MedQ-Deg: A Multidimensional Benchmark for Evaluating MLLMs Across Medical Image Quality Degradations

Jiyao Liu[1*], Junzhi Ning[2*], Chenglong Ma[1*], Wanying Qu[1], Jianghan Shen[2], Siqi Luo[3], Jinjie Wei[1], Jin Ye[2], Pengze Li[2], Tianbin Li[2], Jiashi Lin[2], Hongming Shan[1], Xinzhe Luo[4], Xiaohong Liu[3], Lihao Liu[2], Junjun He[2†], and Ningsheng Xu[1†]

[1] Fudan University, Shanghai, China
jiyaoliu.fudan@gmail.com
[2] Shanghai AI Lab, Shanghai, China
[3] Shanghai Jiaotong University, Shanghai, China
[4] Imperial College London, London, UK
Project: https://uni-medical.github.io/MedQ-Robust-web
Dataset: https://huggingface.co/datasets/jiyaoliufd/MedQ-DEG-Bench

Abstract. Despite impressive performance on standard benchmarks, multimodal large language models (MLLMs) face critical challenges in real-world clinical environments where medical images inevitably suffer various quality degradations. Existing benchmarks exhibit two key limitations: (1) absence of large-scale, multidimensional assessment across medical image quality gradients and (2) no systematic confidence calibration analysis. To address these gaps, we present **MedQ-Deg**, a comprehensive benchmark for evaluating medical MLLMs under image quality degradations. MedQ-Deg provides multi-dimensional evaluation spanning 18 distinct degradation types, 30 fine-grained capability dimensions, and 7 imaging modalities, with 24,894 question-answer pairs. Each degradation is implemented at 3 severity degrees, calibrated by expert radiologists. We further introduce *Calibration Shift* metric, which quantifies the gap between a model's prediction-consistency-based perceived confidence and actual performance to assess metacognitive reliability under degradation. Our comprehensive evaluation of 40 mainstream MLLMs reveals several critical findings: (1) overall model performance degrades systematically as degradation severity increases, (2) models exhibit a Dunning-Kruger-like overconfidence pattern, maintaining inappropriately high confidence despite severe accuracy collapse, and (3) models display markedly differentiated behavioral patterns across capability dimensions, imaging modalities, and degradation types. *We hope MedQ-Deg drives progress toward medical MLLMs that are robust and trustworthy in real clinical practice.*

Keywords: Medical Multimodal Large Language Models · Robustness · Image Corruption · Benchmark

*Equal contribution. †Corresponding author.

1 Introduction

Multimodal Large Language Models (MLLMs) [3, 18, 36, 37] have recently demonstrated remarkable performance on medical vision-language benchmarks, in some cases approaching or even surpassing human experts [15, 49]. However, these impressive results largely rely on carefully curated high-quality medical images. In real clinical environments, medical images are frequently degraded due to noise, motion artifacts, or hardware limitations, raising a critical question: can MLLMs remain reliable under such imperfect conditions?

This vulnerability becomes evident when we consider real-world clinical environments that are far more complex than standard benchmark datasets suggest. Medical images in nature suffer from various degradations, including noise from low-dose scanning, motion artifacts from patient movement, field inhomogeneity from aging equipment, and illumination variations in endoscopy, particularly in low-resource settings or smaller clinics where advanced imaging equipment is unavailable. As illustrated in Fig. 3, models exhibit consistent accuracy decline across increasing degradation severity degrees. While human physicians can usually navigate these imperfections and still reach a correct diagnosis, the reliability of MLLMs in such scenarios remains largely unproven, creating a significant barrier to safe deployment in healthcare.

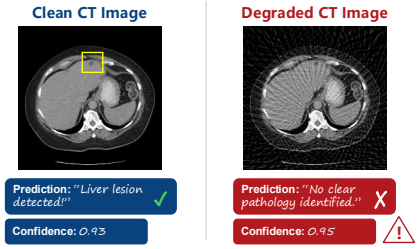


Fig. 1: Illustration of Dunning-Kruger-like overconfidence. An MLLM correctly identifies a liver lesion in the clean CT image but yields erroneous predictions with similarly high confidence when the image is corrupted while the lesion is still visible.

More alarmingly, our pilot investigations reveal a systematic pattern of metacognitive failure in MLLMs when confronted with degraded medical images. Models not only suffer accuracy drops but also exhibit a striking inability to recognize the boundaries of their own competence, maintaining inappropriately high confidence while giving erroneous predictions. As illustrated in Fig. 1, an MLLM correctly detected the liver lesion in a computed tomography (CT) image when no degradation is present; however, when the same image is degraded (*e.g.*, with sparse-view CT artifacts due to the subsampling of raw measurement data [19]), the model yields false-negative predictions while assigning similarly high confidence. This disconnect between capability and self-assessment manifests across diverse degradations and severity degrees, suggesting a fundamental deficit in model awareness rather than isolated failures.

We refer to this phenomenon as an **AI Dunning-Kruger Effect**, using the term as an analogy to the cognitive bias where individuals with limited ability overestimate their competence [12]. This metacognitive blindness, *i.e.*, the inability to "know what it does not know," poses particularly severe risks

in clinical deployment, as overconfident erroneous inferences may prevent the activation of necessary human oversight and mislead physicians into trusting unreliable AI recommendations.

Despite these critical vulnerabilities, existing evaluation frameworks remain inadequate for systematically assessing model performance under image quality variations. Current benchmarks suffer from two fundamental limitations: (1) *Absence of large-scale multidimensional assessment under medical image degradations.* General image quality benchmarks [23, 34] typically focus on natural image corruptions (*e.g.*, Gaussian noise, blur, JPEG compression) that fail to capture medical imaging-specific degradations such as motion artifacts in magnetic resonance imaging (MRI). Although recent work [8, 48] has begun exploring medical MLLM performance under degradations, they lack the scale and comprehensiveness necessary for systematic evaluation. (2) *No systematic confidence calibration analysis.* Existing medical benchmarks provide only coarse-grained overall accuracy metrics, without quantifying model confidence calibration and metacognitive awareness under degradation, which are precisely the dimensions essential for ensuring safe clinical deployment.

To systematically analyze the aforementioned phenomena and address these gaps, we propose **MedQ-Deg**, a benchmark for comprehensively evaluating medical MLLM robustness against clinically realistic image degradations. As shown in Fig. 2. Our benchmark enables both fine-grained capability evaluation under diverse medical image degradations and systematic analysis of model confidence calibration, addressing the critical dimensions missing from existing evaluation frameworks. Our contributions are threefold:

(1) A systematic benchmark with a hierarchical evaluation framework. We construct MedQ-Deg, which features a three-tier framework along two axes: *capability* (spanning from mid-level categories through 6 clinical tasks to 30 fine-grained skills) and *degradation* (covering 18 subcategories across 7 imaging modalities, with each instantiated at 3 severity degrees, each validated by expert radiologists). This hierarchical design enables structured and clinically meaningful assessment across image quality degradation types and degrees.

(2) Quantitative evidence of Dunning-Kruger-like overconfidence in medical MLLMs. We introduce *Calibration Shift*, a metric that quantifies the gap between a model's perceived confidence and its actual accuracy under degradation. Using this metric, we provide large-scale empirical evidence that medical MLLMs remain markedly overconfident even as their true capabilities deteriorate, and this overconfidence systematically widens with increasing degradation severity. This metacognitive failure reveals that current models lack the self-awareness required for safe clinical deployment.

(3) Comprehensive evaluation across models and multidimensions. We conduct extensive evaluation of 40 mainstream MLLMs spanning commercial MLLMs, open-source general models, and medical-specialized models. Our multidimensional analysis examines performance degradation across capability dimensions, degradation categories, and imaging modalities, providing the most

Table 1: Comparison with related benchmarks. "–" indicates the information is not reported in the original publication.

Benchmark	Publication	Total QA	#Modality	#Capability	#Degrad.
MedMNIST-C [11]	Arxiv 2024	520k (train & eval.)	9	1 (classification)	17
RobustMedCLIP [17]	MIUA 2025	137k (train & eval.)	5	1 (classification)	7
Pathology Corruptions [53]	MICCAI 2022	327k (train & eval.)	1	1 (classification)	9
ROOD-MRI [5]	NeuroImage 2023	1,475 (eval.)	1	1 (segmentation)	11
VLM Artefacts [8]	npj Digit. Med. 2025	6,600 (eval.)	3	1 (detection)	5
RobustMed-Bench [48]	Arxiv 2025	1,500 (eval.)	3	–	6
MedQ-Deg (Ours)	**This paper**	**24,894** (eval.)	**7**	**30**	**18**

comprehensive characterization of medical MLLM behavior under image quality variations to date.

2 Related Work

2.1 Medical Multimodal Large Language Models and Benchmarks

The application of Multimodal Large Language Models in medical domains has witnessed rapid advancement. To better address medical-specific challenges, a series of *medical-specialized models* have emerged, including GMAI-VL-R1 [39], UniMedVL [30], Hulu-Med [18], Lingshu [43], HealthGPT [24], and MedGemma [36]. These models are typically pre-trained or fine-tuned on large-scale medical image-text datasets to enhance domain-specific understanding.

To systematically evaluate these medical MLLMs, numerous benchmarks have been developed targeting various medical scenarios. GMAI-MMBench [49] integrates multiple medical datasets covering clinical tasks, department categorization, and perception granularity. OmniMedVQA [15] aggregates various medical VQA datasets to provide comprehensive evaluation. Modality-specific benchmarks include PathVQA [13] for pathology images and VQA-RAD [20] for radiology. SLAKE [25] provides bilingual medical VQA with segmentation masks. General multimodal benchmarks such as MMMU [50] and MMMU-Pro [51] encompass multidisciplinary tasks including medical content. Recent benchmarks MedXpertQA [55] for expert-level exam-style question answering. Despite their comprehensive coverage, these benchmarks primarily evaluate performance on standard clean data without systematically considering robustness to image degradations that commonly occur in real-world clinical environments.

2.2 Evaluating MLLMs Under Image Degradations

In computer vision, ImageNet-C [14] pioneered systematic evaluation by introducing 15 common corruption types for image classification. This framework has been extended to evaluate MLLMs under various image quality conditions [22,34], revealing that even state-of-the-art models exhibit significant performance degradation when confronted with corrupted natural images.

Early efforts in medical imaging have explored degradation robustness within narrow scopes. MedMNIST-C [11] extends the MedMNIST suite with common corruptions but does not include QA evaluation or capability assessment. RobustMedCLIP [17] examines the robustness of medical vision-language models yet is limited to contrastive learning settings without a structured benchmark. Zhang *et al.* [53] benchmark pathology-specific corruptions for classification tasks but restrict their scope to a single modality. ROOD-MRI [5] targets out-of-distribution robustness for MRI segmentation under 11 degradation variants, covering only one modality and one task type. While these works establish important precedents, they are largely confined to limited modalities, single tasks, or traditional vision models, leaving the robustness of MLLMs across diverse medical imaging settings substantially unexplored.

Recent benchmarks have begun examining MLLM robustness to natural image degradation, including R-Bench [23] and Bench-C [40]. Some works expand the evaluation to medical MLLM behavior under image perturbations. Liu *et al.* [26–28] evaluate MLLMs' capability in assessing medical image quality. Cheng *et al.* [8] constructed a VLM artifacts benchmark across multiple modalities and degradation types, but evaluation is confined to disease detection without broader clinical reasoning assessment. Xu *et al.* [48] proposes RobustMed-Bench, targeting image degradation robustness, but omits capability decomposition, severity calibration, and confidence analysis. Together, these efforts remain limited in scope, covering only a handful of modalities or degradation types without a comprehensive assessment across diverse medical tasks and capability dimensions. Tab. 1 provides a structured comparison of MedQ-Deg against representative medical image degradation benchmarks, highlighting the gaps that motivate our work.

3 The MedQ-Deg Benchmark

MedQ-Deg benchmark is designed to evaluate how image quality degradation affects the diagnostic performance and metacognitive reliability of existing MLLMs. As illustrated in Fig. 2, the benchmark is built upon a dual-tier hierarchy that stratifies both model capabilities (Section 3.1) and clinical degradation patterns (Section 3.2), supported by a human-curated data pipeline that ensures clinical realism (Section 3.3) and a novel suite of metacognitive metrics designed to quantify model overconfidence under stress (Section 3.4). By doing so, we move beyond simple accuracy scores to provide multidimensional performance assessment and fine-grained capability characterization.

3.1 Medical MLLM Capability Hierarchy

To comprehensively assess MLLM clinical competence, we construct a three-tier capability hierarchy grounded in the cognitive workflow of a practising clinician (Fig. 2, left). Tasks are sourced from three top-tier medical benchmarks, namely GMAI-MMBench [49], OmniMedVQA [15], and MedXpertQA [55], with

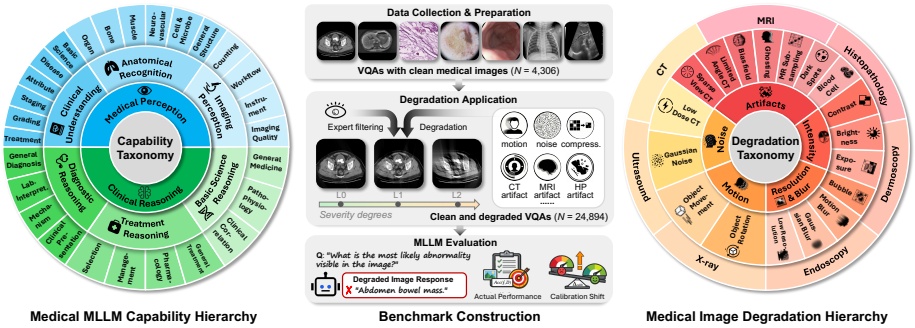


Fig. 2: Overview of the MedQ-Deg benchmark framework. Two orthogonal hierarchies structure the evaluation: a *capability hierarchy* (left) decomposing clinical competence into 30 fine-grained skills across 6 tasks, and a *degradation hierarchy* (right) covering 18 degradation types across 7 modalities. The data pipeline (middle) applies each degradation at three expert-calibrated severity degrees (L0–L2).

redundant items merged and the task structure reorganized to align with clinically meaningful capability dimensions. Full task definitions and statistics are provided in the supplementary material (Section A). The hierarchy is organized as follows:

(1) High-level Capabilities ($\mathcal{T}^{(1)}$): We partition the capability space into two fundamental categories: *medical perception* and *clinical reasoning*. The former focuses on extracting visual and structural information from medical images, while the latter involves higher-level clinical interpretation and decision-making. **(2) Mid-level Tasks** ($\mathcal{T}^{(2)}$): We decompose each high-level capability into six clinically relevant dimensions that mirror the cognitive workflow of a practising clinician: *anatomical recognition*, *imaging perception*, *clinical understanding*, *basic science reasoning*, *diagnostic reasoning*, and *treatment reasoning* constitute the reasoning branch. **(3) Fine-grained Skills** ($\mathcal{T}^{(3)}$): Each mid-level task is further stratified into specific clinical skills, yielding 30 fine-grained capabilities that span the full spectrum of medical image understanding competencies. These decompositions follow the clinical diagnostic process, progressing from raw image perception to actionable clinical decision-making, and ensuring that each dimension captures a distinct and non-redundant cognitive step.

3.2 Medical Image Degradation Hierarchy

We organize degradations into a two-level hierarchy. At the top level, five categories are grounded in clinical imaging physics: artifacts, intensity jitter, "resolution & blur", motion interference, and noise. Within each category, types are further partitioned by formation mechanism into general (cross-modality, *e.g.*, motion blur) and modality-specific corruptions (*e.g.*, low-dose CT noise), as illustrated in the right panel of Fig. 2. Detailed definitions and statistics are provided in the supplementary material (Section A).

We define a degradation operation $\mathcal{C}_i^s$ mathematically as:

$$\mathcal{I}_c = \mathcal{C}_i^s(\mathcal{I}_0; \theta_i^s), \tag{1}$$

where $\mathcal{I}_0$ denotes an original diagnostic-quality image, $i \in \{1, 2, ..., N\}$ indexes the degradation variants, and $s \in \{0, 1, 2\}$ represents the severity degree. The severity parameters θ_i^s are calibrated independently for each degradation type. For each modality–degradation pair, we sample approximately 20 cases at 9 parameter levels and ask three board-certified radiologists to independently select the feasible L1 range, where diagnostic features remain intact, and the feasible L2 range, where diagnosis becomes challenging yet feasible. The overlap among the three experts defines the final parameter interval; cases outside the consensus interval are not used for benchmark generation. L0 retains the original clean image without degradation. All degradation variants are implemented using TorchIO [33], Torch Radon [35], SciPy ndimage [44], and Froodo [38]. As all corruptions in MedQ-Deg are synthetically generated, their alignment with real-world clinical degradations will be validated in Section 4.6.

3.3 Dataset Construction

After cross-source deduplication based on image hash and question-text similarity, we obtain a pool of unique VQA pairs (L0) from three established benchmarks: OmniMedVQA [15], GMAI-MMBench [49], and MedXpertQA [55]. For each VQA pair, we randomly apply three modality-specific degradations at both severity degrees L1 and L2 on the image.

To maintain clinical validity, we implemented a human-in-the-loop filtering process in which three board-certified radiologists independently reviewed each degraded image–question pair under two criteria: (1) the applied corruptions must not completely obliterate the diagnostic features required to answer the question, and (2) the question must not be trivially answerable from text cues alone without reference to the image. Samples failing either criterion were discarded. Only unanimously retained pairs were included in the final benchmark. The unanimous agreement rate in this filtering stage was 92.5%, and approximately 8.3% of generated pairs were removed through this process, ensuring that every retained sample demands genuine visual reasoning under degradation. The resulting benchmark comprises 24,894 QA pairs in total.

3.4 Evaluation Metrics

To assess model robustness and metacognitive capabilities under image degradation, we utilize a combination of accuracy and uncertainty metrics for quantifying both actual performance and calibration shift.

Actual Performance. For each VQA pair in our dataset, models generate predictions through multiple-choice selection. We evaluate performance using accuracy as the primary metric:

$$\mathrm{Acc}(f, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \mathbb{I}\big[\hat{y}(x) = y^*(x)\big], \tag{2}$$

where $\mathcal{D}$ denotes the evaluation dataset, $y^*(x)$ represents the ground-truth answer, $\hat{y}(x)$ is the model's prediction, and $\mathbb{I}[\cdot]$ is the indicator function.

Perceived Confidence Proxy. We measure model confidence using voting-based prediction consistency. This proxy captures how consistently a model selects the same answer across stochastic trials, and should be interpreted as a practical confidence proxy rather than direct access to a model's internal probability. The voting distribution is $p_i(x) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{I}\big[y^{(t)} = i\big]$ for $i \in \{1, \ldots, K\}$, where K is the number of answer choices and $y^{(t)}$ denotes the prediction in trial t. We quantify prediction uncertainty using normalized entropy: $H(x) = -\sum_{i=1}^{K} p_i(x) \log p_i(x)$, with maximum entropy $H_{\max} = \log K$ for complete uncertainty. Sample-level confidence is defined as the inverse of normalized entropy:

$$C(x) = 1 - H(x)/\log K, \tag{3}$$

where $C(x) \in [0, 1]$, with $C(x) = 1$ indicating perfect prediction consistency and $C(x) \approx 0$ for high uncertainty.

Calibration Shift. Model calibration shift on dataset $\mathcal{D}$ is quantified as the gap between perceived confidence and actual accuracy:

$$\Delta_{\text{calib}}(f, \mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{x \in \mathcal{D}} C(x) - \text{Acc}(f, \mathcal{D}). \tag{4}$$

A positive Δ_{calib} indicates overconfidence (model overestimates its capability), zero represents well-calibrated confidence, and negative values indicate underconfidence.

We use Δ_{calib} to characterize two distinct forms of Dunning-Kruger-like overconfidence. *Intra-Model DKE* occurs when a model's accuracy collapses as image quality drops from L0 to L2, but the calibration shift remains stable or increases, defined as:

$$\text{Acc}(f, \mathcal{D}_{\text{L0}}) > \text{Acc}(f, \mathcal{D}_{\text{L2}}) \ \wedge \ \Delta_{\text{calib}}(f, \mathcal{D}_{\text{L0}}) \leq \Delta_{\text{calib}}(f, \mathcal{D}_{\text{L2}}), \tag{5}$$

where $\mathcal{D}_{\text{L0}}$ and $\mathcal{D}_{\text{L2}}$ denote evaluation sets at L0 and L2 severity degree respectively. This reveals a metacognitive failure where the model is "unaware" of its own declining capability. *Inter-Model DKE* describes a capability-correlated trend that lower-performing models show higher calibration shift compared to higher-performing models; this cross-model pattern may also reflect training or architecture differences and is therefore interpreted more cautiously than the intra-model trend:

$$\text{Acc}(f_1, \mathcal{D}) < \text{Acc}(f_2, \mathcal{D}) \ \wedge \ \Delta_{\text{calib}}(f_1, \mathcal{D}) > \Delta_{\text{calib}}(f_2, \mathcal{D}). \tag{6}$$

4 Experiments

This section presents a comprehensive evaluation of representative MLLMs through our hierarchical benchmark. Based on the experimental results, we systematically characterize the performance of these models and provide in-depth

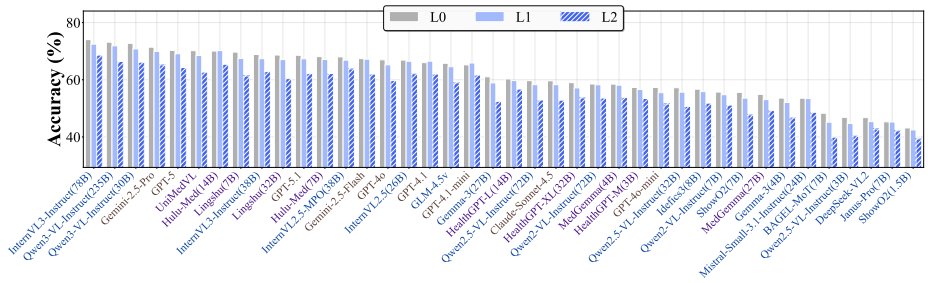


Fig. 3: Model performance across medical image quality degradation. Accuracy of 40 MLLMs evaluated at three severity degrees (L0–L2). X-axis labels are colored by model category: blue for open-source general models, brown for commercial MLLMs, and purple for medical-specialized models.

interpretation of their strengths and limitations across diverse evaluation scenarios. Specifically, we begin by examining overall robustness across degradation severity degrees (Section 4.2), then analyze how performance varies across clinical capability dimensions (Section 4.3) and degradation types (Section 4.4), characterize the overconfidence phenomenon that emerges as degradation increases (Section 4.5), and finally validate that the simulated degradations align with real-world clinical distributions (Section 4.6).

4.1 Experimental Setup

We evaluate 40 representative MLLMs spanning three groups: (1) **9 *Commercial MLLMs***, including GPT-5 [37], GPT-5.1 [32], GPT-4o-mini, GPT-4o [16], GPT-4.1-mini, GPT-4.1 [31], Gemini-2.5-Pro [9], Gemini-2.5-Flash [9] and Claude-Sonnet-4.5 [2]; (2) **21 *Open-source general models***, including InternVL3-Instruct (78B, 38B) [54], InternVL2.5-MPO (38B) [7], InternVL2.5 (26B) [7], Qwen3-VL-Instruct (235B, 30B) [3], Qwen2.5-VL-Instruct (72B, 32B, 3B) [4], Qwen2-VL-Instruct (72B, 7B) [45], Gemma-3 (27B, 4B) [41], Idefics3 (8B) [21], Mistral-Small-3.1-Instruct (24B) [1], ShowO2 (7B, 1.5B) [47], BAGEL-MoT (7B) [10], DeepSeek-VL2 [46], GLM-4.5v [42], and Janus-Pro (7B) [6]. (3) **10 *Medical-specialized models***, including UniMedVL (14B) [30], Hulu-Med (14B, 7B) [18], Lingshu (32B, 7B) [43], HealthGPT-XL/L/M (32B, 14B, 3B) [24], and MedGemma (27B, 4B) [36].

To ensure statistically significant assessments of model confidence and uncertainty, we perform $T = 3$ independent inference trials for performance experiment and $T = 10$ for calibration shift experiment. All models are evaluated with temperature 1.0. Detailed are provided in the supplementary material (Section B).

Table 2: Capability-level performance (L1&L2) with degradation magnitude. Average accuracy across the mid-level capability dimensions, computed over L1 and L2 degradations. Numbers in parentheses show performance drop from L0 to L1&L2. "CU": Clinical Understanding. "IP": Imaging Perception. "AR": Anatomical Recognition. "BS": Basic Science. "Diag.": Diagnosis. "Treat.": Treatment.

		Medical Perception			Clinical Reasoning		
Model	**Avg.**	**CU**	**IP**	**AR**	**BS**	**Diag.**	**Treat.**
Commercial MLLMs							
GPT-5	64.5(↓3.5)	67.2(↓4.3)	36.5(↓0.9)	67.4(↓4.4)	75.0(↓6.0)	62.1(↓2.9)	78.8(↓2.3)
Gemini-2.5-Pro	58.2(↓3.8)	66.9(↓3.4)	30.0(↓6.5)	72.0(↓5.7)	60.7(↓8.3)	60.1(↓1.3)	59.6(↑2.6)
Gemini-2.5-Flash	50.7(↓2.9)	63.6(↓4.8)	25.5(↓1.2)	71.2(↑0.1)	58.9(↓5.7)	38.9(↓3.2)	46.1(↓2.8)
GPT-5.1	49.6(↓5.9)	65.9(↓3.3)	30.5(↓6.0)	65.8(↓4.2)	50.0(↓9.5)	47.0(↓8.5)	38.5(↓3.9)
GPT-4.1	48.9(↓2.3)	66.5(↓2.8)	32.5(↓1.9)	63.8(↓1.1)	50.0(↓4.1)	38.4(↓2.1)	42.3(↓1.8)
GPT-4o	47.5(↓3.2)	64.8(↓5.6)	32.0(↓2.5)	61.9(↓4.3)	42.9(↑0.2)	37.4(↓3.9)	46.1(↓3.0)
GPT-4.1-mini	47.0(↓1.6)	65.9(↓1.9)	34.0(↓2.0)	63.5(↓1.8)	33.9(↓1.2)	44.4(↓2.7)	40.4(↓0.1)
Claude-Sonnet-4.5	45.1(↓3.1)	62.8(↓3.8)	22.0(↓4.5)	44.3(↓6.3)	51.8(↑2.3)	35.9(↓4.6)	53.9(↓1.4)
GPT-4o-mini	31.6(↓2.9)	57.3(↓3.8)	24.5(↓6.7)	50.4(↓5.7)	12.5(↓0.1)	29.3(↓0.2)	15.4(↓0.9)
Avg. Commercial MLLMs	49.2(↓3.2)	64.5(↓3.7)	29.7(↓3.6)	62.3(↓3.7)	48.4(↓3.6)	43.7(↓3.3)	46.8(↓1.5)
Open-Source General Models							
Qwen3-VL-Instruct(235B)	43.8(↓4.2)	73.0(↓3.7)	40.5(↓8.1)	66.8(↓4.9)	37.5(↓2.3)	29.8(↓4.9)	15.4(↓1.5)
InternVL3-Instruct(78B)	39.3(↓3.8)	75.9(↓3.1)	35.0(↑0.3)	66.3(↓3.1)	21.4(↓9.1)	29.3(↓7.0)	7.7(↓0.6)
GLM-4.5v	38.5(↓3.5)	66.4(↓3.6)	25.5(↓1.7)	57.7(↓6.3)	37.5(↓5.1)	32.3(↑0.1)	11.5(↓4.3)
Qwen3-VL-Instruct(30B)	38.1(↓4.1)	73.0(↓2.6)	25.5(↓4.7)	66.1(↓8.1)	14.3(↓9.1)	30.3(↑1.8)	19.2(↓2.1)
InternVL2.5-MPO(38B)	36.5(↓1.6)	72.6(↓2.3)	30.0(↓0.1)	57.0(↓4.1)	25.0(↓1.8)	22.7(↓0.5)	11.5(↓0.8)
Mistral-Small-3.1-Instruct(24B)	33.9(↓2.5)	56.3(↓2.9)	25.0(↓0.3)	43.9(↓2.3)	25.0(↓0.7)	33.8(↓1.3)	19.2(↓7.7)
InternVL3-Instruct(38B)	33.4(↓4.0)	71.7(↓3.0)	24.0(↓8.0)	58.3(↓5.8)	10.7(↓4.5)	31.8(↓2.6)	3.9(↓0.3)
InternVL2.5(26B)	33.0(↓2.4)	71.8(↓2.8)	27.5(↓6.4)	55.7(↓4.5)	12.5(↓0.4)	22.7(↓0.1)	7.7(↑0.1)
Qwen2.5-VL-Instruct(72B)	31.8(↓4.0)	63.4(↓4.3)	28.5(↓3.7)	44.5(↓5.7)	19.6(↓2.3)	28.8(↓5.9)	5.8(↓1.9)
Gemma-3(27B)	31.5(↓3.2)	58.7(↓4.9)	23.0(↓0.1)	55.2(↓6.1)	14.3(↓3.6)	18.7(↓4.6)	19.2(↑0.1)
Gemma-3(4B)	31.5(↓4.3)	54.4(↓2.6)	32.5(↓0.1)	43.2(↓7.2)	17.9(↓15.0)	14.1(↓1.2)	26.9(↑0.1)
Qwen2-VL-Instruct(7B)	31.4(↓2.5)	61.7(↓2.8)	30.5(↓1.6)	39.7(↓3.0)	23.2(↑1.5)	19.7(↓3.3)	13.5(↓5.8)
Idefics3(8B)	31.0(↓1.5)	60.0(↓3.0)	25.0(↓0.8)	46.2(↓3.6)	17.9(↓1.1)	25.2(↓0.5)	11.5(↑0.2)
Qwen2.5-VL-Instruct(32B)	29.9(↓1.4)	61.9(↓4.4)	27.5(↓0.1)	39.7(↓4.0)	17.9(↑0.1)	32.3(↓0.1)	0.0(↓0.0)
Qwen2-VL-Instruct(72B)	27.3(↓3.4)	62.8(↓3.3)	24.0(↓1.6)	47.8(↓3.1)	5.4(↓0.3)	20.2(↓0.9)	3.9(↓11.5)
Janus-Pro(7B)	27.1(↓1.8)	50.7(↓1.5)	37.0(↓4.2)	32.7(↓2.7)	30.4(↓1.5)	11.6(↓1.0)	0.0(↓0.0)
ShowO2(7B)	26.8(↓3.0)	58.6(↓4.0)	25.0(↓0.1)	39.2(↓8.2)	7.1(↑0.1)	21.2(↓0.1)	9.6(↓5.7)
ShowO2(1.5B)	26.6(↓2.6)	46.3(↓2.6)	24.0(↓5.3)	33.2(↓1.2)	21.4(↓4.5)	15.7(↓0.5)	19.2(↓1.8)
Qwen2.5-VL-Instruct(3B)	25.4(↓3.4)	48.7(↓5.0)	28.0(↓3.1)	33.6(↓4.5)	26.8(↓6.8)	15.2(↓0.8)	0.0(↓0.0)
BAGEL-MoT(7B)	24.7(↓2.9)	47.7(↓5.6)	32.5(↓2.4)	35.4(↓7.9)	12.5(↓0.5)	16.2(↓0.8)	3.9(↑0.1)
DeepSeek-VL2	23.2(↓1.7)	47.4(↓1.9)	29.5(↓2.8)	42.0(↓3.9)	7.1(↓0.4)	1.5(↑0.3)	11.5(↓1.2)
Avg. Open-Source Models	31.6(↓2.9)	61.1(↓3.3)	28.6(↓2.6)	47.8(↓4.8)	19.3(↓3.2)	22.5(↓1.6)	10.5(↓2.1)
Medical-specialized Models							
UniMedVL(14B)	42.5(↓4.0)	69.2(↓3.6)	40.0(↓3.7)	63.1(↓5.6)	25.0(↓1.6)	32.8(↓7.6)	25.0(↓1.9)
Hulu-Med(7B)	38.8(↓2.4)	71.0(↓4.2)	31.0(↑1.0)	57.5(↓5.0)	28.6(↓3.4)	27.3(↓0.9)	17.3(↓1.9)
Hulu-Med(14B)	38.4(↓1.5)	72.8(↓2.1)	34.0(↓2.3)	64.0(↓4.6)	26.8(↓0.1)	25.2(↓0.1)	7.7(↑0.1)
Lingshu(32B)	36.1(↓3.9)	70.4(↓3.8)	27.5(↑0.5)	56.2(↓8.7)	26.8(↓8.3)	27.8(↓2.9)	7.7(↓0.5)
Lingshu(7B)	35.8(↓5.1)	71.7(↓4.4)	23.5(↓8.6)	56.3(↓8.1)	23.2(↓2.1)	20.7(↓6.3)	19.2(↓1.4)
HealthGPT-L(14B)	35.2(↓2.1)	66.6(↓1.5)	33.0(↓4.6)	45.3(↓4.3)	17.9(↓4.2)	40.9(↑2.8)	7.7(↓0.6)
MedGemma(4B)	35.0(↓1.9)	62.1(↓2.9)	30.0(↓1.9)	48.0(↓2.5)	17.9(↑2.3)	26.8(↓0.6)	25.0(↓5.8)
HealthGPT-M(3B)	32.6(↓1.9)	63.7(↓2.6)	28.5(↓0.8)	42.0(↓1.9)	14.3(↓3.7)	27.8(↓0.7)	19.2(↓1.7)
MedGemma(27B)	31.4(↓2.4)	59.3(↓4.1)	26.5(↓0.1)	38.0(↓3.5)	10.7(↑0.1)	42.4(↓5.6)	11.5(↓1.1)
HealthGPT-XL(32B)	30.5(↓3.2)	61.4(↓3.9)	33.0(↓0.3)	48.5(↓2.6)	17.9(↓9.1)	14.7(↓2.6)	7.7(↓0.5)
Avg. Medical-Specialized	35.6(↓2.8)	66.8(↓3.3)	30.7(↓2.1)	51.9(↓4.7)	20.9(↓3.0)	28.6(↓2.4)	14.8(↓1.5)
Avg. Performance (All Models)	**36.6**(↓3.0)	**63.3**(↓3.4)	**29.4**(↓2.7)	**52.1**(↓4.5)	**26.3**(↓3.2)	**28.8**(↓2.2)	**19.8**(↓1.8)

4.2 Findings on Severity Degrees

Conclusion 1. Most evaluated MLLMs exhibit severe robustness deficiency with nonlinear degradation patterns. As shown in Fig. 3, the majority of the MLLMs suffer significant performance drops as image quality decreases. Notably, even the best-performing model InternVL3-Instruct (78B) across all levels experiences substantial accuracy drops at a degradation severity of L2, showing that robustness deficiency remains a widespread challenge across MLLMs.

Table 3: Degradation-category performance (L1&L2) with degradation magnitude. Average accuracy across five degradation categories, computed over L1 and L2 degradations. Numbers in parentheses show performance drop from L0 to L1&L2. "R&B": resolution and blur issues.

Model	Avg.	Intensity	Noise	R&B	Artifacts	Motion
Commercial MLLMs						
Gemini-2.5-Pro	68.2(↓3.7)	71.1(↓0.9)	69.2(↓2.3)	67.2(↓2.0)	65.1(↓5.3)	68.6(↓8.0)
GPT-5	67.5(↓4.2)	72.2(↓4.0)	66.2(↓3.2)	67.5(↓2.5)	63.1(↓6.4)	68.4(↓4.8)
GPT-5.1	65.7(↓4.2)	70.9(↓2.5)	64.2(↓4.8)	65.7(↓4.2)	60.5(↓6.2)	67.2(↓3.4)
Gemini-2.5-Flash	65.4(↓2.5)	67.5(↓0.6)	65.1(↓1.8)	64.8(↓1.0)	60.7(↓5.0)	68.7(↓4.0)
GPT-4.1	65.0(↓2.1)	69.4(↓3.7)	62.7(↓0.8)	65.1(↓2.4)	60.6(↓3.2)	67.4(↓0.6)
GPT-4.1-mini	64.4(↓1.6)	67.1(↓1.6)	63.3(↑0.8)	64.4(↓1.4)	60.5(↓3.1)	66.8(↓2.5)
GPT-4o	63.4(↓4.5)	68.8(↓1.3)	61.8(↓5.5)	63.2(↓4.6)	58.3(↓6.7)	65.0(↓4.2)
GPT-4o-mini	54.7(↓4.5)	62.3(↓2.7)	53.9(↓4.3)	55.1(↓3.2)	47.8(↓6.0)	54.2(↓6.1)
Claude-Sonnet-4.5	55.7(↓5.1)	59.1(↓2.7)	53.3(↓5.6)	56.6(↓4.5)	55.2(↓7.0)	54.3(↓5.5)
Avg. Commercial MLLMs	63.3(↓3.6)	67.6(↓2.2)	62.2(↓3.1)	63.3(↓2.9)	59.1(↓5.4)	64.5(↓4.3)
Open-Source General Models						
InternVL3-Instruct(78B)	71.4(↓3.2)	75.1(↓3.6)	69.3(↓2.8)	72.7(↓2.0)	64.3(↓5.2)	75.6(↓2.5)
Qwen3-VL-Instruct(235B)	70.0(↓4.1)	75.3(↓2.9)	69.8(↓2.7)	70.7(↓2.2)	63.3(↓6.2)	71.1(↓6.5)
Qwen3-VL-Instruct(30B)	69.4(↓4.2)	76.2(↓1.0)	68.0(↓4.5)	71.3(↓2.2)	61.4(↓7.1)	70.3(↓6.4)
InternVL2.5-MPO(38B)	66.0(↓2.6)	70.1(↓1.4)	64.2(↓1.7)	66.5(↓2.4)	61.9(↓4.4)	67.4(↓3.1)
InternVL3-Instruct(38B)	65.8(↓3.8)	70.5(↓1.6)	64.0(↓4.3)	67.4(↓2.7)	59.9(↓6.2)	67.2(↓4.1)
InternVL2.5(26B)	65.2(↓3.5)	73.6(↓3.2)	62.6(↓1.9)	66.3(↓3.0)	59.9(↓4.1)	63.7(↓5.3)
GLM-4.5v	62.4(↓4.6)	65.5(↓1.7)	63.9(↓0.6)	62.6(↓3.8)	57.2(↓7.4)	63.0(↓9.4)
Qwen2-VL-Instruct(72B)	56.2(↓3.5)	58.5(↓3.5)	54.8(↓2.4)	57.5(↓2.5)	53.2(↓5.4)	57.1(↓3.5)
Gemma-3(27B)	56.5(↓4.4)	58.2(↓1.7)	58.7(↓1.8)	55.2(↓5.7)	51.4(↓7.0)	59.1(↓5.6)
Qwen2.5-VL-Instruct(72B)	55.8(↓4.8)	57.6(↓3.4)	53.0(↓5.1)	57.2(↓3.7)	53.6(↓7.5)	57.5(↓4.1)
Idefics3(8B)	55.3(↓2.9)	65.7(↓0.9)	50.5(↓4.7)	55.7(↓3.1)	48.0(↓2.3)	56.6(↓3.6)
Qwen2.5-VL-Instruct(32B)	53.0(↓4.6)	50.6(↓4.3)	51.8(↓4.2)	54.2(↓2.5)	52.3(↓7.2)	56.0(↓4.7)
Qwen2-VL-Instruct(7B)	53.0(↓3.0)	53.7(↓1.2)	52.6(↓2.7)	53.1(↓2.2)	53.0(↓4.5)	52.5(↓4.3)
Mistral-Small-3.1-Instruct(24B)	51.3(↓2.6)	52.4(↓0.2)	49.0(↓2.8)	51.8(↓0.5)	49.5(↓5.5)	53.7(↓3.8)
ShowO2(7B)	52.0(↓4.9)	62.3(↓1.6)	47.0(↓6.3)	52.8(↓5.9)	45.2(↓4.5)	52.6(↓6.1)
Gemma-3(4B)	50.1(↓4.0)	54.7(↓1.9)	49.9(↓2.4)	50.1(↓5.3)	46.0(↓5.0)	49.9(↓5.6)
Janus-Pro(7B)	44.6(↓1.8)	53.7(↓0.5)	41.2(↓3.7)	46.9(↓1.3)	38.4(↓2.9)	42.7(↓0.5)
Qwen2.5-VL-Instruct(3B)	43.5(↓4.2)	48.0(↓3.5)	39.6(↓6.0)	43.0(↓4.0)	39.3(↓7.0)	47.7(↓0.7)
DeepSeek-VL2	43.3(↓2.5)	33.0(↓1.1)	46.9(↓2.5)	43.0(↓1.9)	45.0(↓3.6)	48.6(↓3.4)
BAGEL-MoT(7B)	44.0(↓5.6)	53.9(↓2.9)	39.9(↓8.2)	44.0(↓5.8)	37.0(↓5.5)	45.1(↓5.6)
ShowO2(1.5B)	41.6(↓1.8)	47.8(↓1.3)	38.9(↓1.7)	42.4(↓2.0)	37.8(↓2.0)	41.0(↓1.8)
Avg. Open-Source Models	55.7(↓3.7)	59.8(↓2.1)	54.1(↓3.5)	56.4(↓3.1)	51.3(↓5.3)	57.1(↓4.3)
Medical-specialized Models						
Hulu-Med(14B)	68.2(↓3.0)	70.2(↓2.3)	66.3(↓2.2)	69.7(↓1.8)	64.0(↓4.9)	70.6(↓4.0)
UniMedVL(14B)	67.1(↓3.8)	76.1(↓0.9)	65.3(↓2.4)	66.1(↓4.2)	60.1(↓6.5)	67.8(↓5.0)
Hulu-Med(7B)	65.4(↓3.9)	69.9(↓1.1)	63.5(↓3.1)	65.8(↓3.8)	61.1(↓6.1)	66.5(↓5.3)
Lingshu(7B)	65.5(↓5.3)	73.7(↓3.1)	62.2(↓3.7)	65.9(↓4.3)	60.0(↓5.2)	65.6(↓10.1)
Lingshu(32B)	65.1(↓4.9)	74.0(↓1.9)	62.1(↓3.8)	65.1(↓4.3)	57.7(↓6.4)	66.7(↓8.2)
HealthGPT-L(14B)	58.8(↓2.6)	63.5(↓2.0)	57.0(↓1.6)	59.5(↓1.8)	55.3(↓4.1)	58.8(↓3.4)
MedGemma(4B)	57.0(↓3.3)	66.0(↓1.8)	54.8(↓1.9)	59.2(↓2.5)	49.3(↓5.8)	55.5(↓4.5)
HealthGPT-XL(32B)	56.3(↓3.6)	62.2(↓2.1)	55.0(↓2.6)	56.1(↓2.0)	51.7(↓6.4)	56.7(↓4.9)
HealthGPT-M(3B)	55.5(↓2.1)	58.9(↓1.4)	54.7(↓0.8)	55.8(↓1.6)	52.3(↓4.3)	56.0(↓2.5)
MedGemma(27B)	52.2(↓3.7)	62.2(↓0.4)	47.4(↓5.1)	52.3(↓2.3)	48.8(↓5.2)	50.2(↓5.4)
Avg. Medical-Specialized	61.1(↓3.6)	67.7(↓1.7)	58.8(↓2.7)	61.5(↓2.9)	56.0(↓5.5)	61.4(↓5.3)
Avg. Performance (All Models)	**58.8**(↓3.6)	**63.5**(↓2.0)	**57.1**(↓3.2)	**59.2**(↓3.0)	**54.2**(↓5.4)	**59.8**(↓4.6)

Furthermore, the performance degradation suggests a nonlinear acceleration pattern beyond a simple monotonic decline. While many models show some tolerance from L0 to L1 severity degrees, the performance deterioration rate increases sharply from L1 to L2. Additional dense sampling on five representative degradation types with nine parameter levels shows the same sharper drop near the severe range, supporting a "cliff effect" where model perception remains relatively stable until a certain degradation threshold is reached.

4.3 Findings on Capability Dimensions

We analyze robustness across the capability dimensions. Tab. 2 provides comprehensive performance.

Conclusion 2. Except for a few top-tier commercial models, most MLLMs perform poorly in clinical reasoning, and existing MLLMs exhibit a certain degree of fragility across all capability dimensions. Across all groups, Clinical Understanding is strongest, while reasoning dimensions (Basic Science, Diagnosis, Treatment) are critically weak, with Treatment planning being the most catastrophic where multiple open-source models collapse to near-zero accuracy.

Commercial MLLMs substantially outperform open-source models in clinical reasoning, whereas leading open-source models remain competitive in Medical Perception, even surpassing commercial models in some dimensions. Medical-specialized models show no consistent advantage over general models. Regarding robustness under degradation, Treatment planning shows the smallest average performance drop among models with non-trivial L0 treatment accuracy, while Anatomical Recognition exhibits the poorest robustness with the largest degradation magnitude; for models already near zero on Treatment planning, this apparent resilience is dominated by floor effects.

4.4 Findings on Degradation-Type Sensitivity

We analyze model sensitivity to different degradation types. Comprehensive degradation heatmaps showing performance drops from baseline (L0) to gradual degradations (L1 and L2) across different models and degradation types are presented in Tab. 3.

Conclusion 3. MLLMs are strikingly vulnerable to physics-based artifacts and motion interference compared to intensity-based degradations. The five degradation categories produce markedly different impacts on model robustness. Degradations involving "intensity jitter" demonstrate the highest resilience with the smallest performance drops, while "noise" and "resolution & blur" show moderate impact. In sharp contrast, "artifact" degradations, which encompass medical physics-specific corruptions such as MRI undersampling artifacts and sparse-view CT artifacts, cause larger drops. Motion-induced degradations such as object rotation, also show substantial impact, ranking second in severity after artifacts. This further indicates that medical artifacts and motion interference are substantially more destructive in corrupting the image semantics than other types of degradations.

This vulnerability hierarchy reveals that current MLLMs lack sufficient understanding of medical imaging-specific degradation mechanisms that are absent from natural image pretraining distributions. Besides, modality-specific sensitivity patterns are further analyzed in the supplementary material (Section C).

4.5 Findings on Overconfidence Under Degradations

Furthermore, to investigate the metacognitive ability of MLLMs, we analyze the relationship between actual performance and model confidence under degradation.

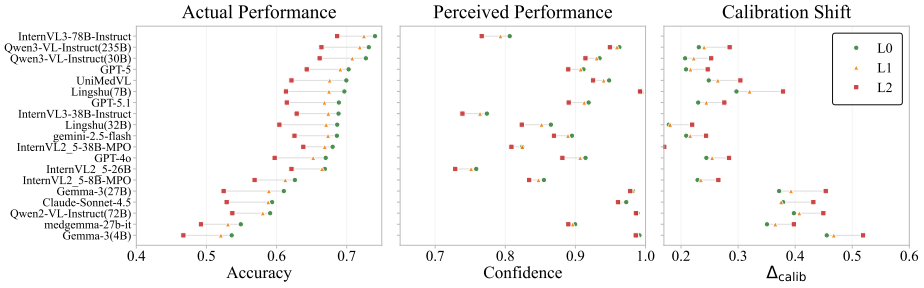


Fig. 4: Comprehensive analysis of model calibration shift across severity degrees. These panels quantify the widening gap between actual accuracy and model certainty under increasing image degradation. As degradation scales from L0 to L2, the stability of perceived performance against collapsing accuracy reveals a fundamental failure in metacognitive awareness across MLLMs.

Following the Calibration Shift proposed in Section 3.4, we measure the gap between perceived confidence and actual performance.

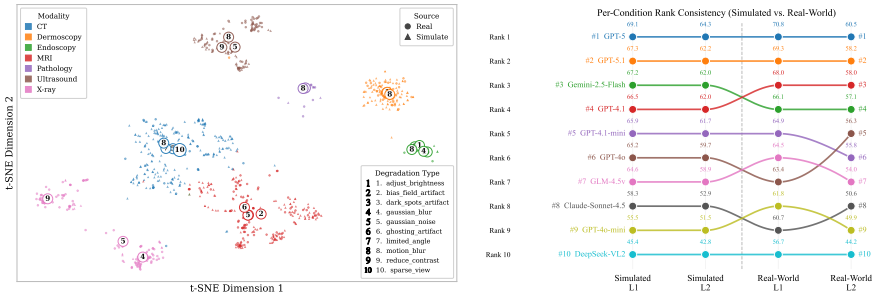
Conclusion 4. All models exhibit severe overconfidence under degradation, demonstrating a Dunning-Kruger-like pattern. As shown in Fig. 4, models maintain inappropriately high perceived confidence despite substantial accuracy degradation. Across all models and quality degrees, calibration shift remains consistently positive and increases systematically from L0 to L2.

Our analysis reveals this phenomenon manifests in two dimensions. *Intra-Model DKE* characterizes metacognitive failure within individual models: as degradation severity increases from L0 to L2, actual performance drops substantially while calibration shift consistently increases, indicating models grow more overconfident precisely when their capabilities deteriorate. This pattern holds across all evaluated architectures under the voting-based proxy. *Inter-Model DKE* describes the inverse relationship across different models: at severe degradation degrees, lower-performing models overall exhibit disproportionately higher calibration shift compared to higher-capability models.

4.6 Distribution Alignment Validation

A key concern for any benchmark that includes simulated images is whether the simulated degradations are representative of those encountered in real-world clinical practice. We address this question through two complementary analyses: (1) a feature-space distribution comparison using t-SNE analysis [29], and (2) a cross-evaluation rank consistency study.

Feature Distribution Analysis. The 2,000 real-world degraded QA pairs are non-overlapping with MedQ-Deg and are drawn from the same source benchmarks, namely OmniMedVQA, GMAI-MMBench, and MedXpertQA. We first use Qwen3-VL-Instruct (235B) to identify low-quality candidate images, then ask the same three radiologists to filter the candidates, assign mild/severe degradation labels, and resample by modality and degradation type to balance the



(a) (b)

Fig. 5: Simulated vs. Real-World Evaluation Consistency. *Left:* t-SNE projection of BiomedCLIP features from simulated and real degraded images. Real and simulated images co-locate within each modality cluster, confirming distribution alignment. *Right:* Bump chart tracing per-model ranks across Simulated-Mild, Simulated-Severe, Real-World-Mild, and Real-World-Severe conditions.

validation set. We extract feature vectors from 2,000 simulated and 2,000 real-world degraded images using BiomedCLIP [52], a vision encoder pretrained on large-scale biomedical image-text pairs. We then project these feature vectors into a two-dimensional space using t-SNE. As shown in Fig. 5 (a), *simulated and real degraded images exhibit highly overlapping feature distributions within each modality cluster*, suggesting that the degradations simulated by our pipeline capture characteristics similar to those observed in real-world data.

Rank Consistency Analysis. We further evaluate whether model rankings on simulated images faithfully predict rankings on the 2,000 real-world degraded clinical QA pairs. As shown in Fig. 5 (b), *per-model rank trajectories remain highly consistent across Simulated-Mild, Simulated-Severe, Real-World-Mild, and Real-World-Severe conditions.* Across all 40 models, Spearman's rank correlation is 0.92 for simulated L1 versus real-world mild degradation and 0.90 for simulated L2 versus real-world severe degradation. Only a few adjacent-rank swaps are observed (*e.g.*, GPT-4.1 and Gemini-2.5-Flash), while the global ordering remains stable, with GPT-5 and GPT-5.1 dominating and DeepSeek-VL2 consistently at the bottom. Together, these results suggest that MedQ-Deg evaluation scores obtained on simulated degradations are useful proxies for real-world clinical performance rankings, although the validation supports ranking consistency rather than exact score equivalence.

5 Discussion

This paper presents **MedQ-Deg**, a large-scale and multidimensional benchmark for evaluating medical MLLMs under clinically realistic image quality degradations. Our evaluation of 40 models reveals that robustness deficiency is widespread and often nonlinear: most models tolerate mild corruptions (L0 to L1) reasonably well but collapse sharply at severity degrees (L1 to L2), suggesting a fundamental

fragility in vision-language integration beyond a certain degradation threshold. Capability-level analysis (Section 4.3) reveals that most models already perform poorly in clinical reasoning tasks at baseline, and this weakness persists under degradation. Notably, Anatomical Recognition exhibits the poorest degradation robustness despite being a perception-oriented task, while Treatment planning shows the greatest resilience, suggesting that degradation sensitivity is shaped more by the granularity of visual detail required than by whether a task involves reasoning or perception. At the degradation level (Section 4.4), physics-based artifacts and motion interference cause substantially larger performance drops than intensity or resolution perturbations, consistent with the absence of such domain-specific corruptions from natural image pretraining distributions.

Perhaps the most clinically alarming finding is the Dunning-Kruger-like overconfidence pattern (Section 4.5): as degradation severity increases, model accuracy collapses while perceived confidence remains high. This metacognitive failure holds across all 40 architectures under the voting-based proxy and is corroborated on a subset by verbalized and logprob-based confidence proxies, posing direct patient safety risks by preventing appropriate human oversight from being triggered. Developing models that are not only accurate but also well-calibrated under degradation is therefore a critical open problem that we hope MedQ-Deg will help drive.

A potential concern for simulation-based benchmarks is ecological validity. Our distribution alignment analysis (Section 4.6) shows that simulated and real degraded images overlap in BiomedCLIP feature space, and that model rankings on simulated data strongly correlate with rankings on real clinical images, supporting MedQ-Deg as a useful proxy for real-world robustness assessment. *We believe MedQ-Deg provides essential infrastructure for developing medical MLLMs that are not only accurate on clean data, but reliable under the imperfect conditions of real clinical practice.*

References

1. AI, M.: Mistral small 3.1. `https://mistral.ai/news/mistral-small-3-1/` (2025), accessed: 2026-01-28
2. Anthropic: Claude Sonnet 4.5. `https://www.anthropic.com/news/claude-sonnet-4-5` (2025)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Boone, L., Biparva, M., Forooshani, P.M., Ramirez, J., Masellis, M., Bartha, R., Symons, S., Strother, S., Black, S.E., Heyn, C., et al.: Rood-mri: Benchmarking the robustness of deep learning segmentation models to out-of-distribution and corrupted data in mri. NeuroImage **278**, 120289 (2023)

6. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-Pro: Unified multimodal understanding and generation with data and model scaling (2025)
7. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling (2025)
8. Cheng, Z., Ong, A.Y., Wagner, S.K., Merle, D.A., Ju, L., Zhang, H., Chen, R., Pang, L., Li, B., He, T., et al.: Understanding the robustness of vision-language models to medical image artefacts. NPJ Digital Medicine **8**(1), 727 (2025)
9. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
10. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., Fan, H.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
11. Di Salvo, F., Doerrich, S., Ledig, C.: Medmnist-c: Comprehensive benchmark and improved classifier robustness by simulating realistic image corruptions. arXiv preprint arXiv:2406.17536 (2024)
12. Dunning, D.: The dunning–kruger effect: On being ignorant of one's own ignorance. In: Advances in experimental social psychology, vol. 44, pp. 247–296. Elsevier (2011)
13. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: PathVQA: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
14. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019)
15. Hu, Y., Li, T., Lu, Q., Shao, W., He, J., Qiao, Y., Luo, P.: OmniMedVQA: A new large-scale comprehensive evaluation benchmark for medical LVLM. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22170–22183 (2024)
16. 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)
17. Imam, R., Marew, R., Yaqub, M.: On the robustness of medical vision-language models: Are they truly generalizable? In: Annual Conference on Medical Image Understanding and Analysis. pp. 233–256. Springer (2025)
18. Jiang, S., Wang, Y., Song, S., Hu, T., Zhou, C., Pu, B., Zhang, Y., Yang, Z., Feng, Y., Zhou, J.T., et al.: Hulu-Med: A transparent generalist model towards holistic medical vision-language understanding. arXiv preprint arXiv:2510.08668 (2025)
19. Jin, K.H., McCann, M.T., Froustey, E., Unser, M.: Deep convolutional neural network for inverse problems in imaging. IEEE Transactions on Image Processing **26**(9), 4509–4522 (2017)
20. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific Data **5**(1), 1–10 (2018)
21. Laurençon, H., Marafioti, A., Sanh, V., Tronchon, L.: Building and better understanding vision-language models: insights and future directions. arXiv preprint arXiv:2408.12637 (2024)
22. Li, C., Wang, Z., Sheng, Y., Zhu, X., Hao, Y., Wang, X.: Res-bench: Benchmarking the robustness of multimodal large language models to dynamic resolution input. arXiv preprint arXiv:2510.16926 (2025)

23. Li, C., Zhang, J., Zhang, Z., Wu, H., Tian, Y., Sun, W., Lu, G., Min, X., Liu, X., Lin, W., Zhai, G.: R-bench: Are your large multimodal model robust to real-world corruptions? IEEE Journal of Selected Topics in Signal Processing (2025)
24. Lin, T., Zhang, W., Li, S., Yuan, Y., Yu, B., Li, H., He, W., Jiang, H., Li, M., xiaohui, S., Tang, S., Xiao, J., Lin, H., Zhuang, Y., Ooi, B.C.: HealthGPT: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. In: Forty-second International Conference on Machine Learning (2025)
25. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th international symposium on biomedical imaging (ISBI). pp. 1650–1654. IEEE (2021)
26. Liu, J., Ning, J., Qu, W., Liu, L., Ma, C., He, J., Xu, N.: Medq-engine: A closed-loop data engine for evolving mllms in medical image quality assessment. arXiv preprint arXiv:2603.19863 (2026)
27. Liu, J., Ning, J., Qu, W., Liu, L., Ma, C., He, J., Xu, N.: Medq-uni: Toward unified medical image quality assessment and restoration via vision-language modeling. arXiv preprint arXiv:2603.18465 (2026)
28. Liu, J., Wei, J., Qu, W., Ma, C., Ning, J., Li, Y., Chen, Y., Luo, X., Chen, P., Gao, X., et al.: MedQ-Bench: Evaluating and exploring medical image quality assessment abilities in MLLMs. arXiv preprint arXiv:2510.01691 (2025)
29. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(86), 2579–2605 (2008)
30. Ning, J., Li, W., Tang, C., Lin, J., Ma, C., Zhang, C., Liu, J., Chen, Y., Gao, S., Liu, L., et al.: UniMedVL: Unifying medical multimodal understanding and generation through observation-knowledge-analysis. arXiv preprint arXiv:2510.15710 (2025)
31. OpenAI: GPT-4.1 model card. Tech. rep., OpenAI (2025), `https://openai.com/index/gpt-4-1/`, accessed: 2026-01-28
32. OpenAI: GPT-5.1 Instant and GPT-5.1 Thinking system card addendum. Tech. rep., OpenAI (2025), `https://openai.com/index/gpt-5-system-card-addendum-gpt-5-1/`, accessed: 2026-01-28
33. Pérez-García, F., Sparks, R., Ourselin, S.: TorchIO: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Computer Methods and Programs in Biomedicine **208**, 106236 (2021)
34. Qiu, X., Kan, M., Zhou, Y., Shan, S.: Benchmarking multimodal large language models against image corruptions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9014–9023 (2025)
35. Ronchetti, M.: Torchradon: Fast differentiable routines for computed tomography. arXiv preprint arXiv:2009.14788 (2020)
36. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: MedGemma technical report. arXiv preprint arXiv:2507.05201 (2025)
37. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: OpenAI GPT-5 system card. arXiv preprint arXiv:2601.03267 (2025)
38. Stieber, J., Fuchs, M., Mukhopadhyay, A.: Froodo: Framework for out-of-distribution detection. arXiv preprint arXiv:2208.00963 (2022)
39. Su, Y., Li, T., Liu, J., Ma, C., Ning, J., Tang, C., Ju, S., Ye, J., Chen, P., Hu, M., et al.: GMAI-VL-R1: Harnessing reinforcement learning for multimodal medical reasoning. arXiv preprint arXiv:2504.01886 (2025)

40. Sui, X., Li, S., Zhu, H., Chen, B., Fang, Y., Sun, X.: Benchmarking corruption robustness of LVLMs: A discriminative benchmark and robustness alignment metric. arXiv preprint arXiv:2511.19032 (2025)
41. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)
42. Team, G.V., Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., et al.: GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning (2026)
43. Team, L., Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
44. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al.: Scipy 1.0: fundamental algorithms for scientific computing in python. Nature Methods **17**(3), 261–272 (2020)
45. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
46. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024)
47. Xie, J., Yang, Z., Shou, M.Z.: Show-o2: Improved native unified multimodal models. arXiv preprint arXiv:2506.15564 (2025)
48. XU, D., Yang, X., Li, Y., Miao, J., Li, J., Heng, P.A.: Perceive and calibrate: Analyzing and enhancing robustness of medical multi-modal large language models. arXiv preprint arXiv:2512.21964 (2025)
49. Ye, J., Wang, G., Li, Y., Deng, Z., Li, W., Li, T., Duan, H., Huang, Z., Su, Y., Wang, B., et al.: GMAI-MMBench: A comprehensive multimodal evaluation benchmark towards general medical ai. Advances in Neural Information Processing Systems **37**, 94327–94427 (2024)
50. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9556–9567 (2024)
51. Yue, X., Zheng, T., Ni, Y., Wang, Y., Zhang, K., Tong, S., Sun, Y., Yu, B., Zhang, G., Sun, H., et al.: MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 15134–15186 (2025)
52. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI **2**(1), AIoa2400640 (2025)
53. Zhang, Y., Sun, Y., Li, H., Zheng, S., Zhu, C., Yang, L.: Benchmarking the robustness of deep neural networks to common corruptions in digital pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 242–252. Springer (2022)

54. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)
55. Zuo, Y., Qu, S., Li, Y., Chen, Z.R., Zhu, X., Hua, E., Zhang, K., Ding, N., Zhou, B.: MedXpertQA: Benchmarking expert-level medical reasoning and understanding. In: Forty-second International Conference on Machine Learning (2025)