | 0.0037           |

## References

1. Lau, J.J.; Gayen, S.; Abacha, A.B.; Demner-Fushman, D. A dataset of clinically generated visual questions and answers about radiology images. *Sci. Data* **2018**, *5*, 180251. <https://doi.org/10.1038/sdata.2018.251>.
2. He, X.; Cai, Z.; Wei, W.; Zhang, Y.; Mou, L.; Xing, E.; Xie, P. Towards Visual Question Answering on Pathology Images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*; Zong, C., Xia, F., Li, W., Navigli, R., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 708–718. <https://doi.org/10.18653/v1/2021.acl-short.90>.
3. Liu, H.; Li, C.; Li, Y.; Lee, Y.J. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 26286–26296. <https://doi.org/10.1109/CVPR52733.2024.02484>.
4. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning, ICML'23*; JMLR: Norfolk, MA, USA, 2023.
5. Bannur, S.; Hyland, S.; Liu, Q.; Perez-Garcia, F.; Ilse, M.; Castro, D.C.; Boecking, B.; Sharma, H.; Bouzid, K.; Thieme, A.; et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2023; pp. 15016–15027.
6. Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; Gao, J. LLaVA-med: Training a large language-and-vision assistant for biomedicine in one day. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS'23*; Curran Associates Inc.: Red Hook, NY, USA, 2023.
7. Hartsock, I.; Rasool, G. Vision-language models for medical report generation and visual question answering: A review. *Front. Artif. Intell.* **2024**, *7*, 1430984. <https://doi.org/10.3389/frai.2024.1430984>.
8. Lin, Z.; Zhang, D.; Tao, Q.; Shi, D.; Haffari, G.; Wu, Q.; He, M.; Ge, Z. Medical visual question answering: A survey. *Artif. Intell. Med.* **2023**, *143*, 102611. <https://doi.org/https://doi.org/10.1016/j.artmed.2023.102611>.
9. Dong, W.; Shen, S.; Han, Y.; Tan, T.; Wu, J.; Xu, H. Generative Models in Medical Visual Question Answering: A Survey. *Appl. Sci.* **2025**, *15*, 2983. <https://doi.org/10.3390/app15062983>.
10. Peng, P.; Fan, W.; Shen, Y.; Yang, X.; Zhou, D. A Global Visual Information Intervention Model for Medical Visual Question Answering. *Comput. Biol. Med.* **2025**, *192*, 110195. <https://doi.org/https://doi.org/10.1016/j.combiomed.2025.110195>.
11. Fan, L.; Gong, X.; Zheng, C.; Tan, X.; Li, J.; Ou, Y. Cycle-VQA: A Cycle-Consistent Framework for Robust Medical Visual Question Answering. *Pattern Recognit.* **2025**, *165*, 111609. <https://doi.org/https://doi.org/10.1016/j.patcog.2025.111609>.
12. Lu, Z.; Zeng, Q.; Lu, M.; Chen, G.; Xia, Y. Bridging the Semantic Gap in Medical Visual Question Answering With Prompt Learning. *IEEE Trans. Med. Imaging* **2025**, *44*, 4605–4616. <https://doi.org/10.1109/TMI.2025.3580561>.
13. Chen, Q.; He, W.; Chen, X.; Cheng, C.; Bian, M.; Xu, H. Large-small model collaboration for medical visual question answering with task aware mixture of experts and relation knowledge distillation. *Image Vis. Comput.* **2026**, *165*, 105820. <https://doi.org/https://doi.org/10.1016/j.imavis.2025.105820>.
14. Mbarek, H.; Elharrouss, O.; Mahersia, H.; Litayem, N. MedFuseT: A Transformer-Based Model for Advancing Medical Visual Question Answering. *IEEE Access* **2026**, *14*, 24700–24714. <https://doi.org/10.1109/ACCESS.2026.3661875>.
15. Zhang, J.; Li, B.; Zhou, S. Hierarchical Modeling for Medical Visual Question Answering with Cross-Attention Fusion. *Appl. Sci.* **2025**, *15*, 4712. <https://doi.org/10.3390/app15094712>.
16. Azandariani, A.K.; Gordon, M.; Kaiser, I.; DadeMatthews, O.; Mirjalili, A.; Spielmann, G.; Kim, H.K. Assessing gluteus medius volume with freehand 3DUS: Validating a practical imaging tool for complex muscle morphology. *Med. Biol. Eng. Comput.* **2026**, *64*, 963–973. <https://doi.org/10.1007/s11517-025-03503-x>.
17. Ebadi, S.; Keller, E. Optimizing Machine Learning Inference Serving Systems: A Survey. In *Proceedings of the 2026 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*; IEEE: New York, NY, USA, 2026; pp. 1–9. <https://doi.org/10.1109/ACDSA67686.2026.11468130>.
18. Alsolami, F.; Alrayzah, A.; Najam, R. MiRA: A Zero-Shot Mixture-of-Reasoning Agents Framework for Multimodal Answering of Science Questions. *Appl. Sci.* **2026**, *16*, 372. <https://doi.org/10.3390/app16010372>.
19. Tayebnaimi, S.; Kiasaleh, K. On the Performance of Multi-Wavelength Underwater Optical Channels in the Presence of Optical Turbulence. In *Proceedings of the 2025 IEEE Texas Symposium on Wireless and Microwave Circuits and Systems (WMCS)*; IEEE: New York, NY, USA, 2025; pp. 1–6. <https://doi.org/10.1109/WMCS63512.2025.11104912>.
20. Tayebnaimi, S.; Kiasaleh, K. Scintillation index analysis for multi-wavelength Gaussian beams in turbulent underwater channels. *Opt. Eng.* **2024**, *63*, 128105. <https://doi.org/10.1117/1.OE.63.12.128105>.
21. Sabir, M.; Sheikh, H.; Heath, R.; Bovik, A. A joint source-channel distortion model for JPEG compressed images. *IEEE Trans. Image Process.* **2006**, *15*, 1349–1364. <https://doi.org/10.1109/TIP.2006.871118>.
22. Chaudhary, U.; Ali, M.; Kumar, A.; Sharma, A.; Jayakody, D.N. Unleashing the Power of Wireless Communication in Healthcare by Empowering Patient Care and Connectivity: A Comprehensive Survey. *IEEE Access* **2025**, *13*, 117239–117299. <https://doi.org/10.1109/ACCESS.2025.3578344>.

23. Güllü, M.; Barışçi, N. Visual Question Answering for Intelligent Communication Systems: A Systematic Review. *IEEE Access* **2026**, *14*, 11607–11630. <https://doi.org/10.1109/ACCESS.2026.3654676>.
24. Güllü, M.; Barışçi, N. Toward Semantic Communication in 6G: A Multimodal Perspective Based on Visual Question Answering Systems. In *Proceedings of the 2025 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAET)*; IEEE: New York, NY, USA, 2025; pp. 37–42. <https://doi.org/10.1109/IICAET67254.2025.11265486>.
25. Liu, S.; Li, N.; Deng, Y.; Quek, T.Q.S. Goal-Oriented Semantic Communication for Wireless Visual Question Answering. *IEEE J. Sel. Areas Commun.* **2025**, *43*, 4247–4261. <https://doi.org/10.1109/JSAC.2025.3623184>.
26. Goyal, S.; Doddapaneni, S.; Khapra, M.M.; Ravindran, B. A Survey of Adversarial Defenses and Robustness in NLP. *ACM Comput. Surv.* **2023**, *55*, 332. <https://doi.org/10.1145/3593042>.
27. Omar, M.; Choi, S.; Nyang, D.; Mohaisen, D. Robust Natural Language Processing: Recent Advances, Challenges, and Future Directions. *IEEE Access* **2022**, *10*, 86038–86056. <https://doi.org/10.1109/ACCESS.2022.3197769>.
28. Liu, X.; Li, Y.; Li, K. Enhancing the Robustness of AI-Generated Text Detectors: A Survey. *Mathematics* **2025**, *13*, 2145. <https://doi.org/10.3390/math13132145>.
29. Song, Z.; Liu, L.; Jia, F.; Luo, Y.; Jia, C.; Zhang, G.; Yang, L.; Wang, L. Robustness-Aware 3D Object Detection in Autonomous Driving: A Review and Outlook. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 15407–15436. <https://doi.org/10.1109/TITS.2024.3439557>.
30. Liu, C.; Dong, Y.; Xiang, W.; Yang, X.; Su, H.; Zhu, J.; Chen, Y.; He, Y.; Xue, H.; Zheng, S. A Comprehensive Study on Robustness of Image Classification Models: Benchmarking and Rethinking. *Int. J. Comput. Vis.* **2025**, *133*, 567–589. <https://doi.org/10.1007/s11263-024-02196-3>.
31. Hendrycks, D.; Dietterich, T. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *Proceedings of the International Conference on Learning Representations*; OpenReview: Alameda, CA, USA, 2019.
32. Zhang, C.; Pan, F.; Kim, J.; Kweon, I.S.; Mao, C. ImageNet-D: Benchmarking Neural Network Robustness on Diffusion Synthetic Object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 21752–21762.
33. Imam, R.; Marew, R.; Yaqub, M. On the Robustness of Medical Vision-Language Models: Are They Truly Generalizable? In *Proceedings of the Medical Image Understanding and Analysis*; Ali, S., Hogg, D.C., Peckham, M., Eds.; Springer: Cham, Switzerland, 2026; pp. 233–256.
34. Hussein, N.; Shamshad, F.; Naseer, M.; Nandakumar, K. PromptSmooth: Certifying Robustness of Medical Vision-Language Models via Prompt Learning. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2024*; Springer Nature: Cham, Switzerland, 2024; Volume LNCS 15012.
35. Mohasel Arjomandi, H.; Khalooei, M.; Amirmazlaghani, M. Low-epsilon adversarial attack against a neural network online image stream classifier. *Appl. Soft Comput.* **2023**, *147*, 110760. <https://doi.org/10.1016/j.asoc.2023.110760>.
36. He, Y.; Yu, G.; Dai, H. Robustness in Wireless Distributed Learning: An Information-Theoretic Analysis. *IEEE Trans. Commun.* **2025**, *73*, 11243–11258. <https://doi.org/10.1109/TCOMM.2025.3585576>.
37. Wu, W.; Wang, Y.; Liu, Q.; Wang, G.; Zhang, J. Wavelet-Improved Score-Based Generative Model for Medical Imaging. *IEEE Trans. Med. Imaging* **2024**, *43*, 966–979. <https://doi.org/10.1109/TMI.2023.3325824>.
38. Chen, B.; Huang, X.; Liu, Y.; Zhang, Z.; Lu, G.; Zhou, Z.; Pan, J. Attention-Guided and Noise-Resistant Learning for Robust Medical Image Segmentation. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 4008013. <https://doi.org/10.1109/TIM.2024.3406804>.
39. López-Fernández, J.; Espinosa, P.R.; Romero-Jerez, J.M.; López-Martínez, F.J. A Fluctuating Line-of-Sight Fading Model With Double-Rayleigh Diffuse Scattering. *IEEE Trans. Veh. Technol.* **2022**, *71*, 1000–1003. <https://doi.org/10.1109/TVT.2021.3131060>.
40. Gvozdev, A.S.; Artemova, T.K.; Alishchuk, A.M.; Kazakova, M.A. Closed-Form Hyper-Rayleigh Mode Analysis of the Fluctuating Double-Rayleigh with Line-of-Sight Fading Channel. *Inventions* **2023**, *8*, 87. <https://doi.org/10.3390/inventions8040087>.
41. Ye, F.; Liu, B.; Li, L.; Liu, D. Rate-Distortion-Optimized Deep Preprocessing for JPEG Compression. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 8330–8343. <https://doi.org/10.1109/TCSVT.2025.3550872>.
42. Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X.V.; et al. OPT: Open Pre-trained Transformer Language Models. *arXiv* **2022**, arXiv:2205.01068. <https://doi.org/10.48550/arXiv.2205.01068>.
43. Gopal, B.; Yang, H.; Horton, M.; Chen, Y. SAFER: Sharpness Aware layer-selective Finetuning for Enhanced Robustness in vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2025; pp. 3999–4008.

**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.