MedIQA: A Scalable Foundation Model for Prompt-Driven Medical Image Quality Assessment

Siyi Xun[1], Yue Sun[1], Jingkun Chen[2], Zitong Yu[3], Tong Tong[4], Xiaohong Liu[5](✉), Mingxiang Wu[6], and Tao Tan[1](✉)

[1] Faculty of Applied Sciences, Macao Polytechnic University, Macao, China
[2] Department of Engineer Science, University of Oxford, Oxford, UK
[3] Great Bay University, Dongguan, China
[4] College of Physics and Information Engineering, Fuzhou University, Fuzhou, China
[5] Shanghai Jiao Tong University, Shanghai, China
[6] Department of Radiology, Shenzhen People's Hospital, Shenzhen, China
Corresponding author(✉) : Xiaohong Liu (xiaohongliu@sjtu.edu.cn), Tao Tan(taotan@mpu.edu.mo)

Abstract. Rapid advances in medical imaging technology underscore the critical need for precise and automated image quality assessment (IQA) to ensure diagnostic accuracy. Existing medical IQA methods, however, struggle to generalize across diverse modalities and clinical scenarios. In response, we introduce MedIQA, the first comprehensive foundation model for medical IQA, designed to handle variability in image dimensions, modalities, anatomical regions, and types. We developed a large-scale multi-modality dataset with plentiful manually annotated quality scores to support this. Our model integrates a salient slice assessment module to focus on diagnostically relevant regions feature retrieval and employs an automatic prompt strategy that aligns upstream physical parameter pre-training with downstream expert annotation fine-tuning. Extensive experiments demonstrate that MedIQA significantly outperforms baselines in multiple downstream tasks, establishing a scalable framework for medical IQA and advancing diagnostic workflows and clinical decision-making. Our code is available at https://github.com/siyi-xun/MedIQA.

Keywords: Medical image quality assessment · Foundation model · Prompt strategy · Upstream and downstream validation.

1 Introduction

Medical image quality assessment (IQA) is critical for reliable diagnosis. However, the heterogeneity of modalities, anatomical regions, and clinical scenarios poses significant challenges. Traditional IQA approaches, often based on handcrafted features or domain-specific models, struggle with generalization in various scenarios [1, 2]. As medical imaging technologies become increasingly complex and data volumes surge, these limitations are further exacerbated.

Recent deep learning advances—especially foundation models pretrained on large-scale data—offer a powerful means to overcome these challenges. Foundation models, pretrained on large datasets, excel at learning universal representations that can be fine-tuned for specific tasks [3–5]. Their ability to generalize across domains and adapt to new tasks with minimal supervision makes them highly suitable for medical IQA. These AI models have demonstrated superior performance in tasks such as segmentation, denoising, artifact detection, and quality scoring [6–14].

Despite their promise, foundation models for medical IQA still face significant challenges: there is a scarcity of high-quality annotated datasets, a need for dynamic adaptation to varying imaging conditions, and difficulty integrating domain-specific knowledge into model architectures [15]. Moreover, the "black box" nature of these models limits their explainability and clinical acceptance.

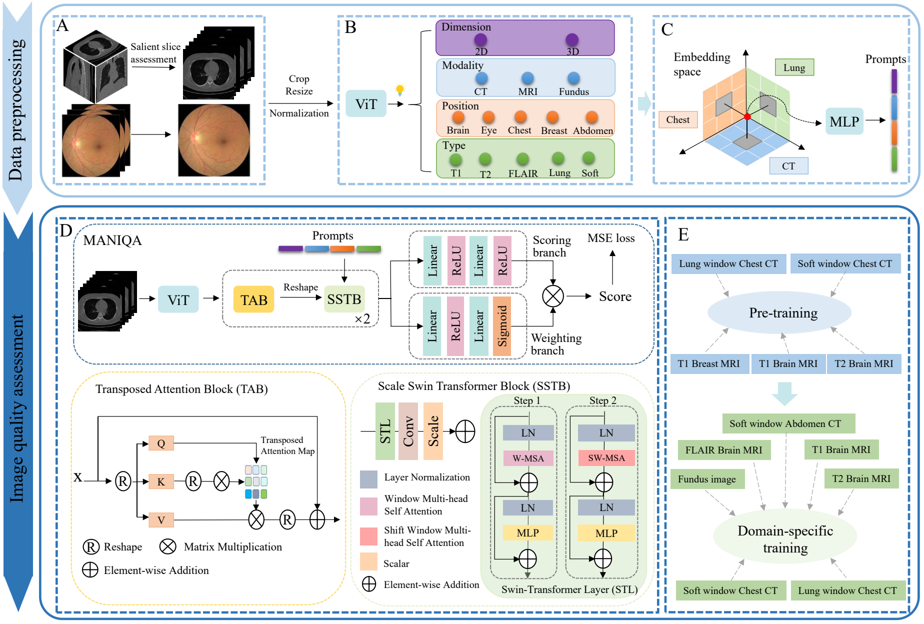

Fig. 1. Overview of the MedIQA workflow. (A) Salient slice assessment. (B) and (C) Prompts generation and encoding. (D) Backbone structure. (E) Training procedure.

Based on these insights, we introduce MedIQA, a prompt-driven and scalable foundation model for medical IQA. Our contributions are as follows: *(1)* We construct a large-scale MedIQA dataset of approximately 15k 2D and 3D radiographic scans, including CT, MRI, and other modalities, with high-quality expert annotations across various anatomical regions. *(2)* We propose a salient slice assessment module to reduce redundant data and suppress background noise, allowing the model to focus on diagnostically relevant regions feature retrieval and enhancing both generalization and efficiency. *(3)* We implement a two-stage training strategy: an upstream pre-training stage using physical parameters (e.g. dose, magnetic field strength) and a downstream fine-tuning stage with expert annotations. This approach creates an explicit link between objective physical characteristics and subjective quality assessment, thereby improving model explainability. *(4)* We integrate domain-specific imaging information (dimension, modality, position, and type) into an automated prompt strategy to ensure the model dynamically adapts to cross-modality multi-organ IQA tasks.

2 Methodology

2.1 Dataset

Existing medical IQA datasets are limited in scale, reliability, and diversity. To overcome these issues, we designed, proposed, and constructed a MedIQA dataset by integrating a large-scale annotated CT dataset and existing medical image datasets. The dataset is divided into pre-training and domain-specific dataset. Fig.2 shows dataset examples, quantities, and distributions. The pretraining dataset contains in-house chest CT and public MRI brain (T1, T2)/breast (T1) datasets (2,500 3D volumes)[16, 17]. Pre-training labels were generated by extracting dose (mAs) and magnetic field strength (Tesla) from image physics parameters with label values positively correlated with image quality [18, 19]. The domain-specific dataset contains five subsets (12,545 cases). All domain-specific images were annotated by radiologists or trained professionals to derive a Mean Opinion Score (MOS): *(1)* Chest-CTIQA (3D). Our first large-scale chest CT IQA dataset (10 readers per volume, 1-5 scale based on sharpness, noise, and contrast). *(2)* LDCTIQAC2023 (2D)[20]. Public abdominal CT IQA dataset (5 readers per image, 0-4 scale based on clinical usability and anatomical clarity). *(3)* ADNI MRI (2D)[21]. To ensure sufficient data volume and balance, different 2D slices were selected from 3D volumes to construct three sub-datasets (T1, T2, FLAIR; 1-4 scale based on artifacts, completeness, and overall image quality). *(4)* Kaggle DR (2D)[22]. High-resolution retinal images under diverse conditions (high quality labeled 1, low quality labeled 0). All datasets were pre-processed for consistency, and normalized for training. MedIQA is the first multimodal IQA dataset addressing diverse quality assessment needs.

2.2 Model

The overall architecture of the model is designed to achieve comprehensive medical IQA, as illustrated in Fig.1. First, the image input module receives and preprocesses input images. For 3D volume, the salient slice assessment module extracts seven salient slices from the sequence, focusing on diagnostic regions feature retrieval. Next, the pre-trained Vision Transformer (ViT) [23] classifier generates encoded prompts for dynamic adaptation to different imaging conditions while matching upstream and downstream tasks. The main framework of

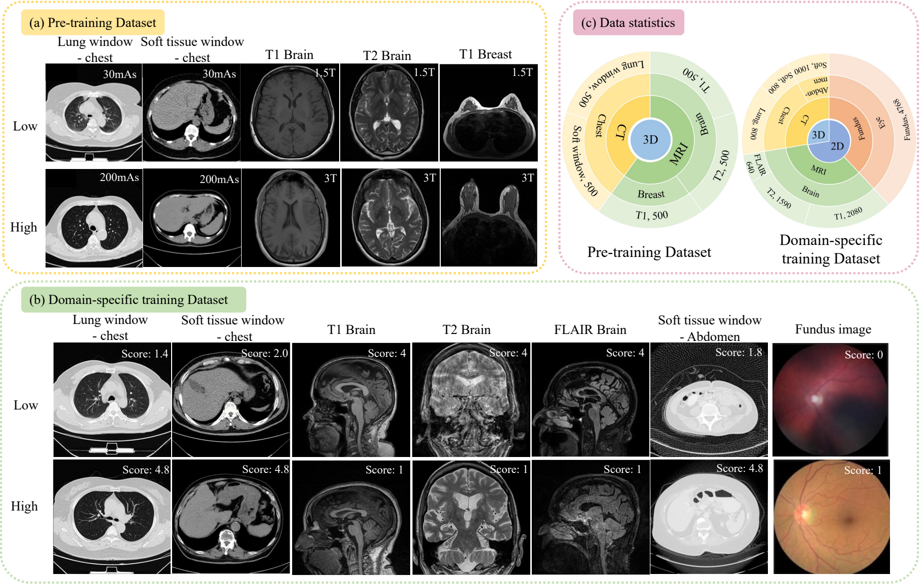

Fig. 2. Examples, quantities, and distributions of the MedIQA-dataset.

the model uses MANIQA [24] as the backbone network for feature extraction and quality assessment. Finally, the model outputs the overall image quality score.

Salient slice assessment for feature retrieval. Due to the minimal quality differences between adjacent slices, continuous sampling often results in redundant data. Therefore, in order to focus on the region of interest while reducing computational complexity, we select seven salient slices from each volume for feature retrieval (Fig.1.A). Specifically, the 3D medical volume $V \in \mathbb{R}^{H \times W \times D}$ is divided into seven regions $R_i = S[v_{i-1} : v_i], i = 1, 2, ..., 7$ along the Z axis by removing the irrelevant slice that does not contain any region of interest. The middle slice $s_i = R_i \left[\left\lfloor \frac{|R_i|}{2} \right\rfloor \right]$ is selected from each region. Compared to existing slice selection strategies, anatomical structure-based equidistant partitioning and geometric centroid prior constraint within regions ensure uniform sampling constraints globally while covering the local critical slice that contains the diagnostic region. All images were min-max normalized to align the intensity distribution, and image sizes were normalized to 224*224 for consistency.

Upstream physical parameters-driven foundation model learning. Physical parameters $p \in \mathbb{R}^k$ such as dose or magnetic field strength are pretrained to help the model learn the effects of these parameters on underlying image features $f \in \mathbb{R}^m$ (noise, contrast, resolution). The objective function is defined by $\underset{\theta}{min}\mathcal{L}_{pre} = \mathbb{E}_{(\tau,p)} \left[\|g_\varphi(E_\theta(\tau)) - p\|_2^2 \right]$, where the encoder $E_\theta : \tau \rightarrow f$, the parameter is θ, $g_\varphi : f \rightarrow p$ is the parameter prediction head, and the explicit association between the feature f and p is constrained by the mean square error (MSE). The explicit association formed by the "parameter → feature" mapping established in the pre-training stage provides a physical basis for the subsequent quality assessment. In the fine-tuning stage, the intermediate features generated by pre-training and strongly related to physical parameters will be reused to make the model decision-making process more transparent.

Prompt-based upstream and downstream matching. Prompts include: dimensional (p_{dim}), modality (p_{mod}), region (p_{reg}), and type (p_{type}). Dimensional prompts are derived from input dimensions, while others are generated using a pre-trained ViT for classification. A 12-layer, 12-head ViT with a size of 3072 is employed as the decoder (Fig.1.B). For integration, four one-hot encoded prompts are concatenated and projected into Swin Transformer Layers (STL) via a fully connected (FC) layer (Fig.1.C). Prompts are added to the feature vector at each decoder layer, enabling dynamic adaptation. Each layer has an independent FC layer, allowing task updates without new branches. The implementation is, $y = x + FC_i(concat(p_{dim}, p_{mod}, p_{reg}, p_{type})), i = 1, 2$. Prompts contribute the same amount in each STL layer. The prompt strategy matches physics-parameterized foundation models with expert-annotated domain knowledge for dual supervision of physical and expert-guided quality characteristics.

Quality assessment. Given the lack of high-quality reference images in medical IQA, no-reference IQA (NR-IQA) has become the optimal choice. Therefore, we adopt MANIQA, a state-of-the-art NR-IQA model, as our backbone (Fig.1.D). MANIQA uses ViT extracts features $F \in \mathbb{R}^{b \times \sum_i C_i \times H_i W_i}$ sent to Transposed Attention Block (TAB) and Scale Swin Transformer Block (SSTB) to implement multi-dimensional attention mechanisms in both channel and spatial. The final score is given by the dual-branch prediction module of the scoring (s) branch and the weighting (w) branch. For 2D images, predict quality score is computed as $q = \frac{\sum_{0<i<N} w_i \times s_i}{\sum_{0<i<N} w_i}$, where N denotes the number of patches for one image. For 3D volumes, features are extracted from seven salient slices, slice-level scores $\mathbf{q} = [q_1, q_2, ..., q_7]$ are obtained by dual-branch structure, corresponding weights $\bar{\mathbf{w}} = [\bar{w}_1, \bar{w}_2, ..., \bar{w}_7]$ are generated by linear layer, and the final image quality score Q is calculated as $Q = \sum_{i=1}^{7} \bar{w}_i q_i$. These weights dynamically adjust based on the importance of each slice, enhancing the model's precision and sensitivity to local quality variations.

2.3 Model Training Procedure

Training process: Two-stage training strategy employs physics-parameter pre-training followed by domain-specific fine-tuning (Fig.1.E). The upstream pre-training task generates predicted distributions for each sample based on physical features, while the downstream fine-tuning task load pre-trained weights and learn expert knowledge from domain-specific data to rapidly minimize discrepancies between predictions and label distributions. Prompts match these two stages, ensuring explainable IQA with cross-domain transferability. Both stages share consistent training settings, differing only in the training data and labels.

Loss Function: The MSE loss function is used to measure the difference between predicted quality scores and true annotations. MSE ensures minimized prediction errors and alignment with expert annotations, while its smoothness promotes stable convergence during optimization. The MSE loss is calculated as $L_{MSE} = \frac{1}{n}\sum(x_i - y_i)^2$, where n is the number of samples, x_i and y_i is the true value and the predicted value of the model.

3 Experiments

3.1 Implementation Details and Metrics

Our experiments are implemented on an Intel Xeon W2245 CPU @ 3.90GHz and an NVIDIA RTX A6000 GPU (48 GB) with Python 3.9 and PyTorch 2.0.0. Hyperparameters included a learning rate of 1e-5, a batch size of 1, and 50 training epochs, optimized using the Adam optimizer. For both phases, the momentum was set to 0.9, weight decay to 0.01, and the dataset was split into training, validation, and test sets in an 8:1:1 ratio across diverse data distributions.

The performance of IQA tasks was evaluated by Spearman rank-order correlation coefficient (SRCC), Pearson linear correlation coefficient (PLCC) and Root mean square error (RMSE). SRCC and PLCC measure the monotonicity and linear correlation of the model, while RMSE assesses the consistency of the model's predictions.

3.2 Experimental Results and Analysis

Classification experiments. To enable automatic prompt generation, we trained VGG and ViT models using additional classification data and evaluated them on the MedIQA dataset. Experimental results show that VGG achieved an average test accuracy of 0.9445, while ViT achieved 0.9969. Meanwhile, ViT's self-attention-based design allowed for higher performance with fewer parameters (86M *VS* 138M). Therefore, we selected ViT for prompt generation.

Upstream foundation experiments. The proposed model was trained and tested on pre-training dataset as well as five sub-benchmarks (chest lung window benchmark, chest soft tissue window benchmark, brain T1 benchmark, brain T2 benchmark and breast T1 benchmark). Table 1 shows the results of our model and other methods. The experimental results show that the proposed model can learn quality features of different modality images, accurately predict quality-related parameters on different benchmarks, and provide a good basis for training downstream tasks.

Downstream tasks. In domain-specific task experiments, we evaluated the model's performance on six benchmarks: 3D chest CT benchmark, 2D brain T1, T2 and FLAIR MRI benchmarks, 2D fundus image benchmark, and 2D synthetic abdominal CT benchmark. Results (Table 1) show that the model performed relatively poorly on 3D chest CT data due to the complexity of high-dimensional data. In contrast, synthetic abdominal CT data achieved the best

Table 1. Analyse the performance on the upstream and downstream tasks. Our-s and Our-e indicate the single baseline and the ensemble model. Best in red and second in blue. $\mathcal{S}$, SRCC; $\mathcal{P}$, PLCC; $\mathcal{R}$, RMSE.

	Upstream – pretraining dataset																	
Backbone	Test			Lung window			Soft window			Brain T1			Brain T2			Breast T1		
	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$	$\mathcal{R}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$	$\mathcal{R}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$	$\mathcal{R}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$	$\mathcal{R}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$	$\mathcal{R}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$	$\mathcal{R}\downarrow$
VGG	0.6623	0.6744	0.2838	0.6857	0.6972	0.2141	0.7054	0.7247	0.2149	0.6838	0.6770	0.2679	0.5914	0.6142	0.4153	0.6452	0.6589	0.3062
Resnet	0.5413	0.5506	0.2994	0.6283	0.6438	0.2112	0.2988	0.2848	0.2569	0.5945	0.6129	0.3112	0.5559	0.5901	0.4083	0.6290	0.6187	0.3121
Swin-transformer	0.6626	0.6758	0.2842	0.7034	0.7221	0.2102	0.7018	0.7021	0.2175	0.6619	0.6771	0.2771	0.5566	0.5774	0.4137	0.6886	0.6792	0.3086
ViT	0.5852	0.5858	0.3044	0.7118	0.7021	0.2175	0.6591	0.6498	0.2339	0.5396	0.5515	0.3313	0.4193	0.4378	0.4284	0.5993	0.5887	0.3211
DeepViT	0.5402	0.5432	0.3108	0.5490	0.5644	0.2396	0.6203	0.6054	0.2376	0.5213	0.5339	0.3338	0.4127	0.4422	0.4327	0.5968	0.5857	0.3217
CNNIQAnet	0.5486	0.5266	0.3202	0.5549	0.5466	0.2471	0.6096	0.5876	0.2404	0.5382	0.5032	0.3423	0.4957	0.5098	0.4499	0.5434	0.4851	0.3404
WaDIQaM	0.6478	0.6298	0.2998	0.7152	0.7225	0.2103	0.7091	0.6891	0.2196	0.6154	0.5930	0.3192	0.5873	0.5546	0.4135	0.6094	0.5657	0.3173
Ours	0.7770	0.8110	0.2609	0.7316	0.7363	0.2029	0.7114	0.7674	0.2031	0.7891	0.8739	0.2443	0.6828	0.6689	0.3843	0.8148	0.9246	0.1957
	Downstream – domain-specific datasets																	
Backbone	Chest-CTIQA			Brain-T1			Brain-T2			Brain-FLAIR			Fundus			LDCTIQAC2023		
	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$	$\mathcal{R}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$	$\mathcal{R}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$	$\mathcal{R}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$	$\mathcal{R}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$	$\mathcal{R}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$	$\mathcal{R}\downarrow$
VGG	0.4166	0.3686	0.2670	0.6172	0.6127	0.2103	0.5313	0.5394	0.3419	0.6283	0.6214	0.3345	0.7368	0.6923	0.2899	0.8350	0.8298	0.1609
Resnet	0.3902	0.3976	0.2026	0.5719	0.5284	0.2571	0.4919	0.5039	0.3467	0.5202	0.5214	0.3556	0.7119	0.6897	0.3772	0.7972	0.7775	0.1599
Swin-transformer	0.4741	0.5004	0.2232	0.8026	0.8694	0.2137	0.6968	0.7012	0.3749	0.6968	0.7012	0.2849	0.8118	0.7755	0.3068	0.9013	0.9009	0.0922
ViT	0.4044	0.4225	0.2040	0.6373	0.6797	0.2589	0.6015	0.5631	0.3051	0.5659	0.5664	0.3364	0.7156	0.6717	0.3777	0.8127	0.7432	0.1124
DeepViT	0.4592	0.4697	0.2964	0.7988	0.8646	0.1822	0.6585	0.6582	0.2775	0.6131	0.5879	0.3203	0.7368	0.6923	0.3899	0.8342	0.8336	0.1257
CNNIQAnet	0.4641	0.4990	0.2530	0.6547	0.6567	0.2161	0.6517	0.6369	0.3027	0.6323	0.6286	0.3429	0.7842	0.7734	0.3002	0.8661	0.8866	0.1036
WaDIQaM	0.4835	0.4810	0.1945	0.7119	0.6897	0.2172	0.7056	0.6717	0.2777	0.6585	0.6582	0.3275	0.7672	0.7775	0.2519	0.8798	0.8583	0.1213
Ours-s	0.4875	0.5255	0.1661	0.8659	0.8905	0.1507	0.7058	0.7035	0.2731	0.7177	0.7184	0.2798	0.8329	0.9144	0.2054	0.9761	0.9759	0.0631
Ours-e	0.7070	0.7455	0.1276	0.8681	0.8985	0.1196	0.8861	0.8912	0.1696	0.7654	0.7578	0.2657	0.8504	0.9320	0.1837	0.9764	0.9762	0.0618

results, as the significant quality variations enabled the model to learn quality features more effectively. T1 and T2 MRI data showed stable performance, while FLAIR MRI data underperformed due to a lack of pre-training data information. Fundus image assessment results were moderate, likely due to resolution and lighting limitations. Future work should focus on optimizing 3D data handling and leveraging synthetic data for pre-training.

Performance Comparisons. To validate the effectiveness of MedIQA, we evaluated it against other models, including VGG [25], ResNet [26], ViT [23], Swin-Transformer [27], DeepViT [28], CNNIQAnet [29], and WaDIQaM [30]. Results (Table 1) demonstrate that our model outperformed all others in image quality assessment tasks. In upstream tasks, the average result of our model (0.7511, 0.7970, 0.2485) is significantly improved (+0.2027, +0.2706, -0.0748) compared with the average result of CNNIQAnet (0.5484, 0.5264, 0.3233). For downstream tasks, the average results of our model (0.8422, 0.8668, 0.1546) improved by 7.79% and 7.88% over the average results of baseline (0.7643, 0.7880, 0.1897). It is also significantly higher than the average result of ResNet (+0.2617, +0.2971, -0.1285). The experimental results show that compared with classical models (VGG and ResNet), our model's multi-dimensional attention mechanisms and salient slice assessment modules better retrieve local and global features, overcoming traditional CNN limitations and significantly improving accuracy. Additionally, our model outperformed other transformer-based models by enhancing generalization through pretraining and prompt strategies, while capturing finer details via salient slice assessment module. It also surpassed specialized IQA models like CNNIQAnet and WaDIQaM. These results validate the effectiveness of our novel designs, offering an efficient solution for IQA tasks.

Ablation Study. We evaluated the impact of pretraining (PT), prompt strategies (PM), and salient slice assessment (SS) performance on domain-specific

Table 2. Ablation study on downstream tasks. Only 3D images use the salient slice assessment module. Best in red and second in blue.

Model	Module			Chest-CTIQA(3D)			Brain-T1(2D)			Brain-T2(2D)			Brain-FLAIR(2D)			Fundus(2D)			LDCTIQAC2023(2D)		
	PT	PM	SS	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$	$\mathcal{R}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$	$\mathcal{R}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$	$\mathcal{R}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$	$\mathcal{R}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$	$\mathcal{R}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$	$\mathcal{R}\downarrow$
1	✗	✗	✗	0.4875	0.5255	0.1661	0.8659	0.8905	0.1507	0.7058	0.7035	0.2731	0.7177	0.7184	0.2798	0.8329	0.9144	0.2054	0.9761	0.9759	0.0631
2	✓	✗	✗	0.5347	0.5541	0.1608	0.8706	0.8978	0.1261	0.8549	0.8559	0.1936	0.8294	0.8247	0.2203	0.8345	0.9438	0.1668	0.9742	0.9757	0.0589
3	✓	✓	✗	0.5464	0.5927	0.1561	0.8681	0.8985	0.1196	0.8861	0.8912	0.1696	0.7654	0.7578	0.2657	0.8504	0.9320	0.1837	0.9764	0.9762	0.0618
4	✓	✓	✓	0.7070	0.7455	0.1276	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/

downstream tasks. Results (Table 2) show that the modules' effects vary by data type. For 3D chest CT data, all modules improved performance, with the salient slice assessment module significantly enhancing 3D feature learning (+0.2195, +0.2200, -0.0385). For 2D T2 data, all modules had positive effects (+0.1803, +0.1877, -0.1035), as prompt strategy and pretraining helped capture quality features. However, for 2D FLAIR data, the prompt strategy caused performance degradation, likely due to mismatches between FLAIR's unique quality characteristics and the prompt strategy's design, leading the model to learn irrelevant features. For 2D T1, 2D fundus, and 2D synthetic CT data, additional modules had minimal impact, as these datasets' simplicity or consistency already enabled strong performance. Future work should optimize prompt designs for different modalities to improve generalization and performance.

4 Discussion

By integrating large-scale cross-modality MedIQA dataset, prompt strategy and salient slice assessment module and upstream and downstream matching, MedIQA captures both global and local quality features, ensuring robust assessments. Compared to other methods, our model shows significant improvements and provides a scalable framework for medical IQA. MedIQA's design enables its use as a pre-/post-processing module in quality control systems to flag suboptimal images. In addition, preliminary experiments revealed that CT image quality affects AI detection of lung nodules. Thus, the relationship between medical image quality and disease detection rates will be a focus of our future research.

Despite promising results, the study has limitations. First, pretraining data may not fully capture variability across modalities or clinical scenarios, necessitating more annotated data or unsupervised learning methods. Second, the prompt strategy, while effective, relies heavily on high-quality prompt design, requiring further optimization for diverse tasks. Third, salient slice assessment module may miss subtle quality changes in long sequences, potentially underperforming in extreme conditions (e.g., excessive noise or missing images). Future research will focus on: (1) Expanding datasets to include diverse modalities and scenarios for better generalizability; (2) Developing explainable architectures to build clinician trust; and (3) Integrating the model into clinical workflows and validating its impact on diagnostic accuracy and efficiency.

5 Conclusion

In this paper, we proposed a scalable foundation model for medical IQA. First, we constructed the MedIQA dataset, a large-scale, multi-modal, and multi-organ dataset with DICOM tags and plentiful manually annotated quality scores, providing a robust foundation for model learning. Second, we designed a salient slice assessment module to focus on diagnostically relevant regions and enhance efficiency, and implement a two-stage training strategy to bridge physical parameters with expert annotations, improving explainability. We also designed domain-specific automated prompts for cross-modality multi-organ IQA tasks. Experimental results demonstrate superior performance across multiple IQA benchmarks and solved the limitations of traditional methods. Our approach provides a scalable solution for clinical applications. Future research will focus on mitigating data scarcity, optimizing prompt strategies, and refining salient slice assessment to further enhance the model's practicality and applicability.

Acknowledgments. This work is supported by Science and Technology Development Fund of Macao (0021/2022/AGJ) and the National Natural Science Foundation of China under Grant 62301310.

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

References

1. Chow, L.S., Paramesran, R.: Review of medical image quality assessment. Biomedical signal processing and control **27**, 145–154 (2016)
2. Zhang, Z., Zhou, Y., Li, C., Zhao, B., Liu, X., Zhai, G.: Quality assessment in the era of large models: A survey. ACM Transactions on Multimedia Computing, Communications and Applications (2024)
3. Zhang, S., Metaxas, D.: On the challenges and perspectives of foundation models for medical image analysis. Medical image analysis **91**, 102996 (2024)
4. He, Y., Huang, F., Jiang, X., Nie, Y., Wang, M., Wang, J., Chen, H.: Foundation model for advancing healthcare: challenges, opportunities and future directions. IEEE Reviews in Biomedical Engineering (2024)
5. Rasoolzadeh, N., Zhang, T., Gao, Y., van Dijk, J.M., Yang, Q., Tan, T., Mann, R.M.: Multimodal breast mri language-image pretraining (mlip): An exploration of a breast mri foundation model. In: Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care. pp. 42–53. Springer (2024)
6. El-Shafai, W., El-Nabi, S.A., Ali, A.M., El-Rabaie, E.S.M., Abd El-Samie, F.E.: Traditional and deep-learning-based denoising methods for medical images. Multimedia Tools and Applications **83**(17), 52061–52088 (2024)
7. Allman, D., Reiter, A., Bell, M.A.L.: Photoacoustic source detection and reflection artifact removal enabled by deep learning. IEEE transactions on medical imaging **37**(6), 1464–1477 (2018)
8. Wu, H., Wu, X., Li, C., Zhang, Z., Chen, C., Liu, X., Zhai, G., Lin, W.: T2i-scorer: Quantitative evaluation on text-to-image generation via fine-tuned large multi-modal models. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 3676–3685 (2024)

9. Lin, Z., Li, S., Wang, S., Gao, Z., Sun, Y., Lam, C.T., Hu, X., Yang, X., Ni, D., Tan, T.: An orchestration learning framework for ultrasound imaging: Prompt-guided hyper-perception and attention-matching downstream synchronization. Medical Image Analysis p. 103639 (2025)
10. 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)
11. Chen, J., Huang, W., Zhang, J., Debattista, K., Han, J.: Addressing inconsistent labeling with cross image matching for scribble-based medical image segmentation. IEEE Transactions on Image Processing (2025)
12. Chen, J., Zhang, J., Debattista, K., Han, J.: Semi-supervised unpaired medical image segmentation through task-affinity consistency. IEEE Transactions on Medical Imaging **42**(3), 594–605 (2023)
13. 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: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22. pp. 48–56. Springer (2019)
14. Yang, H., Coyner, A.S., Guretno, F., Mien, I.H., Foo, C.S., Campbell, J.P., Ostmo, S., Chiang, M.F., Krishnaswamy, P.: A minimally supervised approach for medical image quality assessment in domain shift settings. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1286–1290. IEEE (2022)
15. Shi, C., Rezai, R., Yang, J., Dou, Q., Li, X.: A survey on trustworthiness in foundation models for medical image analysis. arXiv preprint arXiv:2407.15851 (2024)
16. Knoll, F., Zbontar, J., Sriram, A., Muckley, M.J., Bruno, M., Defazio, A., Parente, M., Geras, K.J., Katsnelson, J., Chandarana, H., et al.: fastmri: A publicly available raw k-space and dicom dataset of knee images for accelerated mr image reconstruction using machine learning. Radiology: Artificial Intelligence **2**(1), e190007 (2020)
17. Chitalia, R., Pati, S., Bhalerao, M., Thakur, S.P., Jahani, N., Belenky, V., McDonald, E.S., Gibbs, J., Newitt, D.C., Hylton, N.M., et al.: Expert tumor annotations and radiomics for locally advanced breast cancer in dce-mri for acrin 6657/i-spy1. Scientific data **9**(1), 440 (2022)
18. Payne, J.T.: Ct radiation dose and image quality. Radiologic Clinics **43**(6), 953–962 (2005)
19. Wu, C.W., Chuang, K.H., Wai, Y.Y., Wan, Y.L., Chen, J.H., Liu, H.L.: Vascular space occupancy-dependent functional mri by tissue suppression. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine **28**(1), 219–226 (2008)
20. Lee, W., Wagner, F., Maier, A., Wang, A., Jongduk, B., Scott, H., Choi, J.H.: Low-dose computed tomography perceptual image quality assessment grand challenge dataset. In: Medical Image Computing and Computer Assisted Intervention (2023)
21. Wyman, B.T., Harvey, D.J., Crawford, K., Bernstein, M.A., Carmichael, O., Cole, P.E., Crane, P.K., DeCarli, C., Fox, N.C., Gunter, J.L., et al.: Standardization of analysis sets for reporting results from adni mri data. Alzheimer's & Dementia **9**(3), 332–337 (2013)
22. Dugas, E., Jared, J., Cukierski, W.: Diabetic retinopathy detection (2015). URL https://kaggle. com/competitions/diabetic-retinopathy-detection **7**

23. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
24. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1191–1200 (2022)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
27. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
28. Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Jiang, Z., Hou, Q., Feng, J.: Deepvit: Towards deeper vision transformer. arXiv preprint arXiv:2103.11886 (2021)
29. Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no-reference image quality assessment. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1733–1740 (2014)
30. Bosse, S., Maniry, D., Müller, K.R., Wiegand, T., Samek, W.: Deep neural networks for no-reference and full-reference image quality assessment. IEEE Transactions on image processing **27**(1), 206–219 (2017)