# MedQ-Engine: A Closed-Loop Data Engine for Evolving MLLMs in Medical Image Quality Assessment

Jiyao Liu[1†], Junzhi Ning[2†], Wanying Qu[1], Lihao Liu[2], Chenglong Ma[1], Junjun He[2✉], and Ningsheng Xu[1✉]

[1] Fudan University, Shanghai, China
[2] Shanghai Artificial Intelligence Laboratory, Shanghai, China

**Abstract.** Medical image quality assessment (Med-IQA) is a prerequisite for clinical AI deployment, yet multimodal large language models (MLLMs) still fall substantially short of human experts, particularly when required to provide descriptive assessments with clinical reasoning beyond simple quality scores. However, improving them is hindered by the high cost of acquiring descriptive annotations, and by the inability of one-time data collection to adapt to the model's evolving weaknesses. To address these challenges, we propose **MedQ-Engine**, a closed-loop data engine that iteratively *evaluates* the model to discover failure prototypes via data-driven clustering, *explores* a million-scale image pool using these prototypes as retrieval anchors with progressive human-in-the-loop annotation, and *evolves* through quality-assured fine-tuning, forming a self-improving cycle. Models are evaluated on complementary *perception* and *description* tasks. An entropy-guided routing mechanism triages annotations to minimize labeling cost. Experiments across five medical imaging modalities show that MedQ-Engine elevates an 8B-parameter model to surpass GPT-4o by over 13% and narrow the gap with human experts to only 4.34%, using only 10K annotations with more than 4× sample efficiency over random sampling. We will release our dataset, model and code publicly.

**Keywords:** Image Quality Assessment · Multimodal Large Language Models · Data Engine

## 1 Introduction

Medical image quality assessment (Med-IQA) is essential for reliable diagnosis, yet the heterogeneity of modalities, anatomical regions, and clinical scenarios poses significant challenges [26]. Existing approaches either produce modality-agnostic scalar scores [16, 26] or are confined to specific modalities [6, 13, 25]. Multimodal large language models (MLLMs) offer a promising cross-modality

[†]Equal contribution. [✉]Corresponding author.

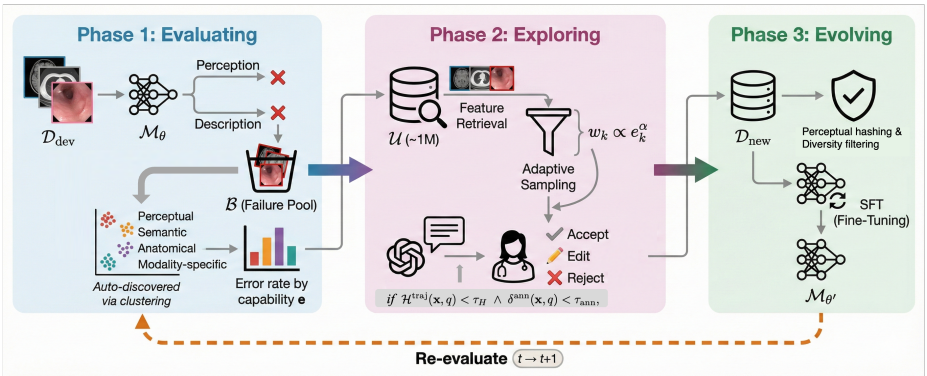


**Fig. 1. Overview of MedQ-Engine.** Our closed-loop data engine iteratively improves MLLMs for Med-IQA via three phases: *Evaluating* (failure clustering on a dev set), *Exploring* (prototype-based retrieval and human-in-the-loop annotation), and *Evolving* (quality-assured fine-tuning with re-evaluation).

alternative, capable of generating descriptive assessments that identify degradation types, analyze visual impact, and evaluate degradation severity [28, 32]. Yet recent benchmarking reveals that MLLMs still exhibit substantial deficiencies in Med-IQA, with systematic performance gaps compared to human experts [14,15]. Notably, MedQ-Bench [15] shows that *MLLM errors concentrate in specific capability-modality intersections rather than being uniformly distributed*, suggesting that targeted remediation is far more efficient than uniform data augmentation. However, translating this insight into practice faces two obstacles: a *cost-value dilemma* where simple scoring provides minimal training signal while comprehensive expert descriptions are prohibitively expensive, and *static data collection* that cannot adapt to evolving model weaknesses as new bottlenecks emerge after each improvement.

To address these challenges, we propose **MedQ-Engine**, a closed-loop data engine that systematically improves MLLM capabilities for Med-IQA through three iterative phases (Fig. 1): (1) *Evaluating*, which assesses the model on a dedicated development set and clusters failure cases into *failure prototypes* that capture dominant error patterns; (2) *Exploring*, which uses these prototypes as retrieval anchors to expand training data from a large-scale image pool (∼1M images), combined with human-in-the-loop verification of AI-generated pre-annotations; (3) *Evolving*, which fine-tunes on quality-assured data and re-evaluates, forming a closed loop that progressively resolves model weaknesses.

**Contributions.** (1) We propose MedQ-Engine, the first closed-loop data engine for Med-IQA that transforms data-driven error analysis into systematic model improvement through iterative evaluate-explore-evolve phases. (2) We introduce a data-driven failure discovery mechanism with error-weighted adaptive sampling, combined with a human-in-the-loop annotation paradigm that maximizes information gain per expert minute. (3) Extensive experiments and ablation studies across five medical imaging modalities demonstrate that MedQ-

Engine substantially narrows the gap with human experts using only 10K samples.

## 2 Method

### 2.1 Problem Formulation

Following [15], we formulate Med-IQA as two complementary tasks. *Perception task* evaluates quality-related visual perception through multi-choice questions, including: (1) *Yes-or-No* questions for binary quality classification (e.g., "Does this image contain artifacts?"), (2) *What* questions for multi-class identification of degradation types, and (3) *How* questions for severity assessment. *Description task* requires generating comprehensive responses that cover modality and anatomy identification, degradation characterization, technical attribution, visual impact assessment, and an overall quality judgment.

Let $\mathcal{D}_{\text{test}}$ denote the evaluation benchmark and $\mathcal{U}$ denote the unlabeled image pool ($|\mathcal{U}| \approx 10^6$). We additionally construct a development set $\mathcal{D}_{\text{dev}}$, held out independently from $\mathcal{D}_{\text{test}}$, to enable iterative failure analysis without data leakage. Our goal is to select a training subset $\mathcal{D}_{\text{train}} \subset \mathcal{U}$ under annotation budget $B$ that maximally improves $\mathcal{M}_\theta$ (an autoregressive LLM policy parameterized by $\theta$) on $\mathcal{D}_{\text{test}}$. Since exhaustive search over this combinatorial space is intractable, MedQ-Engine approximates the solution through an iterative evaluate-explore-evolve loop that progressively identifies and resolves the most impactful failure modes.

### 2.2 Phase 1: Evaluating

**Failure Case Collection.** We evaluate $\mathcal{M}_\theta$ on $\mathcal{D}_{\text{dev}}$ $R$ times across perception and description tasks spanning multiple modalities. A sample is identified as a failure case and added to the failure pool $\mathcal{B}$ when its error rate across $R$ runs exceeds a threshold $\gamma$, indicating a persistent weakness rather than stochastic variance. Each failure case is associated with a capability dimension label vector $\mathbf{c} = (c_1, \ldots, c_K)$.

**Data-driven Failure Clustering.** Rather than imposing predefined error categories, we characterize failure patterns directly from model behavior. For each failure case in $\mathcal{B}$, we represent it as a feature vector that combines its visual content and question-answer information. We then apply agglomerative clustering on these feature vectors and select the number of clusters via the silhouette criterion, yielding $N_c$ cluster centroids $\{\mathbf{p}_1, \ldots, \mathbf{p}_{N_c}\}$ that serve as *failure prototypes*. In Phase 2, only the visual component of each prototype is used as a retrieval anchor to query the unlabeled image pool.

**Capability Dimension Analysis.** We aggregate $\mathcal{B}$ by capability dimensions to compute error rate distribution:

$$\mathbf{e} = (e_1, \ldots, e_K), \quad e_k = \frac{|\{b \in \mathcal{B} : c_k(b) = 1\}|}{|\{s \in \mathcal{D}_{\text{dev}} : c_k(s) = 1\}|}, \tag{1}$$

which reveals systematic model weaknesses and guides subsequent data collection.

### 2.3 Phase 2: Exploring

**Prototype-based Retrieval.** We construct the ∼1M images pool $\mathcal{U}$ with five modalities, sourced from FastMRI [30], AAPM CT-MAR [10], MR-ART [18], HistoArtifacts [9], EndoCV2020 [19], EyeQ [8], GMAI-MMBench [27], Omni-iMedVQA [11], and Radiopaedia[3], and encoded using BiomedCLIP [31]. Rather than retrieving per individual failure case, we use the $N_c$ failure prototypes from Phase 1 as query anchors. Since each prototype is derived from image and Q-A features jointly, we extract its visual component $\mathbf{p}_j^{\text{vis}}$ for retrieval against the image-only embeddings in $\mathcal{U}$:

$$\mathcal{N}(\mathbf{p}_j) = \{\mathbf{x} \in \mathcal{U} : \cos(\mathbf{p}_j^{\text{vis}}, f_{\text{enc}}(\mathbf{x})) > \tau_{\text{sim}}\}. \tag{2}$$

Adaptive sampling weights $w_k \propto e_k^{\alpha}$ prioritize weak capability dimensions, forming annotation set $\mathcal{S}$ with $|\mathcal{S}| = B$.

**Progressive Human-in-the-loop Annotation.** We design a progressive strategy that adapts human effort to annotation difficulty across iterations, substantially reducing expert cost as the model evolves.

*Cold start (t=0):* GPT-4o [1] pre-annotates the initial 2K samples and domain experts review all annotations via Accept / Edit / Reject, building a high-quality seed dataset $\mathcal{D}_{\text{new}}^{(0)}$.

*Self-evolution (t>0):* For each new sample, we obtain both a self-annotation $\hat{y}^{\text{self}}$ from $\mathcal{M}_{\theta^{(t)}}$ and a reference annotation $\hat{y}^{\text{GPT}}$ from GPT-4o, then route it into one of three paths based on two signals: (i) model confidence, measured by trajectory-level entropy [2]:

$$\mathcal{H}^{\text{traj}}(\mathbf{x}, q) = -\frac{1}{|\hat{y}^{\text{self}}|} \sum_{l=1}^{|\hat{y}^{\text{self}}|} \log \mathcal{M}_{\theta^{(t)}}(\hat{y}_l^{\text{self}} \mid \mathbf{x}, q, \hat{y}_{<l}^{\text{self}}), \tag{3}$$

and (ii) oracle-agreement $\delta^{\text{ann}} \triangleq \mathcal{R}(\hat{y}^{\text{self}}, \hat{y}^{\text{GPT}})$. The routing logic is: **(a)** if the model is *uncertain* ($\mathcal{H}^{\text{traj}} \geq \tau_H$), adopt $\hat{y}^{\text{GPT}}$; **(b)** if *confident yet disagrees* with the oracle ($\mathcal{H}^{\text{traj}} < \tau_H \wedge \delta^{\text{ann}} < \tau_{\text{ann}}$), escalate to expert review; **(c)** if *confident and consistent*, adopt $\hat{y}^{\text{self}}$ directly.

### 2.4 Phase 3: Evolving

**Quality Assurance.** We ensure clinical reliability through: (1) *Deduplication* via perceptual hashing [29]; (2) *Diversity filtering* using TF-IDF [20] thresholds to remove near-duplicate descriptions.

[3] `https://radiopaedia.org/`

**Model Fine-tuning.** We fine-tune $\mathcal{M}_\theta$ using supervised instruction tuning with full parameter updating:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(\mathbf{x},q,y)\sim\mathcal{D}_{\text{new}}}\left[\sum_{i=1}^{|y|}\log p_\theta(y_i \mid y_{<i}, \mathbf{x}, q)\right]. \tag{4}$$

**Closed-loop Iteration.** After fine-tuning, the updated model $\mathcal{M}_{\theta'}$ re-enters the Evaluating phase, updating failure pool $\mathcal{B}^{(t+1)}$ and error distribution $\mathbf{e}^{(t+1)}$. Iteration terminates when performance on $\mathcal{D}_{\text{dev}}$ plateaus.

## 3 Experiments

### 3.1 Experimental Setup

**Datasets.** We construct our unlabeled image pool $\mathcal{U}$ from five medical imaging modalities: MRI, CT, endoscopy, fundus photography, and histopathology, totaling approximately 1 million images from public repositories. We additionally curate a development set $\mathcal{D}_{\text{dev}}$ ($\sim$2k samples) for iterative failure analysis, which is held out independently from the test benchmark to prevent data leakage. To ensure strict data integrity, we verify that $\mathcal{U}$, $\mathcal{D}_{\text{dev}}$, and $\mathcal{D}_{\text{test}}$ are sourced from disjoint patient cohorts and scanning sessions; perceptual-hash deduplication confirms zero image overlap across all three splits. The Med-IQA abilities of MLLMs after instruction tuning are quantitatively evaluated on MedQ-Bench [15] in two complementary tasks: *Perception*, by measuring the accuracy of answering multi-choice questions (MCQ) related to image quality attributes; and *Description*, which examines how MLLMs can describe image quality issues in natural language.

**Evaluation Metrics.** For *Perception*, we report type accuracy measuring correct identification of degradation categories. For *Description*, we employ four expert-evaluated dimensions: Completeness, assessing coverage of quality issues; Preciseness, measuring description accuracy; Consistency, evaluating alignment between analysis and conclusion; and Quality Accuracy, verifying correctness of clinical impact assessment.

**Comparison Methods.** We compare against state-of-the-art MLLMs spanning four categories: (1) *Open-source general MLLMs*: Qwen2.5-VL-Instruct (7B/32B/72B) [4] and InternVL3 (8B/38B) [5]; (2) *Medical-specialized MLLMs*: BiMediX2 [17], Lingshu [24], and MedGemma [21]; (3) *Closed-source MLLMs*: Mistral-Medium-3 [12], Claude-4-Sonnet [3], Gemini-2.5-Pro [7], Grok-4 [23], and GPT-4o [1]. We also include human expert and non-expert performance as reference baselines.

**Implementation Details.** We use Qwen2.5-VL-7B [4] and InternVL3-8B [5] as base models $\mathcal{M}_\theta$, with GPT-4o [1] as the annotation oracle. We set similarity threshold $\tau_{\text{sim}}$=0.75, evaluation runs $R$=5, and failure frequency threshold $\gamma$=0.6. Fine-tuning uses learning rate $2\times10^{-5}$ with cosine schedule over 3 epochs. The annotation budget per iteration is $B$=2000 samples, with the first iteration

**Table 1.** Performance comparison on the MedQbench+ benchmark. Green : open-source general MLLMs; Blue : medical-specialized MLLMs; Orange : closed-source MLLMs; Yellow : after MedQ-Engine optimization.

| **Sub-categories** | **Perception** | | | | **Description** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Yes/No↑** | **What↑** | **How↑** | **Overall↑** | **Comp.↑** | **Prec.↑** | **Cons.↑** | **Qual.↑** | **Overall↑** |
| *random guess* | 50.00 | 28.48 | 33.30 | 37.94 | | | | | |
| *Non-experts* | 67.50 | 57.50 | 57.50 | 62.50 | - | - | - | - | - |
| *Human experts* | 88.50 | 77.50 | 77.50 | 82.50 | - | - | - | - | - |
| BiMediX2-8B [17] | 44.98 | 27.52 | 27.81 | 35.10 | 0.376 | 0.394 | 0.281 | 0.670 | 1.721 |
| Lingshu-32B [24] | 50.36 | 50.39 | 51.74 | 50.88 | 0.624 | 0.697 | 1.932 | 1.059 | 4.312 |
| MedGemma-27B [21] | 67.03 | 48.06 | 50.72 | 57.16 | 0.742 | 0.471 | 1.579 | 1.262 | 4.054 |
| Mistral-Medium-3 [12] | 65.95 | 48.84 | 52.97 | 57.70 | 0.923 | 0.729 | 1.566 | 1.339 | 4.557 |
| Qwen2.5-VL-32B [22] | 67.38 | 43.02 | 58.69 | 59.31 | 1.077 | 0.928 | **1.977** | 1.290 | 5.272 |
| Claude-4-Sonnet [3] | 71.51 | 46.51 | 54.60 | 60.23 | 0.742 | 0.633 | 1.778 | 1.376 | 4.529 |
| InternVL3-38B [5] | 69.71 | 57.36 | 52.97 | 61.00 | 0.964 | 0.824 | 1.860 | 1.317 | 4.965 |
| Gemini-2.5-Pro [7] | 75.13 | 55.02 | 50.54 | 61.88 | 0.878 | 0.891 | 1.688 | 1.561 | 5.018 |
| Qwen2.5-VL-72B [22] | 78.67 | 42.25 | 56.44 | 63.14 | 0.905 | 0.860 | 1.896 | 1.321 | 4.982 |
| Grok-4 [23] | 73.30 | 48.84 | 59.10 | 63.14 | 0.982 | 0.846 | 1.801 | 1.389 | 5.017 |
| GPT-4o [1] | 78.48 | 49.64 | 57.32 | 64.79 | 1.009 | 1.027 | 1.878 | 1.407 | 5.321 |
| Qwen2.5-VL-7B [22] | 57.89 | 48.45 | 54.40 | 54.71 | 0.715 | 0.670 | **1.855** | 1.127 | 4.367 |
| Qwen2.5-VL-7B-10k† | 86.56 | 68.22 | 62.78 | 74.02 | 0.950 | 0.937 | 1.839 | 1.520 | 5.246 |
| (Ours) | +28.67 | +19.77 | +8.38 | +19.31 | +0.24 | +0.27 | -0.016 | +0.39 | +0.88 |
| InternVL3-8B [5] | 72.04 | 47.67 | 52.97 | 60.08 | 0.928 | 0.878 | **1.858** | 1.317 | 4.983 |
| InternVL3-8B-10k† | **89.61** | **73.64** | **67.48** | **78.16** | **1.054** | **1.091** | 1.846 | **1.661** | **5.652** |
| (Ours) | +17.57 | +25.97 | +14.51 | +18.08 | +0.13 | +0.21 | -0.012 | +0.34 | +0.67 |

($t$=0) applying full human review and subsequent iterations applying entropy-guided selective review. All experiments are conducted on 8×A100 GPUs.

### 3.2 Main Results

Table 1 presents benchmark results on MedQbench, from which several findings emerge. **First**, our optimized InternVL3-8B-10k achieves 78.16% overall perception accuracy, surpassing GPT-4o (64.79%) by over 13 percentage points and narrowing the gap with human experts (82.50%) to only 4.34%. **Second**, both 7B/8B-scale optimized models consistently outperform all larger counterparts, including 32B, 72B, and closed-source models, demonstrating that targeted data curation can compensate for orders-of-magnitude differences in model scale. **Third**, medical-specialized MLLMs underperform general-purpose models of comparable size, suggesting that existing medical pretraining does not transfer effectively to the quality assessment domain. On the description task, both optimized models achieve the highest overall scores, with notable improvements in Completeness and Quality Accuracy.

### 3.3 Analysis

We examine MedQ-Engine along four dimensions: qualitative improvements, sample efficiency, component contributions, and annotation cost.

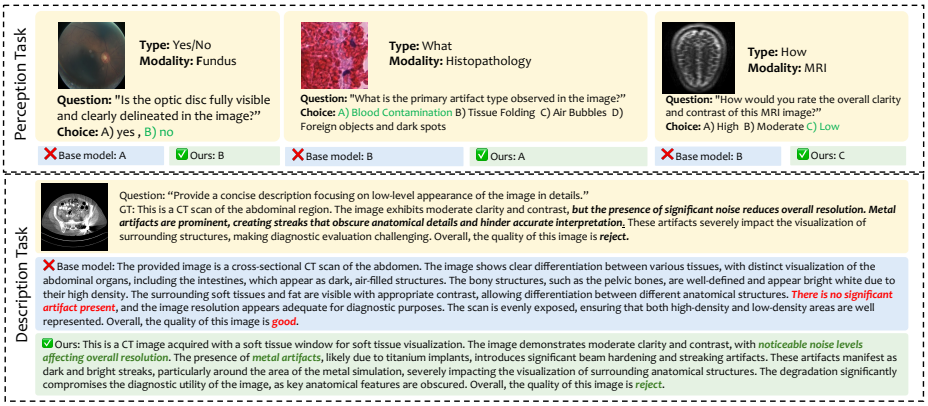


**Fig. 2. Qualitative comparison** between InternVL3-8B (blue) and InternVL3-8B-10k† (green).

**Qualitative Case Studies.** Figure 2 illustrates that, compared to base models which produce generic descriptions, MedQ-Engine generates anatomically specific assessments with clinical reasoning and actionable recommendations.

**Sample Efficiency.** We compare our failure-driven sampling against random sampling under identical annotation budgets. MedQ-Engine consistently outperforms random sampling across all data scales (Figure 3 (c)(f)). Notably, our method with only 10K samples surpasses random sampling at 40K samples and rapidly approaches the scaling upper bound, demonstrating more than 4× sample efficiency.

**Component Ablation.** Table 2 isolates each component's contribution (InternVL3-8B, 10K samples). Human-in-the-loop verification contributes the largest gain, particularly critical for description quality. Capability-driven QA generation outperforms simple seed-rule QA, and adaptive sampling proves essential for prioritizing weak dimensions. Removing all engine components (random baseline) drops performance by 9.46%, underscoring the cumulative value.

**Annotation Statistics.** Table 3 summarizes annotation statistics across iterations. The high GPT-4o Accept rate (63%) at cold-start reduces per-sample review time from 5.1 min (from scratch) to 0.5 min, a 10× speedup. The progressive strategy further reduces average human review to only 18% of samples in subsequent iterations, cutting overall expert cost by more than 5× compared to full human review.

### 3.4 Data Scaling Analysis

We investigate how model performance scales with training data size across different question types and degradation severity levels (Figure 3).

**Degradation Severity Analysis.** Base models perform well on artifact-free images but struggle with actual quality issues, where mild and severe degradation accuracy drops by over 30% (Figure 3 (a)(b)). MedQ-Engine substantially

**Table 2.** Ablation study using InternVL3-8B with 10K samples.

| Variant | Perc.↑ | Desc.↑ |
|---|---|---|
| Full MedQ-Engine | **78.16%** | **5.65** |
| w/o human-in-the-loop | 73.25% | 5.32 |
| w/o capability-driven QA | 75.43% | 5.38 |
| w/o adaptive sampling | 74.87% | 5.35 |
| w/o quality assurance | 76.51% | 5.49 |
| Random sampling (10K) | 68.70% | 5.28 |

**Table 3.** Human-in-the-loop annotation statistics. †Selective review triggered by Sec. 2.3.

| Metric | $t$=0 | $t$>0 | w/o pseudo-label |
|---|---|---|---|
| Annotation oracle | GPT-4o | Self + GPT-4o | — |
| Human review rate | 100% | 18%† | 100% |
| Accept rate | 63% | — | — |
| Edit rate | 29% | — | — |
| Reject rate | 8% | — | — |
| Avg. review time | 0.5 min | 0.5 min | 5.1 min |

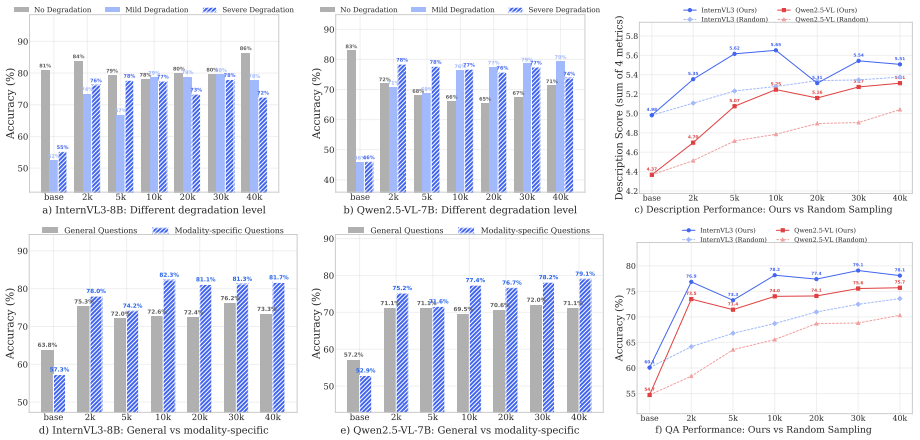


**Fig. 3. Data scaling analysis across two model architectures.** *Top row*: Performance on degradation severity levels and description scores as training data scales from 2K to 40K. *Bottom row*: General vs. modality-specific question performance, and failure-driven vs. random sampling comparison.

narrows this gap, with the largest improvements on mild and severe cases, validating that failure-driven data collection effectively targets the most challenging scenarios.

**General vs. Modality-specific Questions.** After training, modality-specific questions consistently outperform general questions across both models (Figure 3 (d)(e)). This suggests that domain-specific training data enables models to develop specialized knowledge about modality-characteristic degradations, while general quality reasoning remains more challenging.

## 4 Conclusion

We present MedQ-Engine, a closed-loop data engine that systematically improves MLLM capabilities for medical image quality assessment. By clustering failure cases into prototypes, using them as retrieval anchors for targeted data expansion, and employing progressive human-in-the-loop annotation with entropy-guided selective review, MedQ-Engine reduces human involvement to

18% of samples while achieving more than 4× sample efficiency over random sampling. With only 10K annotations, an 8B-parameter model surpasses GPT-4o by over 13% and approaches human expert performance across five medical imaging modalities. Beyond Med-IQA, the evaluate-explore-evolve paradigm offers a general blueprint for data-efficient MLLM adaptation in specialized domains where expert annotations are scarce and model weaknesses are non-uniform.

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

## References


1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Agarwal, S., Zhang, Z., Yuan, L., Han, J., Peng, H.: The unreasonable effectiveness of entropy minimization in llm reasoning. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2024)
3. Anthropic: System card: Claude opus 4 & claude sonnet 4. Tech. rep., Anthropic (2025), https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf
4. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
5. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
6. Chen, Z., Hu, B., Niu, C., Chen, T., Li, Y., Shan, H., Wang, G.: Iqagpt: computed tomography image quality assessment with vision-language and chatgpt models. Visual Computing for Industry, Biomedicine, and Art **7**(1), 20 (2024)
7. 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)
8. Fu, H., Wang, B., Shen, J., Cui, S., Xu, Y., Liu, J., Shao, L.: Evaluation of retinal image quality assessment networks in different color-spaces. In: International conference on medical image computing and computer-assisted intervention. pp. 48–56. Springer (2019)
9. Fuchs, M., Sivakumar, S.K.R., Schöber, M., Woltering, N., Eich, M.L., Schweizer, L., Mukhopadhyay, A.: Harp: Unsupervised histopathology artifact restoration. In: Medical Imaging with Deep Learning. pp. 465–479. PMLR (2024)
10. Haneda, E., Peters, N., Zhang, J., Karageorgos, G., Xia, W., Paganetti, H., Wang, G., Guo, Y., Ma, J., Park, H.S., et al.: Aapm ct metal artifact reduction grand challenge. Medical physics **52**(10), e70050 (2025)
11. 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)

12. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023), https://arxiv.org/abs/2310.06825
13. Liu, J., Gao, S., Li, Y., Liu, L., Gao, X., Xing, Z., Ning, J., Su, Y., Zhang, X.Y., He, J., et al.: Multi-modal mri translation via evidential regression and distribution calibration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 363–373. Springer (2025)
14. Liu, J., Ning, J., Ma, C., Qu, W., Shen, J., Luo, S., Wei, J., Ye, J., Li, P., Li, T., et al.: Medq-deg: A multidimensional benchmark for evaluating mllms across medical image quality degradations. arXiv preprint arXiv:2603.07769 (2026)
15. 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)
16. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing **21**(12), 4695–4708 (2012)
17. Mullappilly, S.S., Kurpath, M.I., Pieri, S., Alseiari, S.Y., Cholakkal, S., Aldahmani, K., Khan, F., Anwer, R., Khan, S., Baldwin, T., et al.: Bimedix2: Bio-medical expert lmm for diverse medical modalities. arXiv preprint arXiv:2412.07769 (2024)
18. Nárai, Á., Hermann, P., Auer, T., Kemenczky, P., Szalma, J., Homolya, I., Somogyi, E., Vakli, P., Weiss, B., Vidnyánszky, Z.: Movement-related artefacts (mr-art) dataset of matched motion-corrupted and clean structural mri brain scans. Scientific data **9**(1), 630 (2022)
19. Polat, G., Sen, D., Inci, A., Temizel, A.: Endoscopic artefact detection with ensemble of deep neural networks and false positive elimination. In: EndoCV@ ISBI. pp. 8–12 (2020)
20. Ramos, J., et al.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning. vol. 242, pp. 29–48. Citeseer (2003)
21. 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)
22. 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)
23. xAI: Grok 4 model card. Tech. rep., xAI (Aug 2025), https://data.x.ai/2025-08-20-grok-4-model-card.pdf, technical Report
24. 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)
25. Xun, S., Jiang, M., Huang, P., Sun, Y., Li, D., Luo, Y., Zhang, H., Zhang, Z., Liu, X., Wu, M., et al.: Chest ct-iqa: A multi-task model for chest ct image quality assessment and classification. Displays **84**, 102785 (2024)
26. Xun, S., Sun, Y., Chen, J., Yu, Z., Tong, T., Liu, X., Wu, M., Tan, T.: Mediqa: A scalable foundation model for prompt-driven medical image quality assessment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 339–349. Springer (2025)
27. 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 bench-

mark towards general medical ai. Advances in Neural Information Processing Systems **37**, 94327–94427 (2024)
28. You, Z., Gu, J., Li, Z., Cai, X., Zhu, K., Dong, C., Xue, T.: Depicting beyond scores: Advancing image quality assessment through multi-modal language models. arXiv preprint arXiv:2312.08962 (2024)
29. Zauner, C.: Implementation and benchmarking of perceptual image hash functions. In: Master's thesis, Upper Austria University of Applied Sciences (2010)
30. Zbontar, J., Knoll, F., Sriram, A., Murrell, T., Huang, Z., Muckley, M.J., Defazio, A., Stern, R., Johnson, P., Bruno, M., et al.: fastmri: An open dataset and benchmarks for accelerated mri. arXiv preprint arXiv:1811.08839 (2018)
31. 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)
32. Zhang, Z., Wu, H., Zhang, E., Zhai, G., Lin, W.: Q-bench++: A benchmark for multi-modal foundation models on low-level vision from single images to pairs. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 10404–10418 (2024)