MedFM-Robust: Benchmarking Robustness of Medical Foundation Models

Xiangxiang Cui[1*], Tianjin Huang[2*], Yifang Wang[3], Lijie Hu[4], and Lu Yin[5†]

[1] Beijing Normal University, China
[2] University of Exeter, United Kingdom
[3] University College London, United Kingdom
[4] Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates
[5] University of Surrey, United Kingdom
l.yin@surrey.ac.uk

Abstract. Medical foundation models have achieved remarkable clinical performance, yet their robustness under real-world perturbations remains underexplored. We present a robustness benchmark comprising 40 perturbation types (12 base, 28 medical-specific) across eight imaging modalities, evaluating five VLMs (LLaVA-Med, MedGemma, MedGemma-1.5, Gemini-2.5-flash and GPT-4o-mini) on VQA, visual grounding, and captioning, alongside two segmentation models (MedSAM, SAM-Med2D) with five fine-tuning strategies. Our findings reveal: (1) Fine-tuning strategy dominates robustness, with LoRA exhibiting nearly double the degradation of full fine-tuning, while SAM-Med2D's Adapter offers favorable efficiency-robustness trade-off. (2) Medical-specific perturbations disproportionately damage segmentation, with 9 of 15 top corruptions being domain-specific. (3) LoRA-tuned visual grounding drops over 40 points, whereas zero-shot captioning remains stable ($<$7% drop). Zero-shot VQA shows model-dependent robustness—medical models drop under 20% while Gemini-2.5-flash drops 54%. General-purpose VLMs achieve higher VQA accuracy but fail on grounding; among medical VLMs, MedGemma demonstrates the best overall stability. These results provide deployment guidelines and underscore the necessity of domain-specific robustness evaluation for medical AI. Our code is available at: https://abnerai.github.io/MedFM-Robust.

Keywords: Medical foundation models · Robustness evaluation · Vision-language models · Medical image segmentation

1 Introduction

Medical foundation models (MedFMs) have emerged as transformative tools in healthcare, demonstrating capabilities across diverse clinical applications [12,21]. These models can be broadly categorized into two paradigms: Medical Vision-Language Models (Med-VLMs) and segmentation foundation models. Med-VLMs

[*] These authors contributed equally to this work.
[†] Corresponding author.

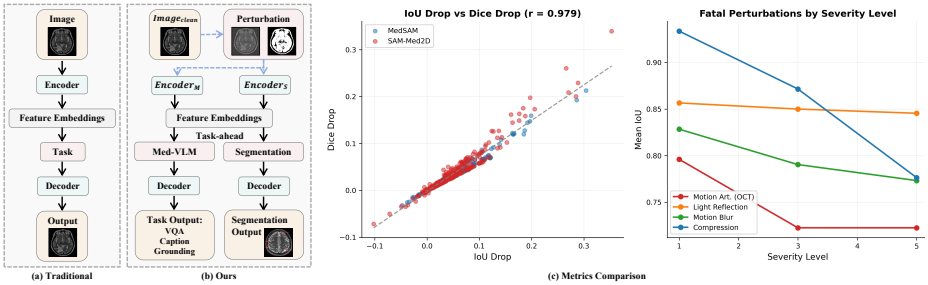

Fig. 1. Overview of our robustness evaluation and representative robustness drops. (a) Traditional clean-image evaluation pipeline. **(b)** Our robustness benchmark applies modality-adaptive perturbations before the encoder and evaluates both Med-VLM tasks and segmentation under matched settings. **(c)** Metrics comparison: IoU drop closely tracks Dice drop, and representative fatal perturbations cause increasing performance degradation with higher severity levels.

range from medical-specialized models such as LLaVA-Med [10] and MedGemma [17], to general-purpose models like GPT-4o [7] and Gemini [19], all capable of medical image understanding tasks including visual question answering (VQA), report generation, and visual grounding. Concurrently, the Segment Anything Model (SAM) [9] has catalyzed a new generation of medical segmentation models, with adaptations like SAM-Med2D [1] and MedSAM [11]. The widespread clinical deployment of these models thus necessitates rigorous evaluation of their reliability under real-world conditions.

Despite impressive benchmark performance, medical foundation models face significant robustness challenges when deployed in clinical practice. Real-world medical images are inherently susceptible to various artifacts and perturbations arising from acquisition conditions, patient factors, and equipment variations [18]. These include motion blur from patient movement, noise from low-dose protocols, and modality-specific degradations such as metal artifacts in CT, bias field inhomogeneity in MRI, speckle noise in ultrasound, and staining variations in pathology slides [20] or or site-specific style shifts [24]. While extensive research has characterized model robustness in natural image domains through corruption benchmarks like ImageNet-C [3], systematic evaluation of medical foundation model robustness remains critically understudied. Existing medical imaging benchmarks predominantly evaluate models on clean, curated datasets, creating a substantial gap between reported performance and real-world reliability.

Addressing this gap presents three fundamental challenges. First, medical imaging encompasses diverse modalities, each exhibiting distinct artifact patterns and degradation mechanisms. Generic perturbation models fail to capture modality-specific characteristics such as CT beam hardening, MRI ghosting artifacts, or histopathology stain variations [16]. Second, the field lacks a unified evaluation framework that comprehensively assesses robustness across both

vision-language understanding tasks (VQA, captioning, grounding) and dense prediction tasks (segmentation), despite these capabilities often being deployed together in clinical systems. Third, the robustness implications of different fine-tuning strategies for medical foundation models remain largely unexplored, despite their widespread adoption in clinical adaptation scenarios.

In this work, we present a comprehensive framework for evaluating and enhancing the robustness of medical foundation models under domain-specific perturbations. Our contributions are threefold:

- We introduce a modality-adaptive perturbation pipeline spanning eight medical imaging modalities, with both base and modality-specific artifacts calibrated into five SSIM-guided severity levels for consistent degradation.
- We establish a unified robustness benchmark for Med-VLMs and SAM-based segmentation models, evaluating VQA, captioning, and grounding across five VLMs (three medical-specialized, two general-purpose), and segmentation across five diverse clinical datasets covering dermoscopy, endoscopy, MRI, ultrasound, pathology, OCT, and CT.
- We find that fine-tuning strategy critically determines robustness: LoRA exhibits the highest degradation across both VLMs and segmentation models, while full fine-tuning and adapter-based methods offer better robustness-efficiency trade-offs. General-purpose VLMs achieve strong zero-shot VQA but fail on grounding tasks requiring fine-tuning.

2 Methods

We present a comprehensive framework for evaluating the robustness of medical foundation models under realistic perturbations, integrating modality-adaptive perturbation generation with unified evaluation protocols for vision-language tasks and rigorous robustness assessment for segmentation models.

2.1 Modality-Adaptive Perturbation Generation

Base Perturbations. We implement 12 base perturbation types applicable across all imaging modalities, categorized into three groups: *noise* (Gaussian, salt-and-pepper, speckle), *degradation* (Gaussian blur, motion blur, brightness, contrast, JPEG compression, pixelation), and *geometric* (rotation, scaling, translation), enabling consistent cross-modality robustness evaluation.

Modality-Specific Perturbations. zhBeyond base perturbations, we design modality-specific perturbations that simulate clinical artifacts across eight imaging modalities. For CT, we model metal-induced streaks and beam-hardening cupping. For MRI, we simulate bias-field inhomogeneity and ghosting. For ultrasound, we generate acoustic shadowing and reverberation. For pathology, we apply stain variations in HSV space. For endoscopy, we add specular reflections and bubbles. For OCT, we simulate shadow, blink, and defocus. For X-ray, we model scatter, exposure variation, and grid patterns. In task-specific experiments, we apply only the perturbations relevant to the modality of each dataset.

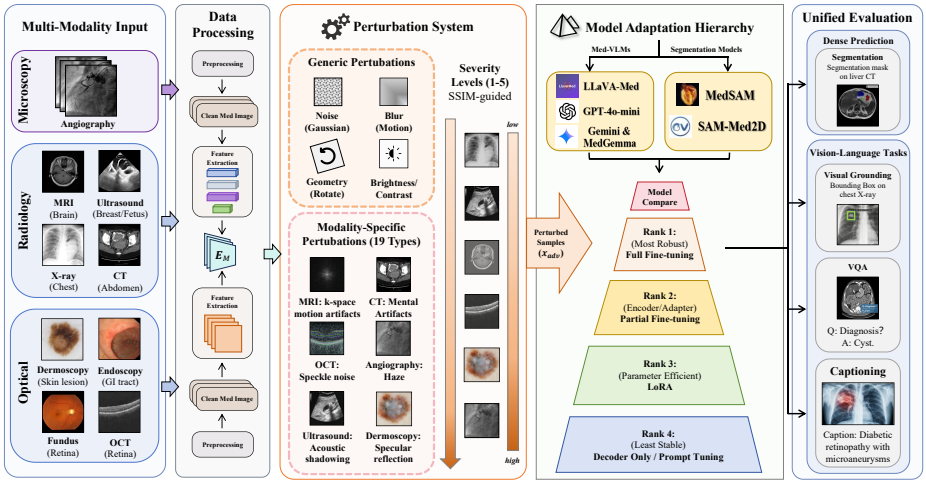

Fig. 2. Overview of our robustness evaluation framework. We generate SSIM-calibrated perturbations across five severity levels, combining base corruptions with modality-specific artifacts. We benchmark three Med-VLMs and two SAM-based segmentation models under a unified protocol, and investigate multiple fine-tuning strategies across VQA, captioning, visual grounding, and segmentation tasks.

SSIM-Guided Severity Calibration. We employ SSIM [23] as a modality-agnostic engineering anchor (not a clinical severity metric) to ensure consistent degradation across five severity levels: Level 1 (SSIM 0.90–0.98), Level 2 (0.80–0.89), Level 3 (0.70–0.79), Level 4 (0.60–0.69), and Level 5 (0.50–0.59). For each perturbation type, we use binary search to find parameters achieving the target SSIM range, caching parameters for efficiency. However, not every perturbation reaches the most severe level, so the maximum number of iterations will be set to avoid excessive search time.

2.2 Vision-Language Model Evaluation

Models and Datasets. We evaluate five foundation models: LLaVA-Med [10], MedGemma [17], MedGemma-1.5, GPT-4o-mini [13] and Gemini-2.5-flash [19]. Evaluation spans three tasks: VQA on OmniMedVQA [5] (accuracy), captioning on ROCOv2 [15] (BLEU, ROUGE-L, CIDEr), and visual grounding on MeCoVQA [6] (IoU@0.5). Among the three tasks, 500 samples were extracted from the dataset for evaluation. For **visual grounding**, we employ LoRA [4] with rank $r = 16$ and $\alpha = 32$ on the vision encoder attention layers. Let $\mathcal{T}_{\text{response}}$ denote the bounding-box coordinate tokens. Training uses response-focused loss that only backpropagates through grounding output tokens:

$$\mathcal{L}_{\text{GND}} = -\frac{1}{|\mathcal{T}_{\text{response}}|} \sum_{t \in \mathcal{T}_{\text{response}}} \log p_\theta(y_t \mid y_{<t}, I, Q), \tag{1}$$

where p_θ is the model distribution parameterised by LoRA weights θ, $y_{<t}$ denotes all preceding tokens, I is the input image, and Q is the query prompt. Gradients are backpropagated only through $\mathcal{T}_{\text{response}}$, preventing instruction tokens from dominating the grounding supervision signal.

2.3 Segmentation Model Robustness Evaluation

Models, Datasets, and Fine-tuning Strategies. We evaluate two SAM-based segmentation foundation models, SAM-Med2D [1] and MedSAM [11], on five datasets spanning diverse clinical scenarios: ISIC 2016 [2] (900 samples, dermoscopy), Kvasir-SEG [8] (1000 samples, endoscopy), Brain Tumor (3064 samples, MRI), and Glaucoma (5977 samples, Disc & Cup). To study robustness enhancement under realistic perturbations, we further investigate five fine-tuning strategies: (1) decoder-only tuning, (2) encoder-only tuning, (3) full encoder tuning, (4) low-rank adaptation (LoRA), and (5) SAM-Med2D adapter tuning.

2.4 Evaluation Protocol

Task-Specific Metrics. For segmentation, we use IoU and Dice coefficient:

$$\text{IoU}_{\text{seg}}(P_m, G_m) = \frac{|P_m \cap G_m|}{|P_m \cup G_m|}, \qquad \text{Dice}_{\text{seg}}(P_m, G_m) = \frac{2|P_m \cap G_m|}{|P_m| + |G_m|}, \tag{2}$$

where $P_m \subseteq \Omega$ and $G_m \subseteq \Omega$ denote the predicted and ground-truth masks for modality m over pixel domain Ω. For VQA, we measure accuracy as $\text{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{a}_i = a_i^*]$. For visual grounding, we report Acc@IoU≥0.5: a prediction is correct iff the box-level overlap $|\hat{b}_i \cap b_i^*|/|\hat{b}_i \cup b_i^*| \geq 0.5$. For captioning, BLEU [14] measures clipped n-gram precision with a brevity penalty, and CIDEr [22] re-weights n-grams by TF-IDF to reward clinically informative phrasing.

Robustness Metric. To quantify performance degradation under perturbations, we report the absolute performance drop. Let $\mathcal{P}_\tau$ denote the set of perturbation types in category $\tau \in \{\text{base, med-specific}\}$, $s \in \{1, \ldots, 5\}$ the SSIM-calibrated severity level, and M the task-specific metric (IoU_{seg} for segmentation, Acc_{VQA} and Acc_{GND} for VQA and visual grounding, BLEU for captioning). Lower values indicate better robustness. For aggregated analysis, we compute the mean drop across all perturbation types within each severity level or perturbation category:

$$\Delta_\tau^{(s)} = \frac{1}{|\mathcal{P}_\tau|} \sum_{p \in \mathcal{P}_\tau} \left(M_{\text{clean}} - M_{\text{perturb}}^{(p,s)} \right). \tag{3}$$

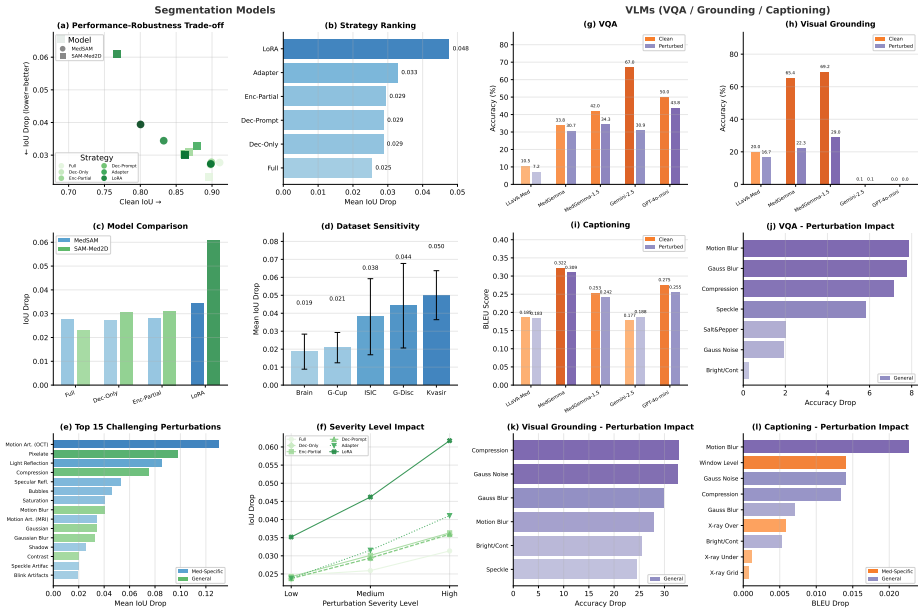

Fig. 3. Comprehensive robustness evaluation of medical image segmentation models and VLMs under perturbations. **Left (Segmentation)**: (a) Performance-robustness trade-off. (b) Strategy ranking. (c) Model comparison. (d) Dataset sensitivity. (e) Top 15 perturbation types. (f) Severity level impact. **Right (VLMs)**: (g-i) Clean vs. perturbed performance on VQA, Grounding, and Captioning. (j-l) Perturbation impact.

3 Experiment

3.1 Experiment Setup

All experiments use PyTorch on NVIDIA A100 GPUs. Segmentation fine-tuning uses AdamW with learning rate 10^{-4}, weight decay 0.01, batch size 32, and 50 epochs with cosine scheduling. We evaluate on clean and perturbed images across five severity levels and report robustness using the absolute performance drop defined earlier.

3.2 Qualitative Results in Segmentation

We evaluate five fine-tuning strategies on two SAM-based medical foundation models (MedSAM and SAM-Med2D) across five segmentation datasets, examining both clean performance and robustness under 19 perturbation types, *e.g.*, Gaussian noise, motion blur, and modality-specific artifacts.

Performance-Robustness Trade-off. Fig. 3(a) reveals a clear trade-off between clean segmentation accuracy and robustness to perturbations. Full fine-tuning achieves the highest clean IoU (0.89–0.91) while maintaining competitive robustness, positioning it in the desirable bottom-right region. In contrast,

Table 1. Per-dataset robustness summary for **MedSAM** (top block) and **SAM-Med2D** (bottom block) across five benchmarks. *Cl.*: clean IoU. ΔB↓: mean IoU drop under base corruptions. ΔM↓: medical-specific corruptions (lower = more robust).

Group	Strategy	Avg. ΔB↓	ISIC 2016			Brain Tumor			Glaucoma Disc			Glaucoma Cup			Kvasir-SEG		
			Cl.	ΔB↓	ΔM↓	Cl.	ΔB↓	ΔM↓	Cl.	ΔB↓	ΔM↓	Cl.	ΔB↓	ΔM↓	Cl.	ΔB↓	ΔM↓
MedSAM																	
Full	**Full**	↓.021	**.953**	↓.020	.034	**.878**	.022	↓.011	**.952**	↓.011	↓.023	**.907**	↓.007	.030	**.862**	.046	↓.031
Full	Enc-Only	.022	.950	.020	.028	.864	↓.013	↓.007	.947	.016	.035	.892	↓.007	.033	.843	.041	.036
PEFT	Dec-Prompt	.022	.950	.020	.028	.864	.014	.008	.946	.016	.035	.891	↓.007	.031	.842	↓.039	.035
PEFT	*LoRA*	.025	.926	↓.015	↓.027	.762	.014	.020	.919	↑.048	↑.084	.761	↑.021	.024	.795	↓.030	.028
Dec	Dec-Only	.022	.949	↓.020	↓.027	.863	↓.013	↓.007	.947	.016	.035	.890	.006	↓.030	.841	↓.039	.035
SAM-Med2D																	
Full	**Full**	↓.019	**.944**	.029	.033	**.859**	.024	.017	**.950**	↓.010	↓.018	**.899**	↓.002	↓.007	**.826**	.042	↓.023
Full	Enc-Only	.035	.932	↑.049	.035	.839	.025	↓.005	.931	.037	.044	.864	.009	.020	.772	.041	.031
PEFT	Adapter	.029	.935	.043	↓.034	.839	.027	.008	.940	.027	.043	.884	.011	.022	.801	↑.062	.036
PEFT	*LoRA*	.051	.921	↑.106	.045	.730	↑.045	.016	.844	↑.063	↑.109	.713	.001	↑.056	.628	.048	↑.059
Dec	Dec-Prompt	.032	.929	.053	.036	.838	.025	↓.004	.927	.037	.041	.855	.008	.017	.760	↓.036	.029
Dec	Dec-Only	.032	.929	.054	.036	.836	↓.024	↓.004	.927	.037	.041	.854	.007	.017	.758	↓.036	.031

LoRA exhibits the largest performance degradation under perturbations despite reasonable clean performance, indicating that parameter-efficient methods may sacrifice robustness for efficiency.

Strategy Ranking. As shown in Fig. 3(b), Full fine-tuning demonstrates the best overall robustness with a mean IoU drop of 0.025, followed by Dec-Only (0.029) and Enc-Only (0.029). LoRA consistently ranks worst with a mean IoU drop of 0.048—nearly double that of Full fine-tuning. This ranking holds across both MedSAM and SAM-Med2D (Fig. 3(c)), though SAM-Med2D generally exhibits higher sensitivity to perturbations than MedSAM across all strategies.

Dataset-Specific Sensitivity. Robustness varies markedly across modalities and datasets. Brain MRI segmentation is the most stable (IoU drop: 0.019), likely benefiting from a more controlled acquisition environment. In contrast, Kvasir endoscopy exhibits the highest sensitivity (IoU drop: 0.050), consistent with the challenging and variable nature of gastrointestinal imaging. Tab. 1 reports the per-dataset results and confirms that full fine-tuning achieves the best clean performance while maintaining robustness across all datasets.

Perturbation Analysis. Failure cases are dominated by modality-specific corruptions. Motion Artifacts (OCT) and Light Reflection induce the performance drops, underscoring the need for domain-specific robustness evaluation. Among general perturbations, Pixelate is the most damaging. Notably, 9 of the top 15 most challenging perturbations are medical-specific, suggesting that standard robustness benchmarks may underestimate deployment risks.

Severity Level Impact. Performance degrades monotonically as perturbation severity increases, but the degradation rate differs across strategies. LoRA exhibits the steepest curve, with IoU drop rising from 0.028 at low severity to 0.065 at high severity. Full fine-tuning shows the flattest curve, suggesting that updating all parameters helps learn more robust feature representations. The gap between strategies widens at higher severity levels, indicating that robustness differences become more pronounced under severe distribution shifts.

3.3 Vision-Language Tasks Evaluation

We evaluate five medical vision-language models (LLaVA-Med, MedGemma, MedGemma-1.5, GPT-4o-mini and Gemini-2.5-flash) on VQA, Visual Grounding, and Captioning. VQA and Captioning are evaluated in a zero-shot setting, while Visual Grounding uses LoRA fine-tuning.

Task-Specific Robustness. Robustness differs sharply across tasks and model types. Visual Grounding, which requires LoRA fine-tuning, degrades severely for medical models: MedGemma drops from 65.4% to 22.3% and MedGemma-1.5 from 69.2% to 29.0%. General-purpose models evaluated zero-shot fail entirely on this task, indicating that Grounding inherently requires task-specific adaptation **for precise localization**. For zero-shot VQA, Gemini-2.5 achieves the highest clean accuracy (67.0%) but suffers the largest drop (36.1 points, 54% relative), while GPT-4o-mini and medical models behave more stably with drops under 8 points. Captioning remains highly robust across all models with drops below 0.02 BLEU **even at high severity**. Since each task is evaluated under its realistic deployment protocol (zero-shot for VQA/captioning. LoRA for grounding, as zero-shot grounding is non-functional), these are independent per-task observations rather than a head-to-head comparison: LoRA adaptation boosts task performance at the expense of robustness, while zero-shot inference preserves pretrained stability (see Fig. 3(g)–(i)).

Perturbation Impact. Perturbations affect tasks in different ways. For LoRA-fine-tuned Grounding, Compression Artifacts (32.9 drop) and Gaussian Noise (32.6 drop) are most damaging, mirroring the sensitivity of fine-tuned segmentation models. For zero-shot VQA, Motion Blur and Gaussian Blur cause the largest drops (around 8 points), while medical-specific perturbations such as Window Level and X-ray artifacts have moderate impact. Captioning remains resilient, with all perturbations causing drops below 0.025 BLEU, suggesting that caption generation relies on higher-level semantic cues that are less sensitive to low-level corruptions (see Fig. 3(j)–(l)).

Model Comparison. General-purpose and medical-specialized models exhibit clear trade-offs. In zero-shot VQA, Gemini-2.5 reaches the highest accuracy (67.0%) but the weakest robustness (54% relative drop), whereas GPT-4o-mini is more balanced (50.0% clean, 12% drop) and performs strongly on captioning. Medical models are more stable, with MedGemma showing the smallest drops on VQA (3.1) and captioning (0.013). For grounding, general-purpose models fail in the zero-shot setting (0–10%), while LoRA-fine-tuned medical models achieve high performance (MedGemma-1.5: 69.2%) at the cost of robustness. This matches the segmentation results: fine-tuning enables specialized capabilities but increases sensitivity to perturbations (see Fig. 3(a)-(c)).

4 Conclusions

We present a robustness benchmark for medical foundation models with 40 perturbations (12 base, 28 modality-specific) across eight modalities, evaluating

segmentation (MedSAM, SAM-Med2D) and VLMs (LLaVA-Med, MedGemma, MedGemma-1.5, Gemini-2.5-flash, GPT-4o-mini). Robustness is mainly determined by fine-tuning strategy: LoRA degrades nearly twice as much as full fine-tuning. Modality-specific corruptions dominate segmentation failures top 9 corruptions. Task formulation further matters: LoRA-tuned Grounding drops $>60\%$, zero-shot Captioning stays $<7\%$, and VQA robustness is model-dependent (medical $<20\%$ vs. Gemini-2.5-flash 54%). For deployment, full fine-tuning is most robust, with SAM-Med2D adapters as a lightweight alternative. General-purpose VLMs excel at zero-shot VQA but fail on Grounding, while MedGemma is the most consistently robust medical VLM.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T.X., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., Sun, H., He, J., Zhang, S., Zhu, M., Qiao, Y.: Sam-med2d (2023), https://api.semanticscholar.org/CorpusID:261339487
2. Gutman, D.A., Codella, N.C.F., Celebi, M.E., Helba, B., Marchetti, M.A., Mishra, N.K., Halpern, A.C.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) pp. 168–172 (2016), https://api.semanticscholar.org/CorpusID:10768153
3. Hendrycks, D., Dietterich, T.G.: Benchmarking neural network robustness to common corruptions and perturbations. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019), https://openreview.net/forum?id=HJz6tiCqYm
4. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net (2022), https://openreview.net/forum?id=nZeVKeeFYf9
5. Hu, Y., Li, T.X., Lu, Q., Shao, W., He, J., Qiao, Y., Luo, P.: Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 22170–22183 (2024), https://api.semanticscholar.org/CorpusID:267657686
6. Huang, X., Shen, L., Liu, J., Shang, F., Li, H., Huang, H., Yang, Y.: Towards a multimodal large language model with pixel-level insight for biomedicine. In: AAAI Conference on Artificial Intelligence (2024), https://api.semanticscholar.org/CorpusID:274655737
7. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
8. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: Ro, Y.M., Cheng, W., Kim, J., Chu, W., Cui, P., Choi, J., Hu, M., Neve, W.D. (eds.) MultiMedia Modeling - 26th International Conference, MMM 2020, Daejeon, South Korea,

January 5-8, 2020, Proceedings, Part II. Lecture Notes in Computer Science, vol. 11962, pp. 451–462. Springer (2020). https://doi.org/10.1007/978-3-030-37734-2_37, https://doi.org/10.1007/978-3-030-37734-2_37

9. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.B.: Segment anything. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 3992–4003 (2023), https://api.semanticscholar.org/CorpusID:257952310
10. 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. ArXiv **abs/2306.00890** (2023), https://api.semanticscholar.org/CorpusID:258999820
11. Ma, J., He, Y., Li, F., Han, L.J., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15** (2023), https://api.semanticscholar.org/CorpusID:260431203
12. Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P.: Foundation models for generalist medical artificial intelligence. Nature (2023)
13. OpenAI: Gpt-4v(ision) system card (2023)
14. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
15. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.: Radiology objects in context (roco): A multimodal image dataset. In: CVII-STENT/LABELS@MICCAI (2018), https://api.semanticscholar.org/CorpusID:53087891
16. Reinhold, J.C., Dewey, B.E., Carass, A., Prince, J.L.: Evaluating the impact of intensity normalization on mr image synthesis. Proceedings of SPIE–the International Society for Optical Engineering **10949** (2018), https://api.semanticscholar.org/CorpusID:54473002
17. Sellergren, A., Kazemzadeh, S., Jaroensri, T., et al.: Medgemma technical report (2025), https://api.semanticscholar.org/CorpusID:280150648
18. Suetens, P.: Fundamentals of medical imaging, 3rd edition (2017), https://api.semanticscholar.org/CorpusID:264848844
19. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
20. Tellez, D., Litjens, G.J.S., Bándi, P., Bulten, W., Bokhorst, J.M., Ciompi, F., van der Laak, J.: Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Medical image analysis **58**, 101544 (2019), https://api.semanticscholar.org/CorpusID:62841444
21. Thirunavukarasu, A., Ting, D., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.: Large language models in medicine. Nature medicine (2023)
22. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
23. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**, 600–612 (2004), https://api.semanticscholar.org/CorpusID:207761262

24. Zhao, X., Huang, W., Wang, X., Zhao, H., Zhuang, L., Jiang, A., Wan, G., Ye, M.: Divide, conquer and unite: Hierarchical style-recalibrated prototype alignment for federated medical segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 28760–28768 (2026)