Beyond Accuracy: Robustness Analysis of Encoders and Fusion in Medical Visual Question Answering

Anonymous Authors[1]

Abstract

Medical Visual Question Answering (Med-VQA) systems are increasingly used in clinical decision support, yet their robustness under distribution shift remains poorly understood. Existing evaluations focus on clean accuracy and provide limited insight into why the models fail. We introduce a diagnostic framework that decomposes robustness failures into encoder instability, cross-modal error propagation, and fusion-induced error amplification. Across SLAKE, PathVQA, and VQA-RAD, we show that encoder choice alone causes up to 12% variation in calibration and up to 25% variation in robustness drop, motivating principled encoder selection before comparing fusion methods. We then evaluate fusion architectures of varying complexity under visual, textual, cross-modal, and fusion-specific clinical perturbations. Cross-modal perturbations consistently produce the largest accuracy drops and the lowest consistency, highlighting modality misalignment as a dominant failure mode. Fusion representation drift strongly correlates with performance degradation (Spearman $\rho = 0.79$), and attention-based fusion increases modality entanglement, revealing an inherent trade-off between fusion expressivity and failure localization. Overall, our results look beyond accuracy-centric evaluation to improve the reliability of multi-modal clinical systems.

1. Introduction

Medical Visual Question Answering (Med-VQA) integrates medical images with clinical text for tasks like diagnosis (Lin et al., 2023). While recent models achieve strong benchmark accuracy, standard end-to-end evaluation masks fundamentally different failure modes: when a system fails, it is unclear if the vision encoder, language encoder, or their multimodal fusion is at fault. Each demands a radically different solution.

Treating fusion mechanisms as black boxes is especially problematic in clinical settings (Wang et al., 2024; Kahl et al., 2024). Multimodal fusion introduces unique brittle failures absent in unimodal models; noise in one modality can leak into fused representations, suppressing informative signals from the other. Without disentangling encoder quality from fusion behavior, we cannot determine if clinical errors stem from weak representations or fusion-induced interference (Hessel & Lee, 2020; Cai et al., 2025).

We address this via a controlled diagnostic evaluation framework. By freezing pretrained encoders and applying targeted clinical perturbations (visual, textual, cross-modal, and fusion-specific), we isolate encoder-level robustness from fusion-induced effects. Our contributions are:

- **Encoder Benchmarking:** Revealing that high clean accuracy and calibration across vision–language backbones often mask deep structural instability.
- **Controlled Fusion Analysis:** Identifying a "compensatory suppression" failure mode where models artificially maintain stability by overriding unstable visual evidence with textual cues.
- **Diagnostic Metrics:** Introducing representation drift, fusion-induced error, and gradient-based sensitivity to quantify multimodal interference invisible to raw accuracy.[1]

2. Related Work

Medical visual question answering (Med-VQA) relies on standard pipelines with separate vision encoders, language encoders, and fusion modules (Canepa et al., 2023). However, prevailing end-to-end evaluations on benchmarks like VQA-RAD, PathVQA, and SLAKE obscure whether failures stem from the individual encoders or their multimodal interactions.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]Code available here: https://anonymous.4open.science/r/Med_VQA_Encoder_and_Fusion_Robustness-4A1C/README.md

While various general and domain-specific encoders exist for vision (He et al., 2016; Huang et al., 2017; Dosovitskiy, 2020; Liu et al., 2021; Zhang et al., 2023; Lu et al., 2025) and language (Alsentzer et al., 2019), they are primarily evaluated on downstream accuracy. This leaves their calibration and robustness under distribution shift largely underexplored. Furthermore, fusion architectures—ranging from simple concatenation to cross-modal transformers—can amplify brittleness via modality dominance or interference under noisy inputs (Hessel & Lee, 2020; Chaudhuri et al., 2025). Motivated by recent robustness and calibration studies (Hendrycks & Dietterich, 2019; Guo et al., 2017; Kahl et al., 2024), our work advances Med-VQA evaluation by explicitly isolating encoder-level effects from fusion-induced failures under controlled perturbations.

3. Methodology

3.1. Overview

Given a medical image I and question Q, Med-VQA predicts an answer $\hat{A} = \arg\max_{a \in \mathcal{A}} p(a \mid I, Q)$. We decompose the model into a visual encoder f_v, a text encoder f_t, and a fusion module g:

$$\mathbf{v} = f_v(I), \quad \mathbf{t} = f_t(Q), \quad \mathbf{h} = g(\mathbf{v}, \mathbf{t}), \tag{1}$$

followed by a linear classifier. This decomposition allows us to isolate whether failures are caused by encoder quality or by the fusion mechanism itself. We evaluate on SLAKE, VQA-RAD, and PathVQA using the official dataset splits.

3.2. Encoder-Level Diagnostics

We first fix fusion to simple concatenation, $\mathbf{h} = [\mathbf{v}; \mathbf{t}]$, followed by LayerNorm and a linear head, and freeze both encoders during training. Only the classifier is optimized, so differences in performance, calibration, and robustness reflect the representational quality of the encoder pair rather than fusion expressiveness. We benchmark diverse vision–language pairings spanning domain-specific and general-purpose models, including BioMedCLIP + PubMedBERT, RadCLIP + SciBERT, Swin-Tiny + BioClinicalBERT, ViT-B + BioClinical ModernBERT, and DenseNet-121 + BERT-Base. Each pair is evaluated on clean data and under the perturbation suite described below.

3.3. Fusion-Level Diagnostics

To study fusion in isolation, we select the most balanced encoder backbone from the encoder study and keep it fixed across all fusion experiments. We then compare fusion mechanisms ranging from weak to strong cross-modal coupling: concatenation, gated sum, GMU, bilinear pooling, BAN, co-attention, and cross-modal transformers. All fusion models are trained under identical optimization settings so that observed differences can be attributed to architecture rather than training protocol.

3.4. Perturbations and Metrics

To evaluate structural robustness beyond standard accuracy, we perform a zero-shot evaluation by exposing the frozen encoders and fusion mechanisms to a suite of 30 controlled, inference-time perturbations. These simulate realistic clinical shifts across four categories: (1) Visual (acquisition noise, artifacts), (2) Textual (clinical synonym replacement, linguistic variability), (3) Cross-Modal (semantic framing and contextual drift), and (4) Fusion-Level (modality dropout). Each perturbation $p \in \mathcal{P}$ is applied at five severity levels $s \in \mathcal{S}$. (See Appendix B for a detailed enumeration).

Beyond standard accuracy and Expected Calibration Error (ECE) (Naeini et al., 2015), we introduce a suite of diagnostic metrics designed to expose failure localization and cross-modal interference:

Overall Robustness Score (OVR): We quantify the average performance degradation across all perturbations and severities. Higher OVR indicates greater systemic stability.

Representation Drift: To measure latent sensitivity under distributional shift, we compute the cosine distance between the clean embedding $\mathbf{h}_{\text{clean}}$ and the perturbed embedding $\mathbf{h}_p$. This can be measured at both the encoder and fusion levels:

$$\text{Drift}_p = 1 - \cos(\mathbf{h}_{\text{clean}}, \mathbf{h}_p)$$

Fusion-Induced Error Rate (FIER): This core metric isolates fusion interference by quantifying cases where multimodal interaction actively overrides correct unimodal reasoning. Lower FIER indicates improved failure containment by the fusion mechanism:

$$\text{FIER} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\Big[(\hat{A}_i^v = A_i \vee \hat{A}_i^t = A_i) \wedge \hat{A}_i^f \neq A_i\Big]$$

where $\hat{A}_i^v$, $\hat{A}_i^t$, and $\hat{A}_i^f$ denote the predictions from the visual encoder, text encoder, and fusion module, respectively.

Failure Localization & Cross-Modal Leakage: We attribute model sensitivity to specific visual (v), textual (t), or fusion (θ_f) components using capacity-normalized gradient norms. To disentangle architectural scale from sensitivity, raw gradients are scaled by the square root of the component's parameter count to derive the Normalized Sensitivity Ratio ($\tilde{G}_i$). Finally, cross-modal leakage is computed via gradient-based sensitivity shifts to measure how perturbations in one modality alter the model's reliance on the other.

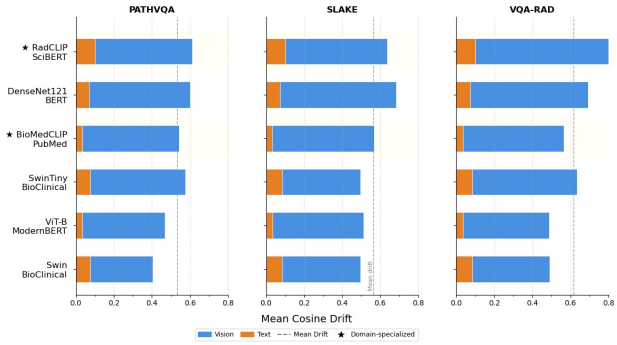


Figure 1. Vision encoders (blue) exhibit 5–10× higher representation drift than text encoders (orange) under comparable noise levels. Domain-specialized vision encoders like RadCLIP (topmost) show higher instability than general-purpose baselines.

4. Experiments and Key Results

Datasets We evaluate our framework on three public Med-VQA benchmarks: SLAKE and VQA-RAD (radiology, emphasizing diagnostic reasoning), and PathVQA (pathology, requiring fine-grained visual recognition). All experiments follow official evaluation protocols and splits.(See Appendix A).

4.1. Encoder-Level Diagnostics Results

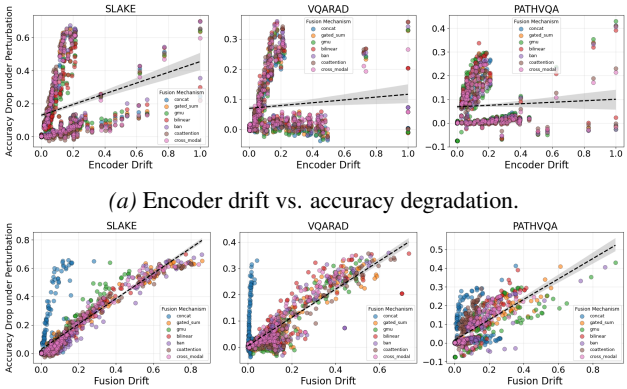
As shown in Table 1, our analysis reveals that clean-data performance is a dangerously poor proxy for reliability under distribution shift. For example, on SLAKE, Swin-Tiny + BioClinicalBERT achieves the highest accuracy (63.20%) and superior clean calibration (ECE=0.066) but exhibits lower overall robustness (OVR=0.809) than BioMedCLIP + PubMedBERT (OVR=0.839), which suffers from significantly worse calibration (ECE=0.144).

More critically, Figure 1 shows striking modality-specific instability: across all architectures, vision encoders exhibit 5–10× higher representation drift than text encoders under comparable noise levels (mean vision drift ≈ 0.50 vs. text drift ≈ 0.08). Domain-specialized encoders do not uniformly solve this; RadCLIP+SciBERT exhibits the highest vision instability across all datasets.

4.2. Fusion-Level Diagnostics & Compensatory Suppression

Fixing the encoder backbone to Swin-Tiny + BioClinicalBERT (which achieved the most balanced performance), we isolate fusion-specific behavior. While clean accuracy saturates across fusion mechanisms (varying by only ±1-2%), calibration and Fusion-Induced Error Rates (FIER) vary substantially, as shown in Table 2.

As detailed in Appendix F (Figure 4), we establish a consistent hierarchy of susceptibility across all architectures:



(a) Encoder drift vs. accuracy degradation.

(b) Fusion drift vs. accuracy degradation.

Figure 2. **Representation drift as a predictor of robustness degradation.** Each point corresponds to a fusion mechanism evaluated under a specific perturbation type and severity. **(Top)** Encoder-level drift shows weak or inconsistent correlation with accuracy drop. **(Bottom)** Fusion-layer drift exhibits strong positive correlation across SLAKE, VQA-RAD, and PathVQA.

$Error_{Cross} \gg Error_{Text} \gg Error_{Vision}$. Visual perturbations produce minimal error amplification (FIER ¡ 0.12), whereas cross-modal perturbations consistently produce the largest accuracy drops and highest FIER values. This exposes a "robustness via blindness" phenomenon—fusion mechanisms detect visual instability during training and learn to suppress visual evidence, defaulting to text-driven shortcuts. Models appear robust to visual noise only because they are ignoring the image, making them catastrophically vulnerable to linguistic and cross-modal drift.

4.3. Representation Drift & Failure Localization

Figure 2 contrasts the representation drift at the encoder and fusion levels. Encoder-level drift shows weak or inconsistent correlation with performance degradation ($\rho = 0.16 - 0.44$). In stark contrast, fusion-layer representation drift strongly correlates with performance degradation ($\rho = 0.79$), establishing the fusion layer as the primary robustness bottleneck.

Finally, capacity-normalized gradient (Figure 3) analysis ($\tilde{G}_i$) reveals a critical trade-off between fusion expressivity and error containment. Across all datasets, textual pathways dominate error sensitivity (45–60%). Structured, mathematically rigid mechanisms (e.g., Bilinear Pooling, BAN) exhibit higher fusion sensitivity but lower FIER; their rigidity acts as an architectural regularizer that localizes errors. Conversely, highly expressive attention-based mechanisms (e.g., Cross-Modal Transformers) show lower fusion sensitivity but significantly higher FIER (1.5-2x higher than Bilinear Pooling). By densely entangling modalities, attention mechanisms allow noise to propagate unconstrained, preventing failure localization.

Table 1. Encoder performance under fixed concatenation fusion across SLAKE, VQA-RAD, and PathVQA.

Encoder Pair	SLAKE			VQA-RAD			PathVQA		
	Acc ↑	ECE ↓	OVR ↑	Acc ↑	ECE ↓	OVR ↑	Acc ↑	ECE ↓	OVR ↑
BioMedCLIP + PubMedBERT	53.00	0.144	**0.839**	28.16	0.162	**0.978**	48.10	0.083	0.903
Swin-Tiny + BioClinicalBERT	**63.20**	**0.066**	0.809	**41.24**	**0.087**	0.930	**51.20**	**0.045**	**0.931**
RadCLIP + SciBERT	61.60	0.105	0.807	34.59	0.154	0.961	47.36	0.063	0.885
DenseNet121 + BERT-Base	61.90	0.069	0.750	30.82	0.206	0.947	48.43	0.066	0.888
ViT-B + BioClinical ModernBERT	53.60	0.141	0.829	31.04	0.153	0.963	50.31	0.063	0.876

Table 2. Fusion-level performance and failure diagnostics across datasets using a fixed encoder backbone (Swin-Tiny + BioClinicalBERT).

Fusion	SLAKE			VQA-RAD			PathVQA		
	Acc (%) ↑	ECE_c ↓	FIER ↓	Acc ↑	ECE_c ↓	FIER ↓	Acc ↑	ECE_c ↓	FIER ↓
Concatenation	70.0	0.104	0.186	37.9	0.228	0.175	50.2	0.039	0.096
Gated Sum	69.9	0.090	0.198	36.4	0.306	0.189	49.8	0.087	0.098
Gated Multimodal Unit (GMU)	71.4	0.086	0.162	36.6	0.309	0.198	51.8	0.096	0.120
Bilinear Pooling	71.5	0.093	0.104	36.6	0.234	**0.073**	48.4	0.092	**0.090**
Bilinear Attention Network (BAN)	72.2	0.100	0.139	37.7	0.279	0.168	47.9	0.029	0.101
Co-Attention	67.9	0.124	0.190	36.6	0.302	0.158	45.5	0.023	0.121
Cross-Modal Transformer	68.3	0.115	0.200	39.0	0.275	0.185	47.6	0.052	0.136

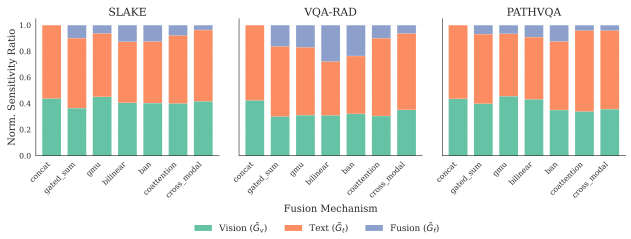


Figure 3. Failure Attribution via Capacity-Normalized Gradients. The dominance of the textual component (orange) across all benchmarks underscores the disproportionate impact of linguistic drift on model reasoning

5. Discussion

Our diagnostic framework reveals that Med-VQA fragility arises from structural interactions between brittle encoders and opportunistic fusion mechanisms. By looking beyond end-to-end accuracy, we identify three governing principles for multimodal clinical systems:

Calibration is Not Robustness: We demonstrate that models can be perfectly calibrated within their training distribution while suffering catastrophic drift under shift. Notably, domain-specialized visual encoders optimize for fine-grained medical morphology but often learn high-frequency textural features that are inherently more susceptible to noise than general-purpose representations.

Compensatory Suppression Illusion: Fusion mechanisms frequently maintain apparent stability by down-weighting visually unstable evidence and defaulting to textual cues. While this "robustness via blindness" minimizes sensitivity to visual noise, it simultaneously increases vulnerability to linguistic and cross-modal perturbations, fundamentally undermining genuine multimodal reasoning.

Expressivity vs. Error Containment: Fusion-layer drift strongly correlates with performance degradation, establishing it as the primary robustness bottleneck. However, highly entangled attention-based architectures amplify error propagation under stress, whereas mathematically constrained structured mechanisms (e.g., Bilinear Pooling) act as architectural regularizers to better localize and contain failures.

Limitations: Our current analysis relies on synthetically generated perturbations, which may not capture the full complexity of real-world clinical shifts (e.g., scanner vendor changes or protocol variations). Furthermore, our evaluation focuses on classification-based Med-VQA; modern generative systems may exhibit qualitatively different failure dynamics, particularly regarding hallucination.

Conclusion: As multimodal models are deployed in safety-critical clinical environments, evaluation must shift from surface-level accuracy to mechanistic reliability. High clean accuracy does not guarantee clinical safety. Future work must prioritize robustness-aware representation learning and principled fusion designs that prevent cross-modal error propagation. Diagnostic frameworks, such as the one introduced here, are essential steps toward developing multimodal medical AI that is not only accurate but interpretable, reliable, and trustworthy under real-world uncertainty.

Impact Statement

This work examines the reliability of Medical Visual Question Answering (Med-VQA) systems to improve transparency in how multimodal clinical models fail. In healthcare, where errors carry severe consequences, evaluating systems based on robustness, calibration, and failure attribution—rather than accuracy alone—is essential.

Our primary impact is methodological. We introduce diagnostic metrics and stress-testing procedures that disentangle encoder behavior from fusion-induced effects. This framework helps researchers understand when Med-VQA systems rely on brittle unimodal shortcuts, how fusion mechanisms propagate or suppress errors, and where failures localize under realistic distribution shifts. By highlighting the architectural trade-offs between expressiveness and failure containment, we offer practical guidance for designing safer, more reliable multimodal systems.

Ultimately, this work aims to narrow the gap between academic benchmarks and real-world clinical deployment requirements. We view robustness-aware evaluation and diagnostics as a vital complement—though not a substitute—for rigorous clinical validation and ethical deployment practices in medical AI.

References

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., and McDermott, M. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pp. 72–78, 2019.

Cai, R., Li, B., Wen, X., Chen, M., and Zhao, Z. Diagnosing and mitigating modality interference in multimodal large language models. *arXiv preprint arXiv:2505.19616*, 2025.

Canepa, L., Singh, S., and Sowmya, A. Visual question answering in the medical domain. In *2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 379–386. IEEE, 2023.

Chaudhuri, A., Dutta, A., Bui, T., and Georgescu, S. A closer look at multimodal representation collapse. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Vf9f7eNX6T.

Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Hessel, J. and Lee, L. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 861–877, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.62. URL https://aclanthology.org/2020.emnlp-main.62/.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Kahl, K.-C., Erkan, S., Traub, J., Lüth, C. T., Maier-Hein, K., Maier-Hein, L., and Jaeger, P. F. Sure-vqa: Systematic understanding of robustness evaluation in medical vqa tasks. *arXiv preprint arXiv:2411.19688*, 2024.

Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M., and Ge, Z. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143:102611, 2023.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Lu, Z., Li, H., Parikh, N. A., Dillman, J. R., and He, L. Radclip: Enhancing radiologic image analysis through contrastive language–image pretraining. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Naeini, M. P., Cooper, G., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Wang, J., Jiang, H., Liu, Y., Ma, C., Zhang, X., Pan, Y., Liu, M., Gu, P., Xia, S., Li, W., Zhang, Y., Wu, Z., Liu, Z., Zhong, T., Ge, B., Zhang, T., Qiang, N., Hu, X., Jiang, X., Zhang, X., Zhang, W., Shen, D., Liu, T., and

Zhang, S. A comprehensive review of multimodal large language models: Performance and challenges across different tasks, 2024. URL https://arxiv.org/abs/2408.01319.

Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

A. Dataset and Split Details

We evaluate our diagnostic framework on three publicly available Med-VQA benchmarks spanning radiology and pathology. All experiments strictly follow the official dataset splits and evaluation protocols provided by the dataset authors.

Table 3. Summary of Med-VQA datasets used in our experiments. All evaluations follow official dataset splits.

Dataset	Domain	# Images	# QA Pairs	Splits
SLAKE	Radiology	642	∼7K	Official train/test
VQA-RAD	Radiology	314	∼2K	Official (80/20 by question type)
PathVQA	Pathology	4.9K	∼32K	Official train/val/test

Evaluation Protocol. For SLAKE, we restrict evaluation to English-language questions and use the official train–test split. For VQA-RAD, we follow the dataset's recommended split strategy based on question type, evaluating on the held-out test set. For PathVQA, we use the predefined training, validation, and test splits released with the dataset.

No additional filtering, relabeling, or resampling is performed beyond these official protocols, ensuring comparability with prior Med-VQA studies.

B. Perturbation Framework Details

This appendix provides a complete specification of the perturbation framework used to evaluate robustness and failure behavior in Med-VQA systems. Perturbations are designed to reflect realistic sources of variation encountered in clinical settings, spanning unimodal corruptions, cross-modal contextual shifts, and fusion-level failures. All perturbations are applied at five severity levels, with higher levels introducing progressively stronger distortions.

Table 4. Perturbation framework used for robustness evaluation. Perturbations are grouped by modality and interaction type, and applied at increasing severity levels.

Perturbation Class	Category	Perturbation Types
Visual	Noise	Gaussian noise, speckle noise, salt-and-pepper noise
	Blur	Defocus blur, motion blur, zoom blur
	Camera	Static rotations, random rotations, pixel-level translations
	Digital	JPEG compression artifacts
Textual	Natural Language	Character-level typos, token swaps, word dropouts, filler word insertions, paraphrasing, synonym replacement, mixed perturbations
	Medical Domain	Medical synonym replacement, abbreviation expansion or contraction, terminology typos, negation insertion, noun dropouts, distracting clinical insertions, mixed perturbations
Cross-Modal Contextual Drift	Temporal	References to prior studies or interval changes
	Procedural	Post-procedural or post-intervention context
	Follow-up	Follow-up or recommendation language
	Clinical	Injection of broader clinical context
	Mixed	Phrases sampled across all contextual drift categories
Fusion-Level	Image Dropout	Partial or complete suppression of visual feature representations
	Text Dropout	Partial or complete suppression of textual feature representations

Table 4 summarizes all perturbation types used in our evaluation along with their categorization. Severity levels control the magnitude or frequency of corruption (e.g., noise intensity, number of text edits, or extent of feature suppression), following established robustness benchmarking practices.

C. Fusion Mechanism Formulations

This appendix provides formal definitions of the fusion mechanisms evaluated in our fusion-level diagnostic experiments. The formulations are included for completeness and reproducibility. All fusion modules operate on frozen visual and textual embeddings produced by pretrained encoders, and differ only in how cross-modal interactions are modeled. Unless otherwise stated, fusion outputs are followed by Layer Normalization and a linear classification head.

Let $\mathbf{v} \in \mathbb{R}^{d_v}$ and $\mathbf{t} \in \mathbb{R}^{d_t}$ denote the visual and textual embeddings, respectively.

Concatenation. Concatenation performs late fusion by directly joining unimodal representations:

$$\mathbf{h} = [\mathbf{v}; \mathbf{t}], \tag{2}$$

where $[\cdot;\cdot]$ denotes vector concatenation. This serves as a minimally expressive baseline with no explicit cross-modal interaction.

Gated Sum. Gated Sum projects both modalities into a shared space and computes a learnable gate to modulate their relative contributions:

$$\tilde{\mathbf{v}} = W_v \mathbf{v}, \quad \tilde{\mathbf{t}} = W_t \mathbf{t}, \tag{3}$$

$$\mathbf{g} = \sigma\big(W_g[\tilde{\mathbf{v}}; \tilde{\mathbf{t}}]\big)\,, \tag{4}$$

$$\mathbf{h} = \mathbf{g} \odot \tilde{\mathbf{v}} + (1 - \mathbf{g}) \odot \tilde{\mathbf{t}}, \tag{5}$$

where $\sigma(\cdot)$ denotes the sigmoid function and $\odot$ denotes element-wise multiplication.

Gated Multimodal Unit (GMU). GMU extends gated fusion by applying nonlinear transformations to each modality prior to gating:

$$\hat{\mathbf{v}} = \tanh(W_v \mathbf{v}), \quad \hat{\mathbf{t}} = \tanh(W_t \mathbf{t}), \tag{6}$$

$$\mathbf{g} = \sigma(W_g[\mathbf{v}; \mathbf{t}])\,, \tag{7}$$

$$\mathbf{h} = \mathbf{g} \odot \hat{\mathbf{v}} + (1 - \mathbf{g}) \odot \hat{\mathbf{t}}. \tag{8}$$

This design allows nonlinear modality-specific transformations while retaining explicit control over modality dominance.

Bilinear Pooling. Bilinear pooling captures multiplicative interactions between modalities:

$$\mathbf{h} = (W_v \mathbf{v}) \odot (W_t \mathbf{t}), \tag{9}$$

where $W_v \in \mathbb{R}^{d_h \times d_v}$ and $W_t \in \mathbb{R}^{d_h \times d_t}$ project inputs to a shared dimension d_h, and $\odot$ denotes element-wise multiplication. This is a computationally efficient approximation to full bilinear pooling $\mathbf{v}^\top W_b \mathbf{t}$.

Bilinear Attention Network (BAN). BAN extends bilinear pooling by introducing attention over joint visual–textual interactions. Let $\mathbf{V} = \{\mathbf{v}_i\}$ denote visual region features and $\mathbf{T} = \{\mathbf{t}_j\}$ denote token-level textual features. BAN computes attention weights over bilinear interactions:

$$\alpha_{ij} \propto \mathbf{v}_i^\top W \mathbf{t}_j, \tag{10}$$

and aggregates attended joint representations to form the fused embedding $\mathbf{h}$. This enables selective emphasis of cross-modal feature pairs.

Co-Attention. Co-attention mechanisms compute bidirectional attention between modalities. Visual features attend to textual features and vice versa:

$$\mathbf{A}_{v \leftarrow t} = \mathrm{Attn}(\mathbf{V}, \mathbf{T}), \tag{11}$$

$$\mathbf{A}_{t \leftarrow v} = \mathrm{Attn}(\mathbf{T}, \mathbf{V}), \tag{12}$$

where $\mathrm{Attn}(\cdot)$ denotes a standard attention operator. The attended representations are combined to produce the fused embedding $\mathbf{h}$.

Cross-Modal Transformer. Cross-modal transformers model fusion using self-attention over joint modality tokens. Visual and textual embeddings are projected into a shared space and concatenated:

$$\mathbf{X} = [\mathbf{V}; \mathbf{T}], \tag{13}$$

which is processed by a transformer encoder with self-attention:

$$\mathbf{H} = \text{Transformer}(\mathbf{X}). \tag{14}$$

The fused representation $\mathbf{h}$ is obtained via pooling over $\mathbf{H}$. This architecture represents the most expressive fusion class considered, enabling deep cross-modal interaction.

Scope Clarification. All fusion mechanisms are evaluated under identical training protocols and operate on frozen encoder representations. Parameter counts, architectural depth, and attention head configurations follow standard instantiations from prior work and are not explicitly matched. Our analysis focuses on the inductive biases introduced by fusion design rather than capacity-controlled comparisons.

Trainable Parameters by Dataset Table 5 reports the number of trainable parameters for each fusion architecture across the three Med-VQA datasets evaluated.

Table 5. Trainable parameters for each fusion architecture by dataset. All architectures use frozen encoders (Swin-Tiny for vision, BioClinicalBERT for text).

Fusion Type	SLAKE	VQA-RAD	PathVQA
Concatenation	327,379	797,701	7,502,095
Gated Sum	2,525,395	2,760,709	6,115,087
GMU	2,525,395	2,760,709	6,115,087
Low-Rank Bilinear	1,344,979	1,580,293	4,934,671
BAN	1,345,748	1,581,062	4,935,440
Co-Attention	3,709,651	3,944,965	7,299,343
Cross-Modal Tfm	12,372,947	16,348,933	19,703,311

The variation in parameter counts across datasets is due to differences in answer vocabulary size (SLAKE: 207 classes, VQA-RAD: 458 classes, PathVQA: 4,998 classes), which determines the final classifier layer dimensions. Cross-Modal Transformer has significantly more parameters due to its multi-layer transformer encoder architecture.

D. Detailed Experimental Results

We report bootstrap-based 95% confidence intervals (CI) for both the fusion mechanism comparison and the encoder architecture ablation studies. Confidence intervals are computed using 10,000 bootstrap resamples and are reported explicitly to avoid redundancy caused by near-identical bootstrap standard deviations across models.

Table 6. **Fusion Mechanism Performance.** Accuracy with 95% confidence intervals across three Med-VQA benchmarks. Interval non-overlap indicates statistically meaningful differences, particularly on PathVQA, while wider intervals on VQA-RAD reflect increased uncertainty due to limited data.

Fusion Method	PathVQA	SLAKE	VQA-RAD
Concatenation	0.502 [0.460, 0.484]	0.700 [0.672, 0.727]	0.379 [0.356, 0.402]
Gated Sum	0.498 [0.458, 0.480]	0.699 [0.671, 0.726]	0.364 [0.341, 0.387]
GMU	**0.518** [**0.510**, **0.534**]	0.714 [0.686, 0.741]	0.366 [0.343, 0.389]
Bilinear Pooling	0.484 [0.463, 0.487]	0.715 [0.687, 0.742]	0.366 [0.343, 0.389]
Bilinear Attention	0.479 [0.458, 0.482]	**0.723** [**0.695**, **0.750**]	0.377 [0.354, 0.400]
Co-Attention	0.455 [0.435, 0.458]	0.679 [0.651, 0.706]	0.366 [0.343, 0.389]
Cross-Modal Transformer	0.476 [0.455, 0.479]	0.683 [0.655, 0.710]	**0.391** [**0.368**, **0.414**]

E. Calibration Under Distribution Shift

In addition to reporting expected calibration error (ECE) under distribution shift, we evaluate the effect of post-hoc calibration via temperature scaling. A single scalar temperature T is learned on clean validation data by minimizing the negative

Table 7. **Encoder Architecture Performance.** Accuracy with 95% confidence intervals comparing frozen vision–language encoder pairings. Swin-based visual encoders demonstrate superior robustness on radiology-focused datasets (SLAKE, VQA-RAD).

Encoder Configuration *(Vision + Text)*	**PathVQA** ($N = 6719$)	**SLAKE** ($N = 1061$)	**VQA-RAD** ($N = 2248$)
BioMedCLIP + PubMedBERT	0.481 [0.460, 0.484]	0.530 [0.491, 0.551]	0.282 [0.247, 0.283]
Swin-Tiny + BioClinicalBERT	0.474 [0.455, 0.479]	**0.632** [**0.593**, **0.651**]	**0.412** [**0.379**, **0.420**]
RadCLIP + SciBERT	**0.512** [**0.510**, **0.534**]	0.616 [0.572, 0.632]	0.346 [0.318, 0.357]
DenseNet121 + BERT-Base	0.484 [0.463, 0.487]	0.619 [0.576, 0.635]	0.308 [0.276, 0.314]
ViT + BioClinicalModernBERT	0.503 [0.500, 0.524]	0.536 [0.498, 0.558]	0.310 [0.279, 0.316]

log-likelihood, following standard practice. This calibrated temperature is then held fixed and applied to all perturbed test sets across visual, textual, cross-modal, and fusion-level perturbations.

Temperature scaling improves calibration on clean data, reducing ECE to 0.0477, indicating that the uncalibrated models are overconfident in-distribution. However, under distribution shift, calibration degrades substantially even after post-hoc calibration. In particular, severe textual and cross-modal perturbations induce large increases in ECE (e.g., >0.5 for high-severity mixed medical text perturbations), while visual perturbations result in comparatively smaller calibration drift.

Importantly, post-hoc calibration does not alter model predictions or accuracy, and thus does not change the relative robustness rankings across fusion mechanisms. Instead, these results indicate that miscalibration under distribution shift is a structural property of the learned representations and fusion mechanisms, rather than a simple consequence of global overconfidence. Overall, while temperature scaling improves in-distribution reliability, it is insufficient to guarantee calibrated confidence under strong semantic or cross-modal distribution shifts.

F. Fusion-Induced Error Rates Under Perturbation

This section provides visual evidence of the fusion-induced error rate (FIER) hierarchy discussed in the main text. As shown in Figure 4, cross-modal perturbations consistently induce the highest error rates across all datasets, followed by textual perturbations. In contrast, visual perturbations produce minimal error amplification.

This consistent pattern across all tested fusion architectures highlights the "robustness via blindness" phenomenon. Because visual encoders exhibit high instability during training, the fusion mechanisms learn to suppress visual representations and over-rely on textual cues. Consequently, the models appear robust to visual noise but become highly vulnerable to cross-modal and linguistic shifts.

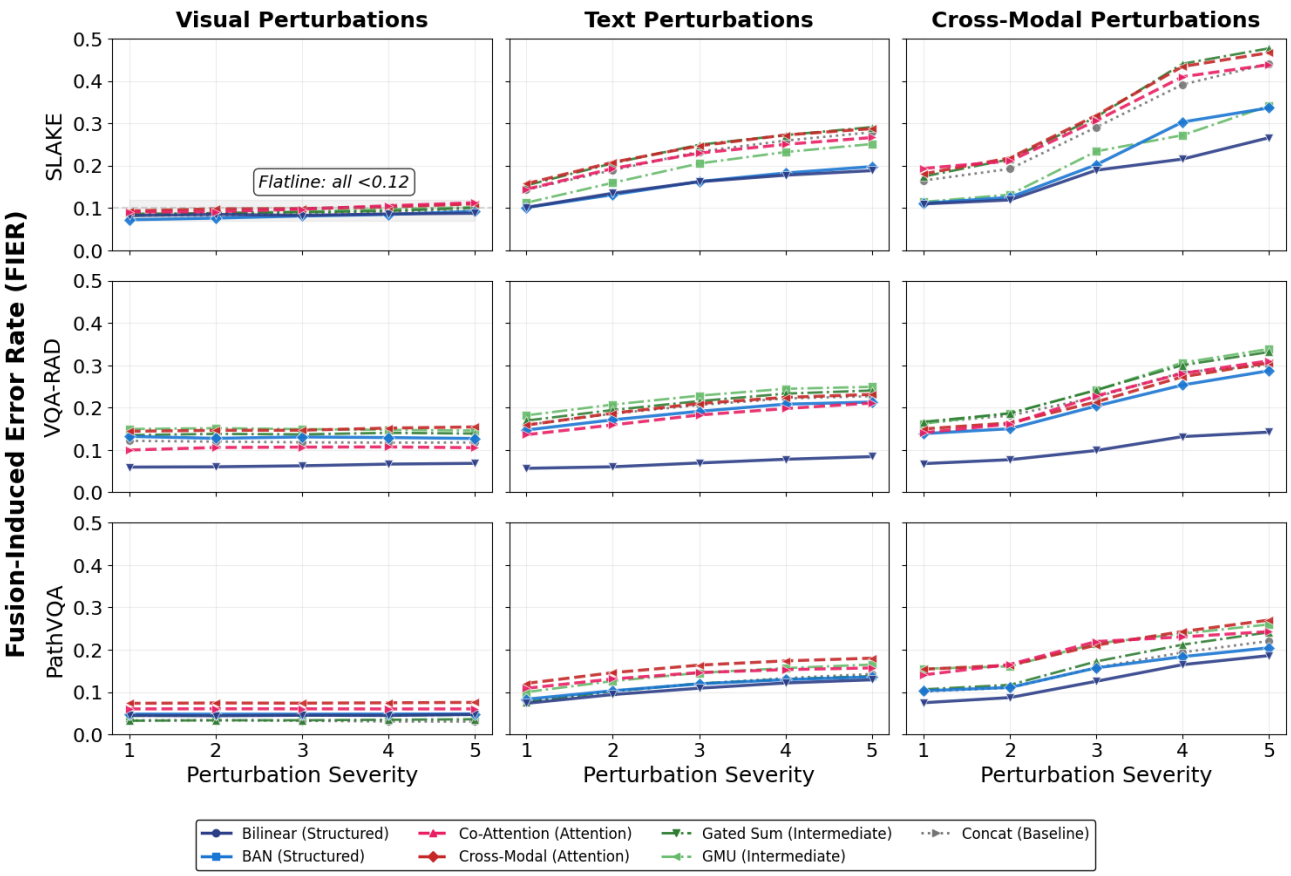


Figure 4. Fusion-induced error rate (FIER) as a function of perturbation severity across datasets and perturbation categories. Cross-modal perturbations induce the highest error rates, followed by textual and visual perturbations, consistently across fusion architectures.