


ORIGINAL ARTICLE

Open Access



IQAGPT: computed tomography image quality assessment with vision-language and ChatGPT models

Zhihao Chen^{1†}, Bin Hu^{2†}, Chuang Niu^{3†}, Tao Chen¹, Yuxin Li^{2*}, Hongming Shan^{1,4,5*}  and Ge Wang^{3*}

Abstract

Large language models (LLMs), such as ChatGPT, have demonstrated impressive capabilities in various tasks and attracted increasing interest as a natural language interface across many domains. Recently, large vision-language models (VLMs) that learn rich vision–language correlation from image–text pairs, like BLIP-2 and GPT-4, have been intensively investigated. However, despite these developments, the application of LLMs and VLMs in image quality assessment (IQA), particularly in medical imaging, remains unexplored. This is valuable for objective performance evaluation and potential supplement or even replacement of radiologists' opinions. To this end, this study introduces IQAGPT, an innovative computed tomography (CT) IQA system that integrates image-quality captioning VLM with ChatGPT to generate quality scores and textual reports. First, a CT-IQA dataset comprising 1,000 CT slices with diverse quality levels is professionally annotated and compiled for training and evaluation. To better leverage the capabilities of LLMs, the annotated quality scores are converted into semantically rich text descriptions using a prompt template. Second, the image-quality captioning VLM is fine-tuned on the CT-IQA dataset to generate quality descriptions. The captioning model fuses image and text features through cross-modal attention. Third, based on the quality descriptions, users verbally request ChatGPT to rate image-quality scores or produce radiological quality reports. Results demonstrate the feasibility of assessing image quality using LLMs. The proposed IQAGPT outperformed GPT-4 and CLIP-IQA, as well as multitask classification and regression models that solely rely on images.

Keywords Deep learning, Medical imaging, Image captioning, Multimodality, Large language model, Vision-language model, GPT-4, Subjective evaluation

[†]Zhihao Chen, Bin Hu and Chuang Niu contributed equally to this work.

*Correspondence:

Yuxin Li

liyuxin@fudan.edu.cn

Hongming Shan

hmshan@fudan.edu.cn

Ge Wang

wangg6@rpi.edu

¹ Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China

² Department of Radiology, Huashan Hospital, Fudan University, Shanghai 200040, China

³ Biomedical Imaging Center, Center for Biotechnology and Interdisciplinary Studies, Department of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, US

⁴ MOE Frontiers Center for Brain Science, Fudan University, Shanghai 200032, China

⁵ Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (Ministry of Education), Fudan University, Shanghai 200433, China

Introduction

In recent years, there have been many advances in the field of large language models (LLMs). LLMs such as PaLM [1], LLaMA [2], and GPTs [3–5] have shown excellent results in natural language processing, including language translation, question answering, and text generation. The most remarkable breakthrough is ChatGPT, which was built upon InstructGPT [6] using labeler-written prompts and reinforcement learning from human feedback [7]. However, LLMs such as ChatGPT are unable to process visual information as they are trained only on textual data. To address this gap, vision-language models (VLMs) [8–14], which synergistically combine the capabilities of LLMs with visual processing, were proposed to capture rich vision-language correspondence. These perform well in various multimodal tasks such as report generation, diagnosis, and vision question answering. In this context, OpenAI launched its new large VLM called GPT-4 [15], with amazing performance on multimodal tasks during dialogues. In addition, MiniGPT-4 [16] integrates an advanced LLM, Vicuna [17], and a pre-trained ViT [18] with a single linear projection layer, leading to performance close to that of GPT-4.

While LLMs and VLMs are powerful in many tasks, few efforts have been made to adapt them for image quality assessment (IQA), which is essential in the development of image reconstruction or enhancement algorithms [19–22]. In medical imaging, IQA plays a crucial role, directly influencing the accuracy and reliability of diagnoses [23, 24]. Particularly, in computed tomography (CT), reconstructed low-dose CT (LDCT) images from various deep-learning methods [25–33] may have blurring or over-smoothing problems, hindering their clinical translation. Therefore, assessing CT image quality before diagnosis is essential. Classic medical IQA methods can be either objective or subjective. Objective assessment methods use mathematical models for quantitative analysis, comparing the similarities or differences between reconstructed images and their references. Over the past decades, several objective IQA metrics have been widely used, including peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and root-mean-square error (RMSE). However, these metrics are usually unsatisfactory in radiological practice, as they do not effectively reflect the diagnostic utilities of images. Subjective IQA methods, on the other hand, involve expert opinions, which more accurately reflect the clinical needs [34]. However, the continuously growing number of CT images per scan poses a major burden on radiologists, who need to carefully assess each image.

In the past few years, deep-learning methods have been developed for diverse IQA tasks, including image perception [35–38], screen content [39], video [40], and medical images [41–44]. Blind pseudo-reference image-based method [36] introduced a no-reference IQA method that creates a pseudo-reference image to facilitate the quality assessment of distorted images. Unified content-type adaptive blind IQA model [37] proposed a unified framework for assessing the quality of compressed images across different content types. However, most of these methods focus on low-level image features, ignoring high-level features, especially hierarchical semantic information that is essential in the clinical context. To address this issue, Gao et al. [44] proposed an IQA network that integrates expert knowledge, combining the overall image quality ratings of radiologists with objective metrics as training labels.

Although using overall ratings as the optimization target can well reflect the overall noise level and fidelity of the image, it cannot meet the requirements of radiologists for extraction of clinically related subtleties, such as the small blood vessels, lymph nodes, and lesions.

Recently, CLIP-IQA [45] used a CLIP model to assess the similarity between images and predefined textual prompts. However, its design for natural images and dependence on simple text prompts limit its applicability for complex medical IQA, especially in evaluating fine structures and small lesions in CT images.

This study developed IQAGPT, a CT IQA system based on an image-quality captioning VLM incorporated with ChatGPT to generate quality scores and summarize quality reports of CT images. First, to train IQAGPT, a dataset of 1,000 image-text pairs named CT-IQA was compiled, in which an experienced radiologist scored CT images of different qualities, similar to the subjective evaluation previously reported [28, 34]. The qualities included image noise, small structures, lesion conspicuity, and diagnostic confidence. To utilize the strengths of LLMs in subjective image evaluation, a prompt template was designed to convert the quality scores to text descriptions. Second, an image-quality captioning model built upon a pre-trained medical VLM [12] was developed and fine-tuned on the CT-IQA dataset using an autoregressive language modeling objective that predicts the next token given previous tokens [3]. Finally, through interacting with ChatGPT, IQAGPT can score CT images and generate quality reports based on the caption from the image-quality captioning model. Figure 1 presents an exemplary dialogue between a user and the proposed IQAGPT.

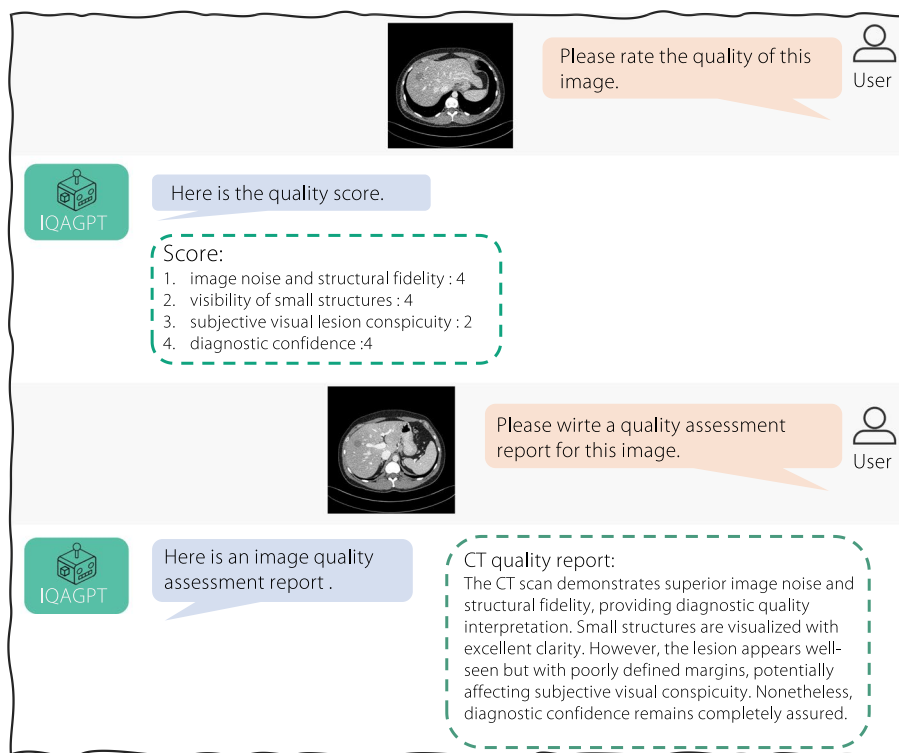


Fig. 1 A dialogue between humans and the proposed IQAGPT. In the dialogues, IQAGPT can output scores and write the quality report based on an input image

In summary, the main contributions of this work are as follows.

- A hybrid large model approach for CT-IQA, which synergizes the objective and subjective image quality evaluation in a clinically important scenario, is introduced.
- An IQA system consisting of VLMs and ChatGPT, termed IQAGPT, which is built on an image-quality captioning model and can output quality scores and reports by interacting with ChatGPT, is developed.
- A CT-IQA dataset for IQA, containing 1,000 image-text pairs professionally annotated according to four common subjective metrics used in diagnosis, was compiled.
- Preliminary results demonstrate the feasibility of assessing CT image quality using IQAGPT, and the resulting text-guided image-quality captioning model outperforms GPT-4 and CLIP-IQA. Furthermore, external evaluations by additional radiologists and performance on new data demonstrate the robustness and generalizability of the proposed method, respectively.

Methods

This study aims to develop a CT IQA system, called IQAGPT, using VLMs and ChatGPT. In CT-IQA dataset subsection, the CT-IQA dataset is detailed. Thereafter, in Image-quality captioning model and Interaction with ChatGPT subsections, the image-quality captioning model and IQAGPT which interacts with ChatGPT, are described, respectively. The implementation details are presented in Implementation details subsection and the performance evaluation of IQAGPT is explained in Evaluation metrics subsection.

CT-IQA dataset

To adapt to the IQA tasks and accurately assess the quality of CT images, an image-text dataset called CT-IQA, was compiled, in which an experienced radiologist subjectively assessed the CT images.

Characteristics

Normal-dose CT (NDCT) slices and corresponding LDCT images at 25% of the normal dose were randomly selected from the 2016 AAPM Grand Challenge dataset [46], which includes abdominal CT scans of 10 anonymous

patients. Specifically, 100 NDCT and LDCT pairs were uniformly selected from 8 patients for training and 25 slice pairs were uniformly selected from the remaining 2 patients for testing. Each scan was acquired using a Siemens SOMATOM Flash scanner and reconstructed with a B30 kernel and 1 mm slice thickness. The NDCT scans were acquired at 120 kV and 200 quality reference mAs (QRM), and the LDCT scans were acquired at 120 kV and 50 QRM. Additionally, some lesions in NDCT and corresponding LDCT were randomly simulated to evaluate subjective visual lesion conspicuity. The selected 125 LDCT images were processed using a modularized denoising model [28] called MAP-NN, producing various intermediate denoised images with associated noise reduction directions. RED-CNN [25], a widely used denoising model, was implemented, and optimized using the MSE loss function. Finally, 1,000 CT slices with different quality were obtained, including 125 NDCT slices with corresponding 125 LDCT slices, 625 reconstructed images with 5 denoising levels from MAP-NN, and 125 reconstructed images from RED-CNN. An abdomen window of all CT scans [-160, 240] HU was employed to visualize abdominal organs. These were normalized into a range of [0, 1]. Figure 2 presents CT images of eight different quality levels from the dataset, including LDCT, NDCT, images denoised with MAP-NN, and images denoised with RED-CNN.

Annotation process

First, a web page was created where all data were randomly displayed, including CT images of eight different levels in the dataset, including LDCT, NDCT, denoised image of MAP-NN, and denoised image of RED-CNN. Subsequently, a radiologist scored these CT images in terms of four metrics used in previous studies [28, 34], defined as follows.

- Image noise and structural fidelity on a four-point scale: 1=better than usual, acceptable for diagnostic interpretation; 2=average, acceptable for diagnostic interpretation; 3=sub-optimal, for limited diagnostic information only; and 4=unacceptable for diagnostic interpretation.
- The visibility of small structures (small blood vessels, adrenal glands, small lymph nodes) on a four-point scale: 1=excellent visualization; 2=acceptable visibility; 3=sub-optimal visibility; and 4=unacceptable visualization.
- Subjective visual lesion conspicuity (N/A=if no lesion) on a four-point scale: 1=well-seen lesion with well-visualized margins; 2=well-seen lesion with poorly visualized margins; 3=poorly seen lesion with poorly visualized margins; and 4=lesion blurred with severe loss of margins.

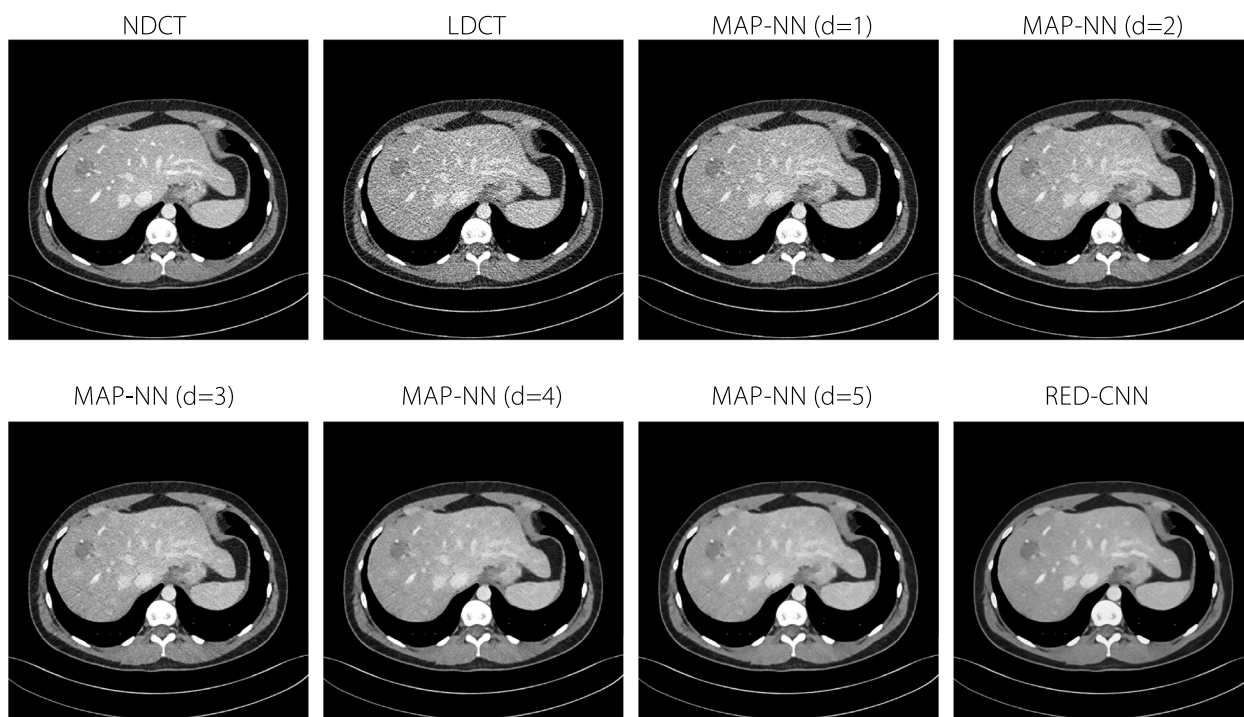


Fig. 2 Examples of images from the CT-IQA dataset

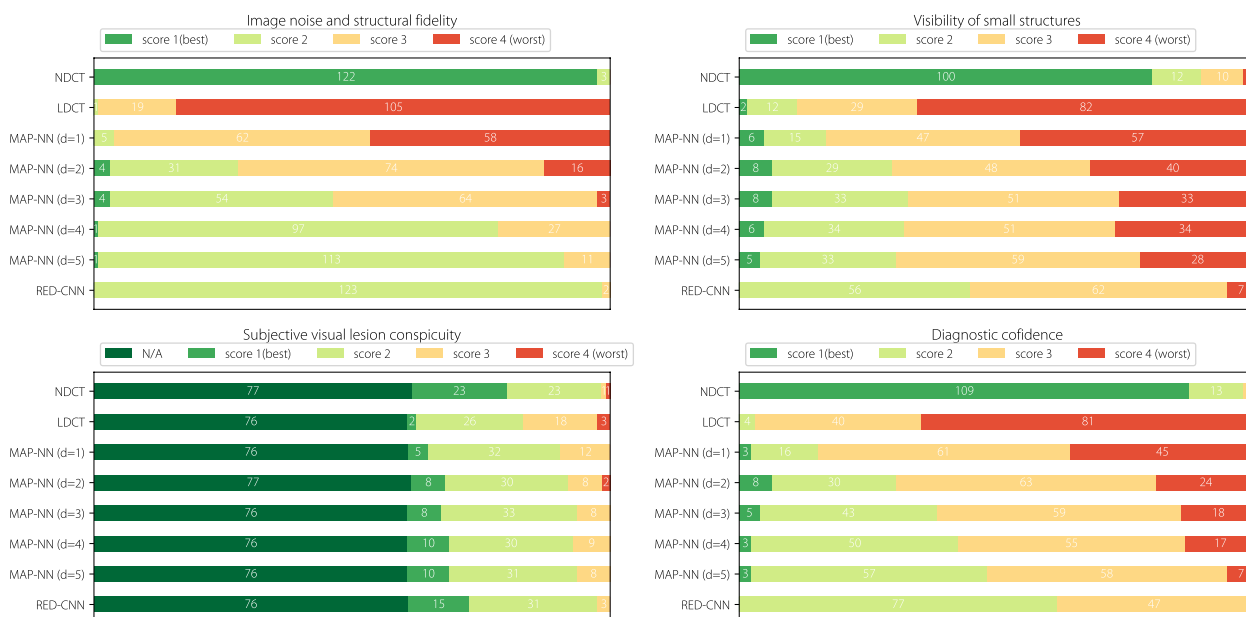


Fig. 3 Score distribution of four metrics assessed by the radiologist in constructing CT-IQA dataset. Scores 1, 2, 3, and 4 are defined in CT-IQA dataset subsection

Table 1 Quantitative performance measures on different quality levels, with NDCT as the reference image

Method	PSNR↑	RMSE↓	SSIM↑
LDCT	21.85 ± 1.25	0.082 ± 0.011	0.7897 ± 0.0250
MAP-NN (d=1)	24.01 ± 1.26	0.063 ± 0.009	0.8171 ± 0.0229
MAP-NN (d=2)	25.56 ± 1.25	0.053 ± 0.007	0.8339 ± 0.0215
MAP-NN (d=3)	26.43 ± 1.22	0.048 ± 0.006	0.8394 ± 0.0209
MAP-NN (d=4)	26.80 ± 1.19	0.046 ± 0.006	0.8365 ± 0.0211
MAP-NN (d=5)	26.89 ± 1.16	0.046 ± 0.006	0.8294 ± 0.0218
RED-CNN	27.16 ± 1.12	0.044 ± 0.005	0.8270 ± 0.0225

- Diagnostic confidence on a four-point scale: 1 = completely confident; 2 = probably confident; 3 = confident only for a limited clinical entity such as a kidney stone, a calcified lesion, or a large lesion; and 4 = poor confidence.

The scoring process was double-blinded; that is, the radiologist did not know the type of images under evaluation; NDCT was not availed for reference. Figure 3 shows the distribution of human expert scores across the aforementioned four metrics for CT images of eight different image qualities: NDCT, LDCT, MAP-NN (d=1), MAP-NN (d=2), MAP-NN (d=3), MAP-NN (d=4), MAP-NN (d=5), and RED-CNN. The denoising level is represented by d.

The objective metrics for images of varying quality were calculated with NDCT as the reference image, shown in Table 1. RED-CNN achieves the best performance in PSNR and RMSE. However, as illustrated by the distribution of annotated scores in Fig. 3, the objective results generated by RED-CNN do not align with the professional preference, which further highlights the necessity to align the annotation of subjective quality-assessment datasets with professional preference.

Image-quality captioning model

Instead of using the rating scores to train the classification or regression model, an image-quality captioning model is developed to summarize the image quality. By doing so, the VLM with semantic text information and image-text fusion can better appreciate the subjective scores than image-only models—this is further discussed in Results section. The proposed model is based on a pre-trained medical VLM and fined-tuned with an autoregressive language modeling objective on the CT-IQA dataset. To leverage the capabilities of LLMs in the subjective image evaluation, we convert scores to quality descriptions using a specific prompt template during training. The prompt template is defined as “Image noise and structural fidelity: {description 1}; Visibility of small structures: {description 2}; Subjective visual lesion conspicuity: {description 3}; Diagnostic confidence: {description 4}.” Every description is the evaluation criterion

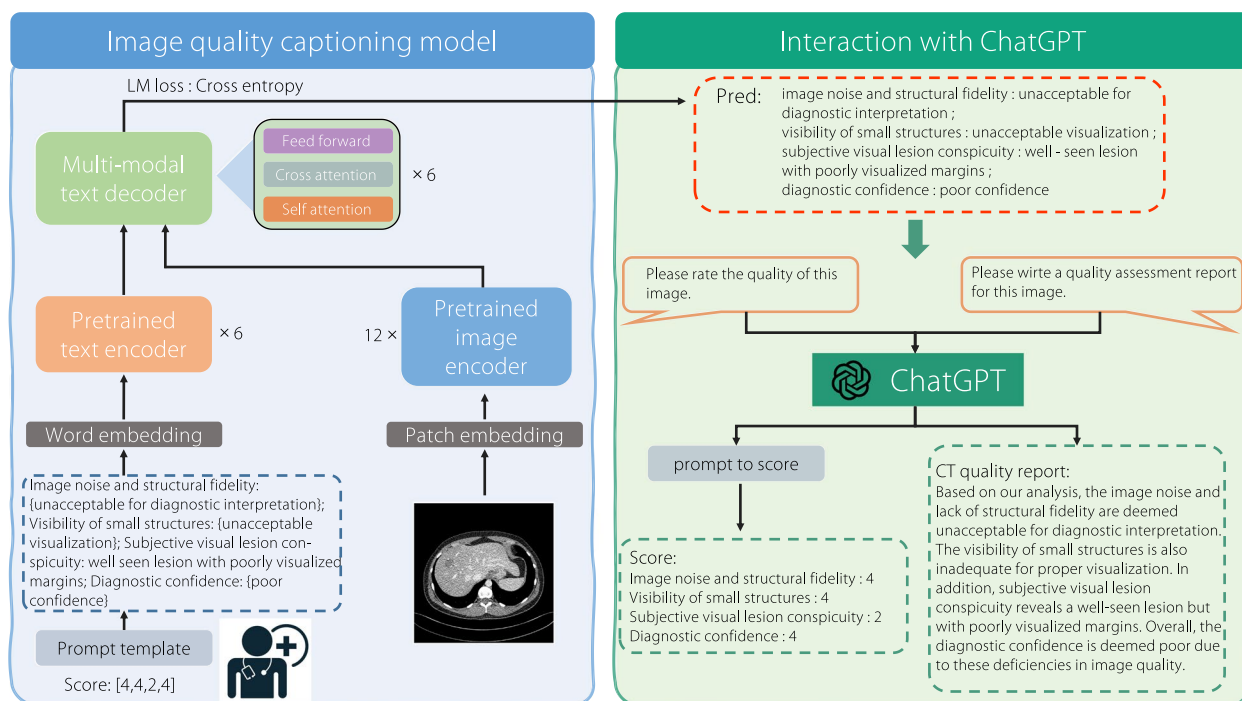


Fig. 4 Overview of IQAGPT. While the left side shows the proposed image-quality captioning model, the right side details the process of the score and report generation through interacting with ChatGPT

corresponding to the score described in CT-IQA dataset subsection. An example of score conversion to quality caption is given in the lower left part of Fig. 4, where the score assessed by the radiologist is [4, 4, 2, 4].

The left side of Fig. 4 presents the overall framework of the proposed image quality captioning model, consisting of an image encoder, text encoder, and multimodal text decoder. The image encoder is a 12-layer visual transformer ViT-S/16 [18] while the text encoder is the first 6 layers of the BERT_{base} [47] model. The multimodal text decoder consists of the last 6 layers of BERT_{base}; its role is to fuse image and text features through cross-modal attention. Some recent studies have incorporated cross-modal attention [48–50]. Min et al. [49, 50] used the normalization and summation fusion function to integrate audio-visual contexts. In contrast, the proposed captioning model leverages the synergy between visual data and textual descriptions through a transformer-based [18] cross-attention mechanism to fuse image features with text features. The image encoder, text encoder, and multimodal decoder have been pre-trained in radiography images and report pairs [12] using four widely used learning objectives in the field of vision-language alignment; more details on these four objectives are in ref. [12]. We then fine-tuned the pre-trained models to predict the next word for IQA on the CT-IQA dataset, using an auto-regressive paradigm. We hypothesize that

this paradigm, combined with the proposed input template, allows LLMs to better comprehend the relationship between different metrics. CT-text pair is denoted as (I, T) , where I represents a CT slice and T is defined as $T = (t_1, t_2, \dots, t_m)$ with m tokens. The objective function is defined as follows:

$$L(I, T) = -(\log P(t_1|I; \theta) + \sum_{i=2}^m \log P(t_i|t_{1:i-1}, I; \theta)) \quad (1)$$

where t_i is the next token to be predicted and $t_{1:i-1} = (t_1, t_2, \dots, t_{i-1})$ represents the sequence of all previous tokens. P is the conditional probability modeled by the image-quality captioning model, and θ represents the trainable parameters of the model.

Interaction with ChatGPT

ChatGPT provides a language interface with remarkable reasoning capabilities across many domains [11]. The proposed IQAGPT enables the interaction between ChatGPT and users to generate more comprehensive output information, as depicted on the right side of Fig. 4. When users upload CT images, they can prompt IQAGPT with requests like “Please rate the quality of this image.” or “Please write a quality-assessment report for this image.” Subsequently, the users receive either quality scores or detailed quality reports. To this end, ChatGPT is used to perform corresponding operations on the output caption from the image-quality captioning model. For

score-related demands, it converts the predicted caption to a score according to the prompt template described in CT-IQA dataset subsection. For report-related demands, it summarizes the predicted caption into a quality-assessment report in a radiology report format.

While it is straightforward to obtain scores using a look-up table, integrating ChatGPT into the proposed model leverages its advanced natural language understanding capabilities to generate detailed and context-aware quality reports, providing the following benefits. (1) Contextual understanding: the ability of ChatGPT to comprehend and generate contextually relevant text ensures that the quality reports are not only accurate but also rich in clinical context, which is more friendly for radiologists. (2) Flexibility: unlike a static look-up table, ChatGPT can be adapted to variations in input data, providing more nuanced and flexible assessments. (3) Scalability: ChatGPT can easily incorporate new quality metrics without requiring significant modifications to the model structure.

Implementation details

All parameters of the proposed model were fine-tuned using a 32 GB NVIDIA V100 GPU. During training, the image-quality captioning model was fine-tuned in IQAGPT for 50 epochs with a batch size of 8, in which we used the AdamW optimizer [51] and a weight decay of 0.02. The initial learning rate was 2.0×10^{-4} , and warm-up [52] in the first 2 epochs had a learning rate of 1.0×10^{-5} , gradually reduced to 1.0×10^{-6} with cosine annealing [53]. For data processing, full-size images were employed within an abdomen window of [-160, 240] HU. The data of 10 patients were split into training and testing datasets at a ratio of 8:2 as described in CT-IQA dataset subsection. The training data were augmented through horizontal flipping and rotation.

Evaluation metrics

To show the effectiveness of IQAGPT, we quantitatively evaluated the performance of generated quality captioning and score. First, captioning results were analyzed using widely recognized metrics in text generation tasks: bilingual evaluation understudy (BLEU-n; “n” means n words) [54], recall-oriented understudy of gisting evaluation (ROUGE-L; “L” means the longest common subsequence) [55], metric for evaluation of translation with explicit ordering (METEOR) [56], and consensus-based image description evaluation (CIDEr; “r” stands for recall) [57]. These metrics measure the similarity between the generated and reference texts, with higher scores for better quality. BLEU measures the quality of machine-translated text compared to a human reference translation. It calculates the precision for the candidate sentence based on n-grams (phrases of n words) with respect to the reference texts. ROUGE-L focuses on the longest common subsequence between the

evaluated text and the reference text. METEOR is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. METEOR calculates the weighted harmonic mean of unigram precision and recall, prioritizing recall over precision. CIDEr quantifies the resemblance of the crafted sentence to multiple reference sentences, considering the agreement among human evaluators. Notably, BLEU-n, ROUGE-L, and METEOR scores range from 0 to 1 while CIDEr scores range from 0 to infinity. In addition, the output text descriptions were converted into scores, the performance was compared in terms of accuracy as the classification evaluation, and the Pearson linear correlation coefficient (PLCC) and Spearman’s rank order correlation coefficient (SROCC) were computed to evaluate the regression.

Results

Evaluation of generated quality captioning

Two examples of the test results are presented in Fig. 5, where the predicted descriptions are converted to scores and quality reports using ChatGPT. It can be observed that IQAGPT consistently generates quality descriptions in excellent alignment with the annotations of radiologists. Furthermore, the reports generated using ChatGPT are consistent with the outputs from the proposed quality captioning model, which effectively overcomes the limitations of the existing VLM dialogue when assessing the quality of medical images. The quantitative captioning performance of IQAGPT and MiniGPT-4 were compared, as depicted in Table 2. GPT-4 [15] was not employed as its latest version, GPT-4 V, was not tailored for interpreting specialized medical imagery such as CT scans. The learnable linear layer in MiniGPT-4 was fine-tuned using the CT-IQA dataset in the experimental settings described by Zhu et al. [16]. IQAGPT achieves better quantitative results in seven metrics. The requirement for vast amounts of training data is a significant challenge for LLMs like MiniGPT-4. Considering the time and resource limitations, only 1,000 image-text pairs were annotated in the CT-IQA dataset to demonstrate the feasibility of the proposed method. MiniGPT-4, with its considerable size of seven billion parameters, struggles with this limited dataset, leading to unstable output and compromised performance. Conversely, the proposed IQAGPT, with a more compact model structure, produces more stable and accurate results.

Evaluation of generated quality score

To validate the efficacy of the proposed image-quality captioning model, the proposed prompt template was employed to transform output text descriptions into scores to assess IQAGPT performance in both



	LDCT		NDCT
<p>Ground-truth</p> <ol style="list-style-type: none"> 1. Image noise and structural fidelity : unacceptable for diagnostic interpretation; 2. Visibility of small structures : unacceptable visualization; 3. Subjective visual lesion conspicuity : no lesion; 4. Diagnostic confidence : poor confidence 		<p>Ground-truth</p> <ol style="list-style-type: none"> 1. Image noise and structural fidelity : better than usual, acceptable for diagnostic interpretation; 2. Visibility of small structures : acceptable visibility; 3. Subjective visual lesion conspicuity: no lesion 4. Diagnostic confidence : completely confident 	
<p>Predicted</p> <ol style="list-style-type: none"> 1. Image noise and structural fidelity: unacceptable for diagnostic interpretation; 2. Visibility of small structures: unacceptable visualization; 3. Subjective visual lesion conspicuity: no lesion; 4. Diagnostic confidence: poor confidence 		<p>Predicted</p> <ol style="list-style-type: none"> 1. Image noise and structural fidelity : better than usual, acceptable for diagnostic interpretation; 2. Visibility of small structures: excellent visualization; 3. Subjective visual lesion conspicuity: no lesion; 4. Diagnostic confidence : completely confident 	
<p>Rating by ChatGPT</p> <ol style="list-style-type: none"> 1. Image noise and structural fidelity: 4 2. Visibility of small structures: 4 3. Subjective visual lesion conspicuity: N/A 4. Diagnostic confidence: 4 		<p>Rating by ChatGPT</p> <ol style="list-style-type: none"> 1. Image noise and structural fidelity: 1 2. Visibility of small structures: 1 3. Subjective visual lesion conspicuity: N/A 4. Diagnostic confidence: 1 	
<p>Report by ChatGPT</p> <p>The CT image reviewed exhibits unacceptable levels of image noise and structural fidelity, rendering it unsuitable for accurate diagnostic interpretation. Additionally, the visualization of small structures is inadequate, making it difficult to identify and analyze relevant anatomical features. No lesions are subjectively identified, but the image lacks any apparent visual lesion conspicuity. As a result of these limitations, diagnostic confidence is poor, and there is limited confidence in the accuracy of the imaging findings.</p>		<p>Report by ChatGPT</p> <p>The CT image under review exhibits better-than-usual image noise and structural fidelity, allowing for accurate diagnostic interpretation. The visibility of small structures is excellent, providing a clear visualization of the relevant anatomical features. No lesions are subjectively identified, and thus the image lacks any apparent visual lesion conspicuity. Overall, diagnostic confidence is high, and there is complete confidence in the accuracy of the imaging findings.</p>	

Fig. 5 Captions predicted using the predicted method and scores and reports generated using ChatGPT

Table 2 Quantitative evaluation of captioning quality using IQAGPT and MiniGPT-4

Method	Parameter	BLEU-1	BLUE-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
MiniGPT-4	7B	0.798	0.733	0.717	0.652	0.516	0.826	3.070
IQAGPT	210 M	0.819	0.777	0.742	0.712	0.546	0.858	3.620

Table 3 Comparison of classification evaluation accuracy between IQAGPT and ViT-C

Classification	ViT-C	IQAGPT
Image noise and structural fidelity	0.545	0.765
Visibility of small structures	0.405	0.620
Subjective visual lesion conspicuity	0.725	0.820
Diagnostic confidence	0.375	0.605
Mean	0.512	0.702

classification and regression tasks. A comparative study on IQAGPT was conducted with an image-only multi-task classification model, using accuracy as a metric. Additionally, IQAGPT was compared with CLIP-IQA [45] and an image-only multi-task regression model, employing PLCC and SROCC. These quantitative results are detailed in Tables 3 and 4. The PLCC and SROCC calculation for the metric of subjective visual lesion conspicuity was not performed as over half of the CT scans

Table 4 Comparison of IQAGPT with CLIP-IQA and ViT-R in the regression evaluation performance in terms of PLCC/SROCC

Regression	CLIP-IQA	CLIP-IQA +	ViT-R	IQAGPT
Image noise and structural fidelity	0.277/0.271	0.742/0.633	0.580/0.460	0.821/0.820
Visibility of small structures	0.121/0.117	0.712/0.696	0.436/0.415	0.743/0.735
Subjective visual lesion conspicuity	-	-	-	-
Diagnostic confidence	0.081/0.069	0.650/0.642	0.504/0.422	0.699/0.689
Mean	0.160/0.114	0.701/0.657	0.531/0.519	0.754/0.748

CLIP-IQA + represents the fine-tuned version of CLIP-IQA

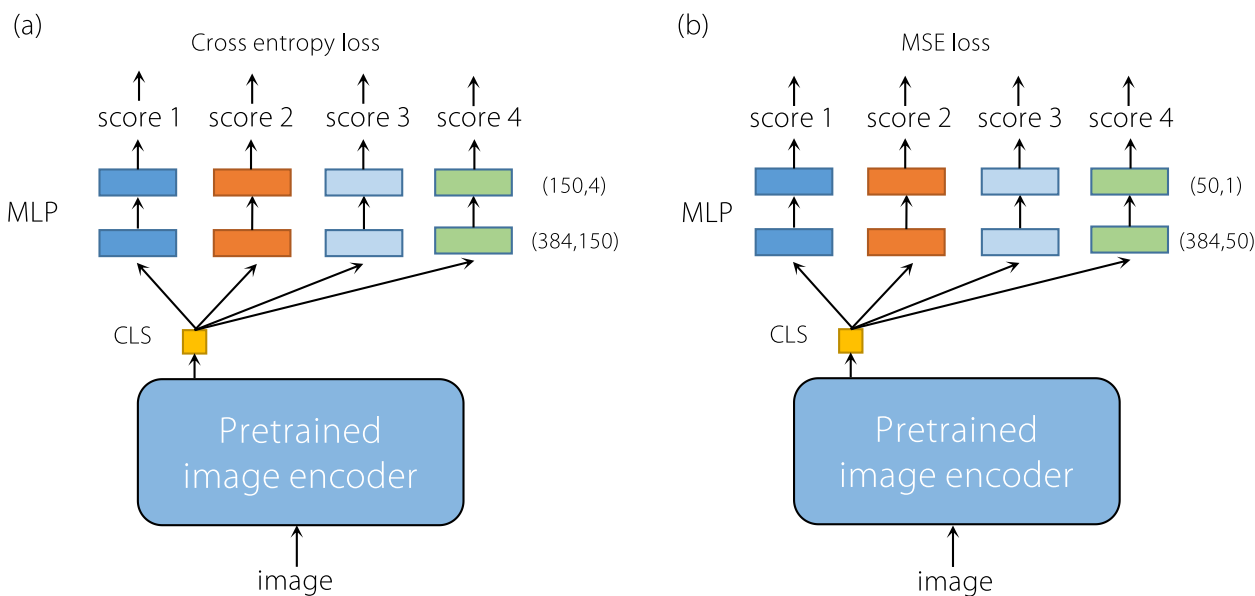


Fig. 6 Flowcharts of (a) multi-task classification model ViT-C and (b) multi-task regression model ViT-R, respectively. CLS tokens are followed by four groups of classifiers, each consisting of two fully connected layers. Scores 1, 2, 3, and 4 are the categories corresponding to the four metrics described in CT-IQA dataset subsection

in the dataset did not contain lesions. The image-only multi-task classification and regression models were named ViT-C and ViT-R respectively. First, the same pre-trained image encoder (ViT-S/16) was employed in IQAGPT to extract image features. Then, four pairs of fully connected layers were implemented following the classification (CLS) token for four metrics, as depicted in Fig. 6. ViT-C and ViT-R employed cross-entropy loss and mean squared error loss respectively. The same training strategy was used with IQAGPT to train CLIP-IQA+, ViT-C, and ViT-R.

Table 3 shows that IQAGPT outperforms the image-only classification model, ViT-C, across four metrics, achieving a notable improvement of 0.19 in mean accuracy. For regression, IQAGPT surpasses ViT-R and CLIP-IQA, as shown in Table 4. Compared with ViT-C and ViT-R, which represent ablation methods without LLM, IQAGPT outperforms them because it uses LLM

to analyze detailed text information instead of using only raw scores as labels. Regarding CLIP-IQA, which uses CLIP to perceive subjective attributes through text prompt pairing, these texts contain only a single adjective, enabling the assessment of global quality attributes such as noisiness and brightness, but insufficient to capture complex details in medical images for diagnosis. In contrast, IQAGPT has complex text descriptions using an autoregressive LLM model. Furthermore, a notable advantage of IQAGPT is its efficiency; unlike CLIP-IQA, which requires separate fine-tuning for each of the four metrics, IQAGPT can simultaneously produce results for all metrics in a single output.

For each image-quality level and metric, accuracy was computed using converted scores, as depicted in Table 5. The relative accuracies associated with intermediate images generated by MAP-NN may not be highly robust owing to their similar features. This aligns with

Table 5 Accuracy for each of the four metrics in eight image-quality levels

Metric	NDCT	LDCT	MAP-NN (1)	MAP-NN (2)	MAP-NN (3)	MAP-NN (4)	MAP-NN (5)	RED-CNN	Mean
Metric 1	1.000	0.800	0.600	0.520	0.480	0.880	0.880	1.000	0.765
Metric 2	0.960	0.680	0.480	0.360	0.520	0.600	0.640	0.720	0.620
Metric 3	0.920	0.920	0.880	0.760	0.800	0.840	0.760	0.800	0.820
Metric 4	0.920	0.760	0.360	0.400	0.320	0.440	0.680	0.960	0.605
Mean	0.950	0.790	0.580	0.510	0.520	0.690	0.740	0.840	0.702

Metric 1: Image noise and structural fidelity; Metric 2: Visibility of small structures; Metric 3: Subjective visual lesion conspicuity; and Metric 4: Diagnostic confidence. MAP-NN (-) provides 5 denoising levels [28]

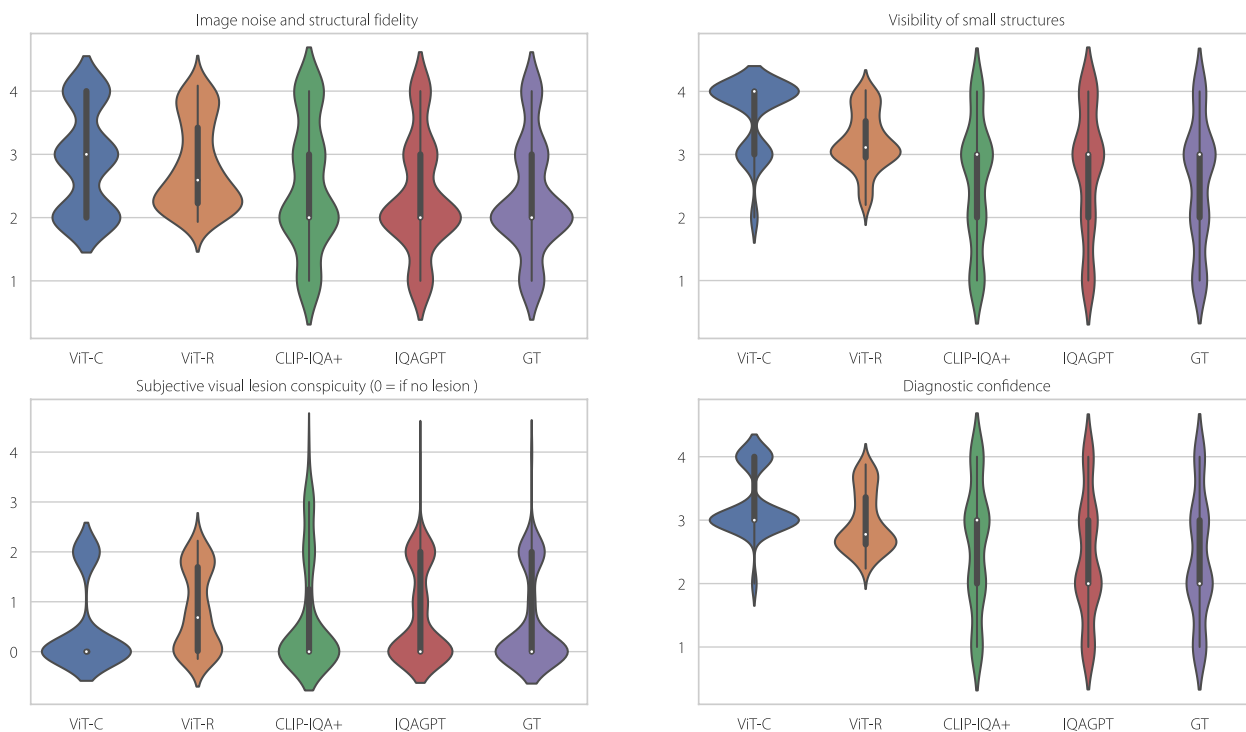


Fig. 7 Scores distribution for four quality metrics using IQAGPT, ViT-C, ViT-R, and CLIP-IQA+. The last column lists the GT scores

the challenges in the subjective evaluation of images with subtle quality differences, a critical aspect of the CT-IQA dataset. This study highlights the complexity of differentiating between similar images.

Furthermore, the score distributions of IQAGPT, ViT-C, ViT-R, and CLIP-IQA+ for four quality metrics are presented in Fig. 7. Notably, our method more closely approximates the groundtruth (GT) compared to ViT-C, ViT-R, and CLIP-IQA+, demonstrating its effectiveness. Overall, IQAGPT has a higher correlation with human perception than the competing methods, marking a significant advancement in CT subjective IQA.

To further demonstrate the effectiveness of IQAGPT, the predicted and GT scores for each quality level and metric are visualized in Fig. 8, demonstrating the

prediction accuracy of IQAGPT. In addition, Fig. 9 presents the predicted scores for NDCT and corresponding denoising results from MAP-NN (d=1) and RED-CNN, along with the calculated PSNR and SSIM. It can be observed that the quantitative results of MAP-NN (d=1) are inferior to that of RED-CNN; however, in professional subjective assessment, these two are similar and considered acceptable. From the perspective of the radiologist, the results of MAP-NN (d=1) suffer from incomplete denoising, leading to some blurred details, while the RED-CNN results exhibit over-smoothing issues due to the use of a pixel-level loss function. In contrast, the scores of the results predicted by IQAGPT are almost identical to the GT ones,

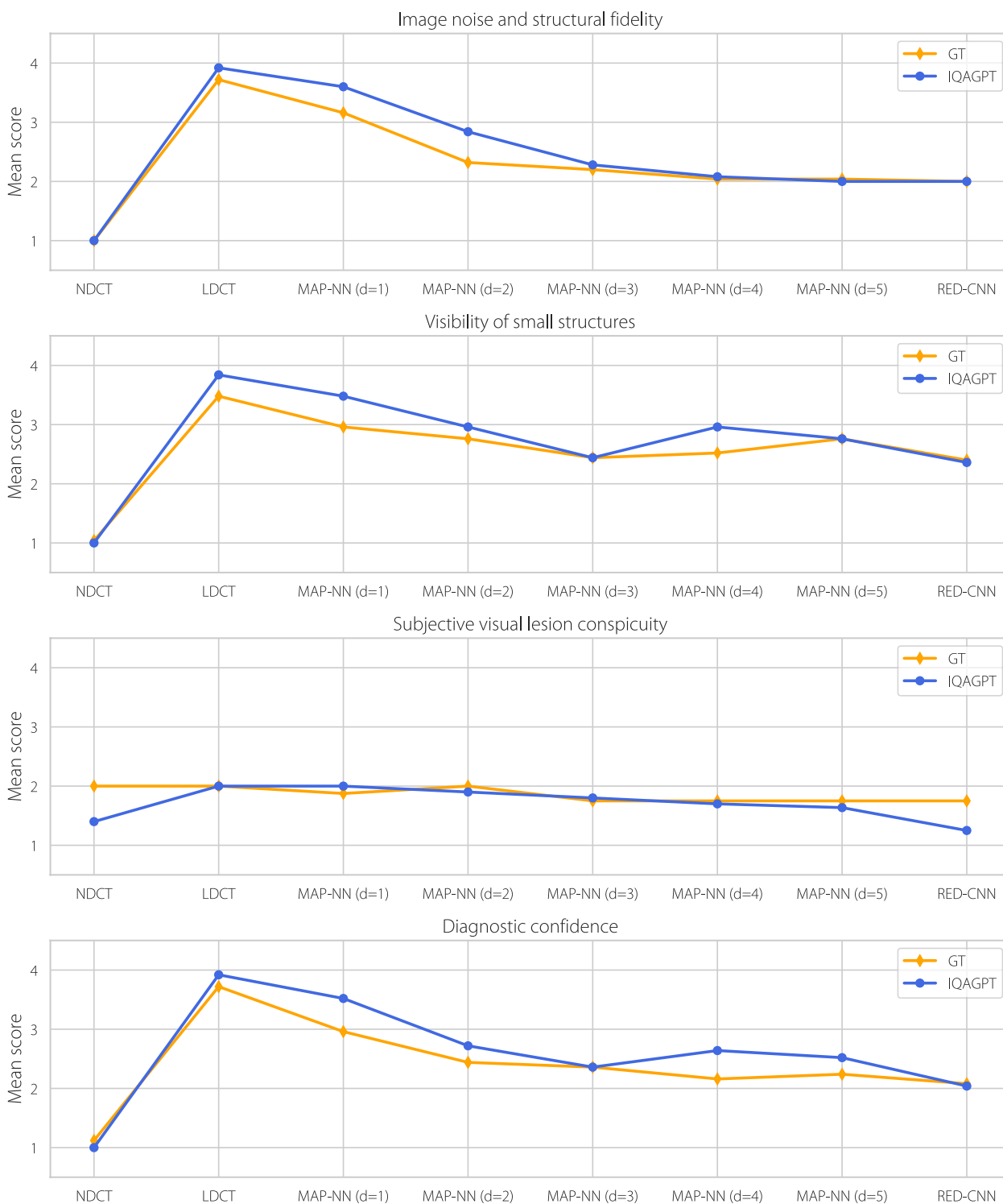


Fig. 8 Mean values of the GT scores and the scores predicted by IQAGPT



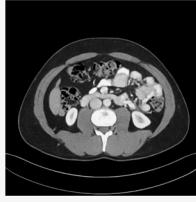
NDCT	MAP-NN (d=1)	RED-CNN
	 PSRN: 24.58 SSIM: 0.8254	 PSRN: 27.34 SSIM: 0.8347
Ground-truth 1. Image noise and structural fidelity : 1 2. Visibility of small structures : 1 3. Subjective visual lesion conspicuity : N/A 4. Diagnostic confidence : 1	Ground-truth 1. Image noise and structural fidelity : 2 2. Visibility of small structures : 2 3. Subjective visual lesion conspicuity : N/A 4. Diagnostic confidence : 2	Ground-truth 1. image noise and structural fidelity: 2 2. visibility of small structures: 2 3. Subjective visual lesion conspicuity: N/A 4. diagnostic confidence: 2
Predicted 1. image noise and structural fidelity: 1 2. visibility of small structures: 1 3. Subjective visual lesion conspicuity: N/A 4. diagnostic confidence: 1	Predicted 1. image noise and structural fidelity: 3 2. visibility of small structures: 2 3. Subjective visual lesion conspicuity: N/A 4. diagnostic confidence: 2	Predicted 1. image noise and structural fidelity: 2 2. visibility of small structures: 3 3. Subjective visual lesion conspicuity: N/A 4. diagnostic confidence: 2

Fig. 9 Scores of three examples predicted using IQAGPT



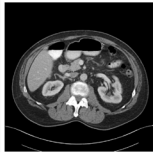
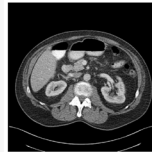
	NDCT	LDCT	NDCT	LDCT
				
ViT-C	Image noise and structural fidelity : █ Visibility of small structures : █ Subjective visual lesion conspicuity : █ Diagnostic confidence : █	Image noise and structural fidelity : 3 Visibility of small structures : █ Subjective visual lesion conspicuity : N/A Diagnostic confidence : 3	Image noise and structural fidelity : 2 Visibility of small structures : 2 Subjective visual lesion conspicuity : N/A Diagnostic confidence : █	Image noise and structural fidelity : █ Visibility of small structures : █ Subjective visual lesion conspicuity : N/A Diagnostic confidence : 2
IQAGPT	Image noise and structural fidelity : 1 Visibility of small structures : █ Subjective visual lesion conspicuity : N/A Diagnostic confidence : 1	Image noise and structural fidelity : 3 Visibility of small structures : 3 Subjective visual lesion conspicuity : N/A Diagnostic confidence : 3	Image noise and structural fidelity : 2 Visibility of small structures : 2 Subjective visual lesion conspicuity : N/A Diagnostic confidence : █	Image noise and structural fidelity : 3 Visibility of small structures : 2 Subjective visual lesion conspicuity : N/A Diagnostic confidence : 2
Ground Truth	Image noise and structural fidelity : 1 Visibility of small structures : 1 Subjective visual lesion conspicuity : N/A Diagnostic confidence : 1	Image noise and structural fidelity : 3 Visibility of small structures : 3 Subjective visual lesion conspicuity : N/A Diagnostic confidence : 3	Image noise and structural fidelity : 2 Visibility of small structures : 2 Subjective visual lesion conspicuity : N/A Diagnostic confidence : 1	Image noise and structural fidelity : 3 Visibility of small structures : 2 Subjective visual lesion conspicuity : N/A Diagnostic confidence : 2

Fig. 10 Scores of four examples predicted using IQAGPT in Mayo2020 dataset. Wrong scoring is highlighted in green

demonstrating that IQAGPT learned IQA expertise consistent with the clinical needs.

Evaluation on new datasets

This study proposes a novel paradigm for IQA by leveraging LLMs to generate text descriptions; however, no datasets specifically annotated for this purpose currently exist. To further verify the generalizability of IQAGPT, we evaluated it on “Low Dose CT Image and Projection Data” latest released by Mayo Clinic in 2020 [58], named Mayo2020 dataset, which includes NDCT and LDCT

images. Owing to the cost of professional annotations, the radiologist annotated several images.

Figure 10 shows that IQAGPT is more consistent with the gold standard of the radiologist than ViT-C. Furthermore, IQAGPT did not blindly categorize NDCT as the best or LDCT as the worst. This is because the noise characteristics of the Mayo2020 differ from those of the CT-IQA dataset used for training in this study. IQAGPT produced results that align with the preference of the radiologist, demonstrating its adaptability across different datasets.

Table 6 Performance of IQAGPT with three radiologist annotations in the regression evaluation in terms of PLCC/SROCC

Regression	R1	R2	R3
Image noise and structural fidelity	0.821/0.820	0.750/0.756	0.755/0.750
Visibility of small structures	0.743/0.735	0.668/0.659	0.680/0.673
Subjective visual lesion conspicuity	-	-	-
Diagnostic confidence	0.699/0.689	0.661/0.672	0.679/0.688
Mean	0.754/0.748	0.694/0.696	0.705/0.704

External evaluation by additional radiologists

The CT-IQA dataset was annotated by one radiologist. To investigate the bias introduced by the radiologist, an external evaluation was conducted, in which two additional radiologists were invited to annotate. Each radiologist independently assessed the CT images using the quality metrics predefined in this study. Considering the high cost and significant time required for professional annotations, this was limited to the test dataset only. Radiologist 1 (R1), with nine years of experience, was the original annotator for both the training and testing sets in previous evaluations. Each of the two additional radiologists, Radiologist 2 (R2) and Radiologist 3 (R3), had two years of experience.

The performance of IQAGPT was evaluated separately for each radiologist’s annotations using PLCC and SROCC. Table 6 shows that the best results are achieved on the test set annotated by R1, indicating the consistency between the training and test sets annotated by R1. The PLCC/SROCC scores for R2 and R3, though dropping a little bit, are still comparable to those of R1, demonstrating the strong robustness of the developed model against external evaluation and verification. Different radiologists have different biases regarding image

quality; however, the high correlation among radiologists shows a small bias between internal and external evaluations.

Ablation on LLMs

To further demonstrate the effectiveness of textual semantic information, the *t*-SNE [59] method was employed to visualize the features of the CLS tokens in the image encoders of IQAGPT, ViT-C, and ViT-R, as illustrated in Fig. 11. Each sample was labeled using the score of the image noise and structural fidelity metric. This visualization demonstrates that IQAGPT distinguishes features of different categories more clearly than ViT-C and ViT-R, and exhibits an ordered sequence in the score-based feature representation. Additionally, the self-attention map of tokens from the multimodal text decoder, depicted in Fig. 12, reveals that each token is interconnected not only with tokens from the same task but also with those from preceding tasks. This finding underscores the merits of textual descriptions in capturing inter-task correlations, enhancing the classification performance.

Interpretation

To provide an interpretation of the proposed quality captioning model, per-word Grad-CAM visualizations are presented in Fig. 13. The Grad-CAM visualizations are highly correlated with where radiologists look at when making decisions. For instance, radiologists tend to concentrate on the global appearance of an image when assessing ‘noise,’ whereas local features gain more attention during evaluations of ‘diagnosis’ or ‘lesions.’

Overall, the above findings indicate that IQAGPT can successfully perform CT subjective quality-assessment tasks. It can predict texts aligned with the GT and also

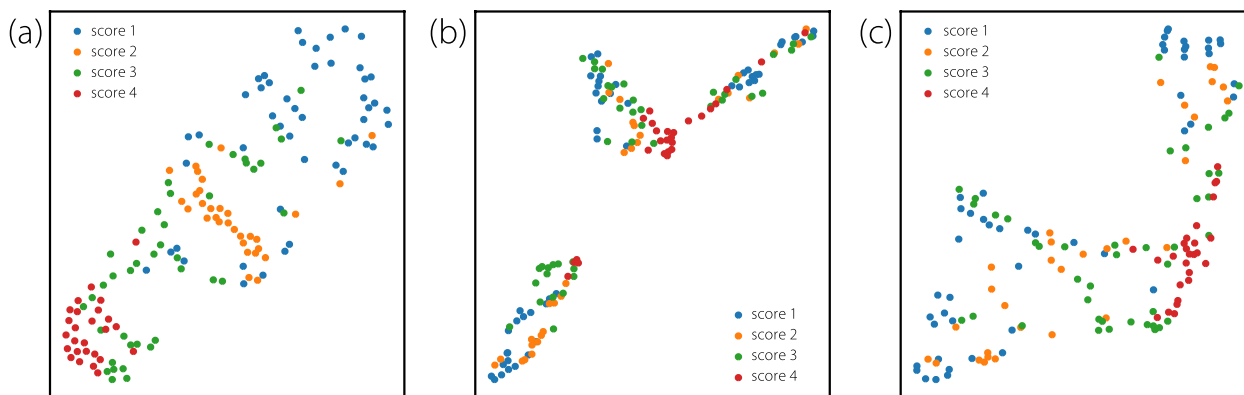


Fig. 11 Feature visualization of the CLS token in the image encoder of (a) IQAGPT, (b) ViT-C, and (c) ViT-R, using the *t*-SNE method. The samples are labeled with categories from the metric of image noise and structural fidelity

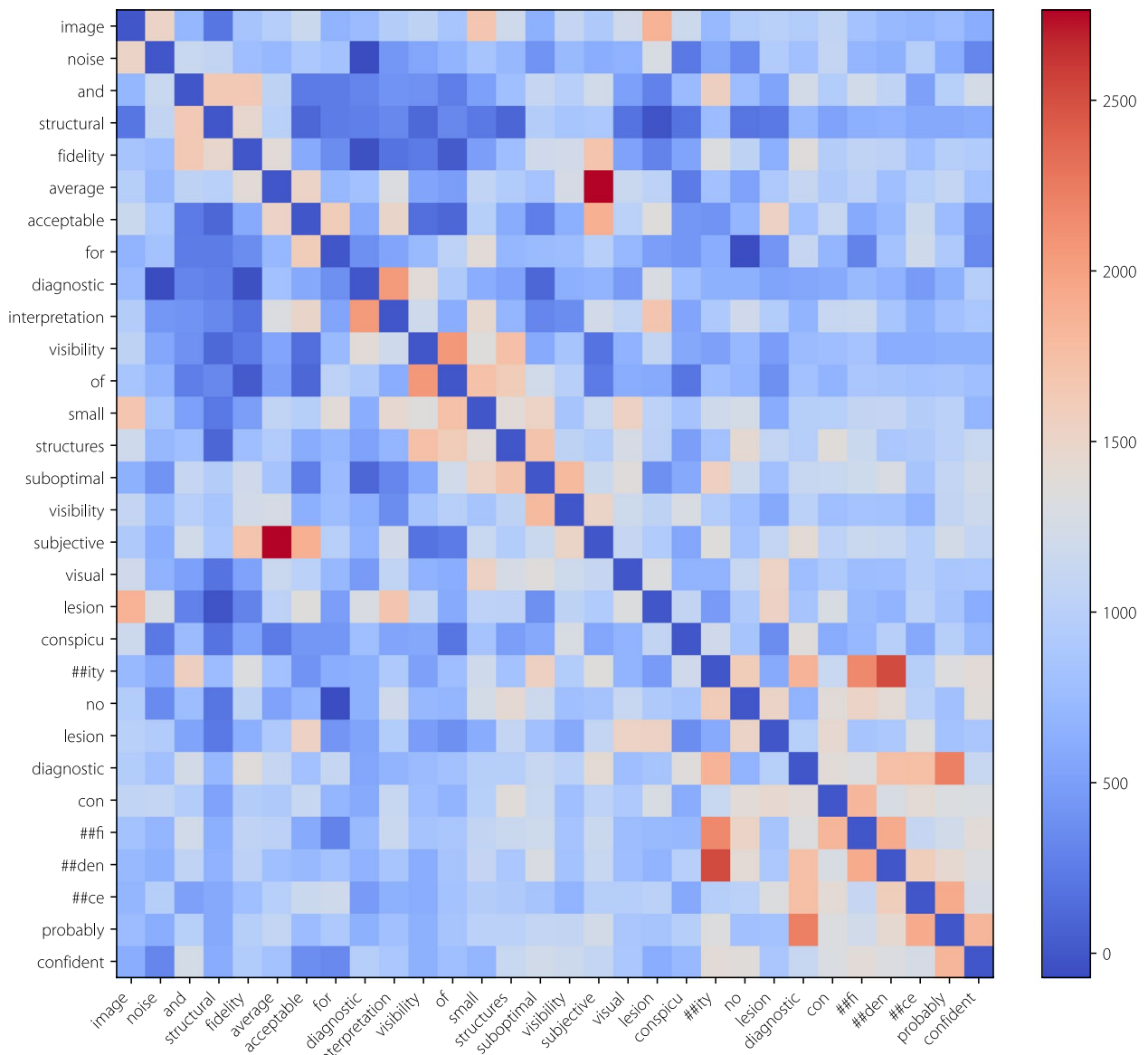


Fig. 12 Self-attention map of tokens from the last layer in the multimodal text decoder

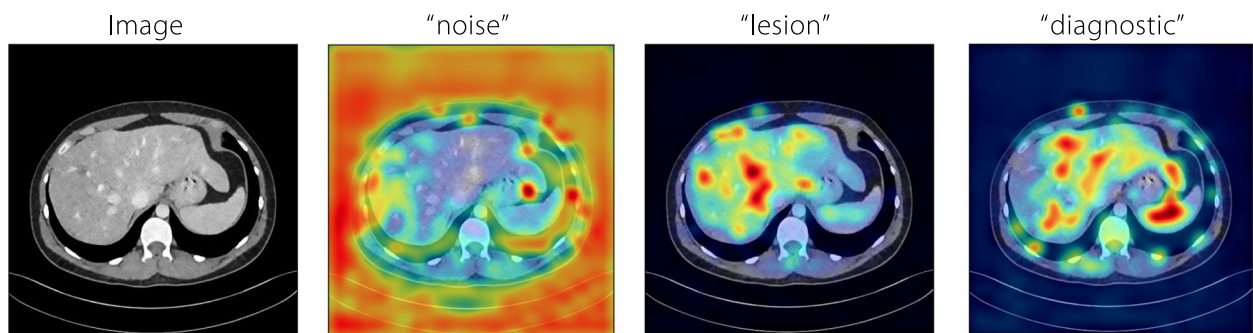


Fig. 13 Grad-CAM visualizations on the cross-attention maps corresponding to individual words

translate these texts into scores and reports using ChatGPT in a clinically meaningful way.

Discussion

This study highlights the efficacy of integrating large models for IQA, with a specific focus on LDCT denoising. It suggests a significant potential to replace the traditional subjective image-quality evaluation procedure conducted by radiologists with large hybrid deep models, which are resource-efficient and time-saving. In other words, the developed IQAGPT is the first attempt in this direction, and IQAGPT not only eases the burden on radiologists by automating CT IQA but also aids radiologists in refining diagnostic performance.

The proposed method was developed on the CT-IQA dataset of 1,000 image-text pairs annotated by a professional radiologist. For this purpose, a prompt template was leveraged to transform quality scores into text descriptions. Having fine-tuned the image-quality captioning model on the CT-IQA dataset, IQAGPT can generate quality descriptions for different CT scans. Using ChatGPT as an interactive interface facilitates user engagement, allowing for versatile outputs including quality scores and comprehensive reports.

Experimental results demonstrate the efficacy of IQAGPT in steadily generating quality descriptions and converting them into scores and reports. Quantitative evaluation using metrics for image captioning, classification, and regression tasks, underscores the superior performance of IQAGPT. In addition, ablation studies show the effectiveness of incorporating LLMs in subjective CT-IQA tasks; IQAGPT can integrate the expertise of radiologists with the advanced capabilities of LLMs. Furthermore, LLM provides an interpretation of generated results using the quality captioning model. While CLIP-IQA also employs LLMs, its limitation to training one metric at a time with simple text prompts restricts its applicability in complex medical IQA scenarios, especially when assessing fine structures and small lesions.

However, it is acknowledged that there are some limitations of the CT-IQA dataset. First, the relatively small size of the dataset might have made the training process of the LLM sub-optimal. Owing to the high cost and significant time required for professional radiologist annotations, the study aims to validate the feasibility of using LLMs for IQA, serving as a rapid communication to demonstrate the potential of the proposed approach. In the future, it is planned to collect more clinical data to conduct larger-scale experiments and further validate current findings. Although the external evaluation indicates a strong correlation between radiologists, the dataset annotated by a single radiologist still introduces a small bias into the model. In the future, it is planned

to use mean calibration or small-scale fine-tuning to adapt to the preferences of different radiologists. Since IQAGPT represents the initial effort in IQA using LLMs, the reliance on the annotation standards of the prior studies may not fully encompass the complexity of image-quality nuances [28, 34], such as body parts and lesion types. In the future, text descriptions could be significantly refined, and different types of CT images could be added for IQA, thereby broadening its applicability and effectiveness in clinical scenarios.

Conclusions

This study presents a pioneering exploration into CT subjective quality assessment, using an innovative amalgamation of VLMs and ChatGPT. We collected CT-IQA, an image-text dataset comprising pairs of CT scans with quality scores annotated by an experienced radiologist. We develop IQAGPT, fine-tuned on a VLM using the CT-IQA dataset, which can integrate with ChatGPT to generate both quality scores and detailed reports. The results of extensive experiments not only demonstrate the feasibility of IQAGPT but also highlight the effectiveness of LLMs, marking a significant potential of integrating LLMs in the field of subjective IQA.

Abbreviations

LLMs	Large language models
VLMs	Vision-language models
IQA	Image quality assessment
CT	Computed tomography
PSNR	Peak signal-to-noise ratio
SSIM	Structural similarity
RMSE	Root-mean-square error
NDCT	Normal-dose CT
LDCT	Low-dose CT
QRM	Quality reference mAs
BLEU-n	Bilingual evaluation understudy for n words
ROUGE-L	Recall oriented understudy of gisting evaluation for longest common subsequence
METEOR	Metric for evaluation of translation with explicit ordering
CIDEr	Consensus-based image description evaluation
PLCC	Pearson linear correlation coefficient
SROCC	Spearman's rank order correlation coefficient
CLS	Classification
GT	Ground-truth
R1	Radiologist 1
R2	Radiologist 2
R3	Radiologist 3

Acknowledgements

We would like to thank Huiming Li from Southeast University Zhongda Hospital and Sirong Piao from Peking Union Medical College Hospital for external data annotation. We also would like to thank three anonymous reviewers for their insightful comments and constructive suggestions, which significantly improved the quality of this work.

Authors' contributions

GW conceived and designed the study, and revised the paper; ZC conducted experiments, performed the analysis and wrote the paper; BH and YL collected and analyzed the data; CN and TC conducted experiments and revised the paper; HS supervised the entire project, secured funding and revised

the paper. All the authors have read and approved the final version of this manuscript.

Funding

This work was supported in part by the National Natural Science Foundation of China, No. 62101136; Shanghai Sailing Program, No. 21YF1402800; and National Institutes of Health, Nos. R01CA237267, R01HL151561, R01EB031102, and R01EB032716.

Availability of data and materials

For accessing the dataset used in this paper, please contact the corresponding author Hongming Shan.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 22 March 2024 Accepted: 24 July 2024

Published online: 05 August 2024

References

- Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A et al (2024) PaLM: Scaling language modeling with pathways. *J Mach Learn Res* 24(1):240. <https://doi.org/10.48550/arXiv.2204.02311>
- Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T et al (2023) LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv: 2302.13971*. <https://doi.org/10.48550/arXiv.2302.13971>
- Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training. <https://openai.com/index/language-unsupervised/>. Accessed 16 Oct 2023
- Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. <https://d4mucfpxyvw.cloudfront.net/better-language-models/language-models.pdf>. Accessed 16 Oct 2023
- Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P et al (2020) Language models are few-shot learners. In: Proceedings of the 34th international conference on neural information processing systems, Curran Associates Inc., Vancouver, 6-12 December 2020. <https://doi.org/10.48550/arXiv.2005.14165>
- Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P et al (2022) Training language models to follow instructions with human feedback. In: Proceedings of the 36th international conference on neural information processing systems, Curran Associates Inc., New Orleans, 28 November-9 December 2022. <https://doi.org/10.48550/arXiv.2203.02155>
- Christiano PF, Leike J, Brown TB, Martic M, Legg S, Amodei D (2017) Deep reinforcement learning from human preferences. In: Proceedings of the 31st international conference on neural information processing systems, Curran Associates Inc., Long Beach, 4-9 December 2017. <https://doi.org/10.48550/arXiv.1706.03741>
- Wang WH, Bao HB, Dong L, Bjorck J, Peng ZL, Liu Q et al (2023) Image as a foreign language: BEIT pretraining for vision and vision-language tasks. In: Proceedings of the 2023 IEEE/CVF conference on computer vision and pattern recognition, IEEE, Vancouver, 17-24 June 2023. <https://doi.org/10.1109/CVPR52729.2023.01838>
- Li JN, Li DX, Savarese S, Hoi S (2023) BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the 40th international conference on machine learning, JMLR.org, Honolulu, 23-29 July 2023. <https://doi.org/10.48550/arXiv.2301.12597>
- Driess D, Xia F, Sajjadi MSM, Lynch C, Chowdhery A, Ichter B et al (2023) PaLM-E: An embodied multimodal language model. In: Proceedings of the 40th international conference on machine learning, JMLR.org, Honolulu, 23-29 July 2023. <https://doi.org/10.48550/arXiv.2303.03378>
- Wu CF, Yin SM, Qi WZ, Wang XD, Tang ZC, Duan N (2023) Visual ChatGPT: talking, drawing and editing with visual foundation models. *arXiv*. 2303.04671
- Park S, Lee ES, Shin KS, Lee JE, Ye JC (2023) Self-supervised multi-modal training from uncurated image and reports enables zero-shot oversight artificial intelligence in radiology. *arXiv preprint arXiv: 2208.05140*. <https://doi.org/10.1016/j.media.2023.103021>
- Niu C, Wang G (2023) CT multi-task learning with a large image-text (LIT) model. *bioRxiv* 2023.04.06.535859. <https://doi.org/10.1101/2023.04.06.535859>
- Lyu Q, Tan J, Zapadka ME, Ponnatapura J, Niu C, Myers KJ et al (2023) Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. *Vis Comput Ind Biomed Art* 6(1):9. <https://doi.org/10.1186/s42492-023-00136-5>
- OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I et al (2023) Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. <https://doi.org/10.48550/arXiv.2303.08774>
- Zhu DY, Chen J, Shen XQ, Li X, Elhoseiny M (2023) MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv*. 2304.10592
- Chiang WL, Li ZH, Lin Z, Sheng Y, Wu ZH, Zhang H et al (2023) Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. <https://vicuna.lmsys.org>. Accessed 14 Apr 2023
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T et al (2021) An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the 9th international conference on learning representations, ICLR, Vienna, 3-7 May 2021. <https://doi.org/10.48550/arXiv.2010.11929>
- Chow LS, Paramesran R (2016) Review of medical image quality assessment. *Biomed Signal Process Control* 27:145-154. <https://doi.org/10.1016/j.bspc.2016.02.006>
- Sarmah M, Neelima A, Singh HR (2023) Survey of methods and principles in three-dimensional reconstruction from two-dimensional medical images. *Vis Comput Ind Biomed Art* 6(1):15. <https://doi.org/10.1186/s42492-023-00142-7>
- Pack JD, Xu MF, Wang G, Baskaran L, Min J, De Man B (2022) Cardiac CT blooming artifacts: clinical significance, root causes and potential solutions. *Vis Comput Ind Biomed Art* 5(1):29. <https://doi.org/10.1186/s42492-022-00125-0>
- Lei YM, Niu C, Zhang JP, Wang G, Shan HM (2024) CT image denoising and deblurring with deep learning: current status and perspectives. *IEEE Trans Radiat Plasma Med Sci* 8(2):153-172. <https://doi.org/10.1109/TRPMS.2023.3341903>
- Niu C, Wang G (2023) Editorial: advances in deep learning techniques for biomedical imaging. *Vis Comput Ind Biomed Art* 6(1):12. <https://doi.org/10.1186/s42492-023-00139-2>
- Al-Hammuri K, Gebali F, Kanan A, Chelvan IT (2023) Vision transformer architecture and applications in digital health: a tutorial and survey. *Vis Comput Ind Biomed Art* 6(1):14. <https://doi.org/10.1186/s42492-023-00140-9>
- Chen H, Zhang Y, Kalra MK, Lin F, Chen Y, Liao PX et al (2017) Low-dose CT with a residual encoder-decoder convolutional neural network. *IEEE Trans Med Imaging* 36(12):2524-2535. <https://doi.org/10.1109/TMI.2017.2715284>
- Yang QS, Yan PK, Zhang YB, Yu HY, Shi YS, Mou XQ et al (2018) Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. *IEEE Trans Med Imaging* 37(6):1348-1357. <https://doi.org/10.1109/TMI.2018.2827462>
- Shan HM, Zhang Y, Yang QS, Kruger U, Kalra MK, Sun L et al (2018) 3-D convolutional encoder-decoder network for low-dose CT via transfer learning from a 2-D trained network. *IEEE Trans Med Imaging* 37(6):1522-1534. <https://doi.org/10.1109/TMI.2018.2832217>
- Shan HM, Padole A, Homayounieh F, Kruger U, Khera RD, Nitiwarangkul C et al (2019) Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction. *Nat Mach Intell* 1(6):269-276. <https://doi.org/10.1038/s42256-019-0057-9>
- Fu L, De Man B (2022) Deep learning tomographic reconstruction through hierarchical decomposition of domain transforms. *Vis Comput Ind Biomed Art* 5(1):30. <https://doi.org/10.1186/s42492-022-00127-y>
- Gao Q, Li ZL, Zhang JP, Zhang Y, Shan HM (2024) CoreDiff: contextual error-modulated generalized diffusion model for low-dose CT denoising and generalization. *IEEE Trans Med Imaging* 43(2):745-759. <https://doi.org/10.1109/TMI.2023.3320812>

31. Chen ZH, Gao Q, Zhang Y, Shan HM (2023) ASCON: Anatomy-aware supervised contrastive learning framework for low-dose CT denoising. In: Greenspan H, Madabhushi A, Mousavi P, Salcudean S, Duncan J, Syeda-Mahmood T et al (eds) Medical image computing and computer assisted intervention – MICCAI 2023. 26th international conference on medical image computing and computer-assisted intervention, Vancouver, October 2023. Lecture notes in computer science, Springer, Heidelberg, pp 355–365. https://doi.org/10.1007/978-3-031-43999-5_34
32. Chen ZH, Niu C, Gao Q, Wang G, Shan HM (2024) LIT-Former: Linking in-plane and through-plane transformers for simultaneous CT image denoising and deblurring. *IEEE Trans Med Imaging* 43(5):1880–1894. <https://doi.org/10.1109/TMI.2024.3351723>
33. Chen ZH, Chen T, Wang CH, Niu C, Wang G, Shan HM (2024) Low-dose CT denoising with language-engaged dual-space alignment. *arXiv*. 2403.06128
34. Singh S, Kalra MK, Hsieh J, Licato PE, Do S, Pien HH et al (2010) Abdominal CT: comparison of adaptive statistical iterative and filtered back projection reconstruction techniques. *Radiology* 257(2):373–383. <https://doi.org/10.1148/radiol.10092212>
35. Zhai GT, Min XK (2020) Perceptual image quality assessment: a survey. *Sci China Inf Sci* 63(11):211301. <https://doi.org/10.1007/s11432-019-2757-1>
36. Min XK, Gu K, Zhai GT, Liu J, Yang XK, Chen CW (2018) Blind quality assessment based on pseudo-reference image. *IEEE Trans Multimedia* 20(8):2049–2062. <https://doi.org/10.1109/TMM.2017.2788206>
37. Min XK, Ma KD, Gu K, Zhai GT, Wang Z, Lin WS (2017) Unified blind quality assessment of compressed natural, graphic, and screen content images. *IEEE Trans Image Process* 26(11):5462–5474. <https://doi.org/10.1109/TIP.2017.2735192>
38. Min XK, Zhai GT, Gu K, Liu YT, Yang XK (2018) Blind image quality estimation via distortion aggravation. *IEEE Trans Broadcast* 64(2):508–517. <https://doi.org/10.1109/TBC.2018.2816783>
39. Min XK, Gu K, Zhai GT, Yang XK, Zhang WJ, Le Callet P et al (2021) Screen content quality assessment: overview, benchmark, and beyond. *ACM Comput Surv* 54(9):187. <https://doi.org/10.1145/3470970>
40. Min XK, Duan HY, Sun W, Zhu YC, Zhai GT (2024) Perceptual video quality assessment: a survey. *arXiv*. 2402.03413
41. Gao Q, Li S, Zhu MM, Li DY, Bian ZY, Lyu QW et al (2019) Blind CT image quality assessment via deep learning framework. In: Proceedings of the 2019 IEEE nuclear science symposium and medical imaging conference, IEEE, Manchester, 26 October–2 November 2019. <https://doi.org/10.1109/NSS/MIC42101.2019.9059777>
42. Lee W, Cho E, Kim W, Choi H, Beck KS, Yoon HJ et al (2022) No-reference perceptual CT image quality assessment based on a self-supervised learning framework. *Mach Learn: Sci Technol* 3(4):045033. <https://doi.org/10.1088/2632-2153/aca87d>
43. Pouget E, Dedieu V (2023) Comparison of supervised-learning approaches for designing a channelized observer for image quality assessment in CT. *Med Phys* 50(7):4282–4295. <https://doi.org/10.1002/mp.16227>
44. Gao Q, Shan HM, Zeng D (2023) GREAT-IQA: Integrating global perception and local task-specific information for CT image quality assessment. In: Proceedings of the 2023 IEEE international conference on medical artificial intelligence (MedAI), IEEE, Beijing, 18–19 November 2023. <https://doi.org/10.1109/MedAI59581.2023.00059>
45. Wang JY, Chan KCK, Loy CC (2023) Exploring CLIP for assessing the look and feel of images. In: Proceedings of the 37th AAAI conference on artificial intelligence, AAAI, Washington, 7–14 February 2023. <https://doi.org/10.1609/aaai.v37i2.25353>
46. McCollough CH, Bartley AC, Carter RE, Chen BY, Drees TA, Edwards P et al (2017) Low-dose CT for the detection and classification of metastatic liver lesions: results of the 2016 low dose CT grand challenge. *Med Phys* 44(10):e339–e352. <https://doi.org/10.1002/mp.12345>
47. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: human language technologies, ACL, Minneapolis, 2–7 June 2019. <https://doi.org/10.18653/v1/N19-1423>
48. Min XK, Zhai GT, Zhou JT, Farias MCQ, Bovik AC (2020) Study of subjective and objective quality assessment of audio-visual signals. *IEEE Trans Image Process* 29:6054–6068. <https://doi.org/10.1109/TIP.2020.2988148>
49. Min XK, Zhai GT, Gu K, Yang XK (2016) Fixation prediction through multi-modal analysis. *ACM Trans Multimedia Comput, Commun, Appl* 13(1):6. <https://doi.org/10.1145/2996463>
50. Min XK, Zhai GT, Zhou JT, Zhang XP, Yang XK, Guan XP (2020) A multi-modal saliency model for videos with high audio-visual correspondence. *IEEE Trans Image Process* 29:3805–3819. <https://doi.org/10.1109/TIP.2020.2966082>
51. Loshchilov I, Hutter F (2019) Decoupled weight decay regularization. In: Proceedings of the 7th international conference on learning representations, ICLR, New Orleans, 6–9 May 2019. <https://doi.org/10.48550/arXiv.1711.05101>
52. Goyal P, Dollár P, Girshick R, Noordhuis P, Wesolowski L, Kyrola A et al (2017) Accurate, large minibatch SGD: training imagenet in 1 hour. *arXiv*. 1706.02677
53. Loshchilov I, Hutter F (2017) SGDR: Stochastic gradient descent with warm restarts. In: Proceedings of the 5th international conference on learning representations, ICLR, Toulon, 24–26 April 2017. <https://doi.org/10.48550/arXiv.1608.03983>
54. Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of association for computational linguistics, ACL, Philadelphia, 7–12 July 2002. <https://doi.org/10.3115/1073083.1073135>
55. Lin CY (2004) ROUGE: A package for automatic evaluation of summaries. In: Proceedings of the text summarization branches out, ACL, Barcelona, 21–26 July 2004
56. Banerjee S, Lavie A (2005) METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, ACL, Ann Arbor, 29 June 2005
57. Vedantam R, Zitnick C L, Parikh D (2015) CIDEr: Consensus-based image description evaluation. In: Proceedings of the 2015 IEEE conference on computer vision and pattern recognition, IEEE, Boston, 7–12 June 2015. <https://doi.org/10.1109/CVPR.2015.7299087>
58. Moen TR, Chen BY, Holmes III DR, Duan XH, Yu ZC, Yu LF et al (2021) Low-dose CT image and projection dataset. *Med Phys* 48(2):902–911. <https://doi.org/10.1002/mp.14594>
59. Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9(86):2579–2605

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.