CAS-IQA: Teaching Vision-Language Models for Synthetic Angiography Quality Assessment

Bo Wang[1,3*], De-Xing Huang[1,2*], Xiao-Hu Zhou[1,2(✉)], Mei-Jiang Gui[1,2], Nu-Fang Xiao[1,4], Jian-Long Hao[5], Ming-Yuan Liu[6(✉)], and Zeng-Guang Hou[1,2]

[1] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] School of Automation and Electrical Engineering, University of Science and Technology Beijing
[4] School of Computer Science and Engineering, Hunan University of Science and Technology
[5] School of Information, Shanxi University of Finance and Economics
[6] Beijing Friendship Hospital, Capital Medical University
Email: xiaohu.zhou@ia.ac.cn, dr.mingyuanliu@pku.edu.cn

Abstract. Synthetic X-ray angiographies generated by modern generative models hold great potential to reduce the use of contrast agents in vascular interventional procedures. However, low-quality synthetic angiographies can significantly increase procedural risk, underscoring the need for reliable image quality assessment (IQA) methods. Existing IQA models, however, fail to leverage auxiliary images as references during evaluation and lack fine-grained, task-specific metrics necessary for clinical relevance. To address these limitations, this paper proposes CAS-IQA, a vision-language model (VLM)-based framework that predicts fine-grained quality scores by effectively incorporating auxiliary information from related images. In the absence of angiography datasets, CAS-3K is constructed, comprising $3,565$ synthetic angiographies along with score annotations. To ensure clinically meaningful assessment, three task-specific evaluation metrics are defined. Furthermore, a Multi-path featUre fuSion and rouTing (MUST) module is designed to enhance image representations by adaptively fusing and routing visual tokens to metric-specific branches. Extensive experiments on the CAS-3K dataset demonstrate that CAS-IQA significantly outperforms state-of-the-art IQA methods by a considerable margin.
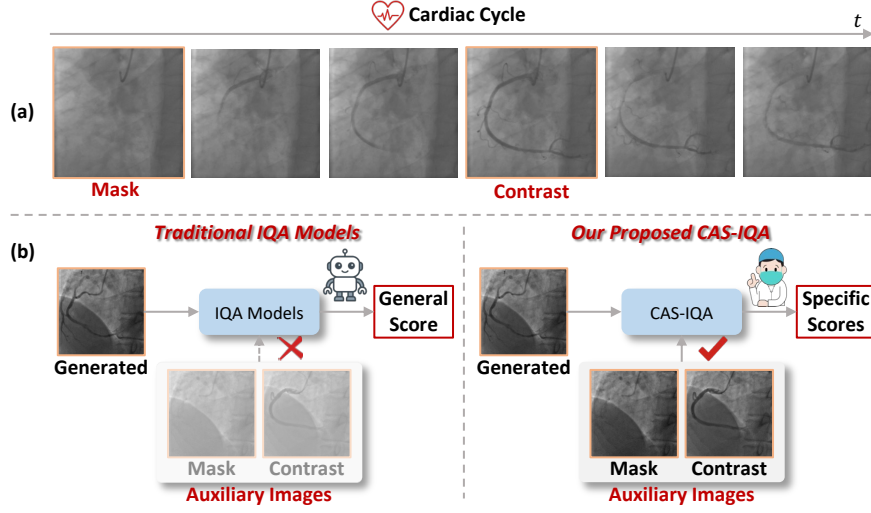
Keywords: X-ray Angiography · Image Quality Assessment · Vision-Language Models.

1 Introduction

X-ray angiography is regarded as the "gold standard" for diagnosing and treating cardiovascular diseases [32]. However, contrast agents used in these procedures

---

[*] Equal contributions.

Fig. 1. (a) An X-ray sequence in real-world clinical scenarios. (b) Comparison between current IQA models and our CAS-IQA.

pose significant health risks to patients, including allergic reactions [1] and renal dysfunction [2]. Recent advances in AI-generated content (AIGC) [18] have demonstrated remarkable capabilities in generating photo-realistic angiographies from non-contrast X-ray images [6], presenting a promising way to reduce reliance on contrast agents in clinical scenarios. Nevertheless, *the fidelity of such synthetic angiographies to essential clinical constraints remains unquantified.* This highlights the urgent need for rigorous quality assessment methods.

Image quality assessment (IQA) is a fundamental task in computer vision [25]. It aims to bridge quantitative mappings between images and corresponding human subjective judgments [15]. Modern IQA methods can achieve strong quality assessment performance on natural images [21,27,25,30], but they falter in our specialized domain for two reasons: **1) Inability to Leverage Auxiliary Images.** Fig. 1 (a) presents a real X-ray sequence in clinical practice. The first image without contrast agents is defined as **Mask**, which is used to generate angiographies (**Generated**) by generative models. The image full with contrast agents is called **Contrast**[7]. Obviously, it is essential to incorporate both **Mask** and **Contrast** when assessing the quality of **Generated**. However, current IQA models typically operate on **Generated** and lack mechanisms to effectively fuse multi-input information; **2) Lack of Fine-grained, Task-specific Metrics.** Current IQA methods often produce a single, abstract quality score, which is insufficient for clinical applications. In practice, synthetic angiographies are primarily used to visualize vascular structures, and radiologists are particularly

---

[7] Note that **Mask** and **Generated** are captured at different moments in the cardiac cycle. The vessels may have undergone slight deformation due to the heartbeat.

concerned with the structural integrity and anatomical accuracy of these vessels. Therefore, a more nuanced assessment framework is needed.

To address these drawbacks, **CAS-IQA** (**C**ontrast-free **A**ngiography **S**ynthesis **I**mage **Q**uality **A**ssessment) is proposed, a novel framework for evaluating the quality of synthetic angiographies. Given the lack of available datasets, CAS-3K is constructed, a high-quality benchmark dataset comprising over $3,500$ synthetic angiographies generated by five state-of-the-art generative models. To comprehensively evaluate synthetic angiographies, three metrics verified by board-certified radiologists are introduced: **1) Vessel Morphology Consistency (VMC)** assesses the structural consistency between **Generated** and **Contrast**; **2) Vessel Branch Detection (VBD)** evaluates whether key vascular branches are correctly generated in **Generated**; **3) Overall Quality (OQ)** considers not only the first two metrics, but also visual defects such as background artifacts or shadowing in **Generated**.

Based on the CAS-3K dataset, CAS-IQA is constructed. Specifically, CAS-IQA is built upon vision-language models (VLMs), as VLMs allow flexible evaluation of different aspects of images through prompt engineering, better aligning with human subjective judgments. Moreover, a <u>M</u>ulti-path feat<u>U</u>re fu<u>S</u>ion and rou<u>T</u>ing (MUST) module is presented, which fuses features of **Mask**, **Contrast**, and **Generated**, and routes them to metric-specific branches for optimized assessment.

Our main contributions are summarized as follows:

- The first high-quality dataset for angiography IQA (CAS-3K) is constructed, consisting of $3,565$ synthetic angiographies with metric-based annotations.
- A novel IQA framework based on vision-language models is proposed to comprehensively evaluate the quality of synthetic angiographies. Moreover, the MUST module is designed to effectively fuse visual cues from multiple image sources and adapt them to different evaluation metrics.
- Extensive experimental results indicate that CAS-IQA significantly outperforms state-of-the-art IQA models on the CAS-3K dataset.

2 Related Works

2.1 Image Quality Assessment

As a fundamental task in computer vision, image quality assessment (IQA) has undergone substantial advancements across diverse applications. Early methods primarily rely on human-crafted metrics such as NIQE [16] and BRISQUE [15], *etc.* However, these approaches heavily depend on prior knowledge and often exhibit limited generalization capabilities [30]. In recent years, the rapid development of deep neural networks (DNNs) has introduced robust solutions for IQA [21,27,28]. Typically formulated as a regression problem, DNN-based IQA methods extract visual features from input images using backbone encoders, and then predict corresponding quality scores. Thanks to strong feature extraction

capabilities of advanced network architectures, these models have achieved impressive performance on standard IQA benchmarks. Nonetheless, they struggle to capture fine-grained semantic content of images, which is crucial for quality assessment [24]. To address this limitation, recent studies have explored the use of vision-language models (VLMs) for IQA [26,25]. By leveraging the advanced multimodal reasoning capabilities of VLMs, these methods have demonstrated substantial progress in aligning quality predictions with human subjective judgments. However, as discussed in Section 1, existing approaches fall short in our specific domain.

2.2   Vision-Language Models

Vision-language models (VLMs) extend large language models (LLMs) by incorporating visual instruction tuning [11], which enables them to jointly process and align visual and textual modalities. This alignment facilitates the development of visual reasoning and understanding capabilities, allowing VLMs to follow multimodal instructions and perform complex vision-language tasks in a unified framework (*e.g.,* LLaVA [11] and mPLUG-Owl2 [29]), including visual question answering (VQA) [12] and image captioning [13], *etc*. In this paper, the proposed CAS-IQA builds upon the powerful visual reasoning capabilities of VLMs to achieve more accurate and human-aligned IQA.

### 2.3   Datasets for Image Quality Assessment

High-quality IQA datasets are the foundation for developing and benchmarking effective IQA models. Considerable efforts have been devoted to constructing diverse IQA datasets spanning both natural and AIGC. Traditional IQA datasets, such as KonIQ-10K [4] and SPAQ [3], primarily focus on natural images captured under a wide range of real-world conditions, offering a broad spectrum of authentic distortions and degradation types. With the rapid emergence of advanced generative models, there has been a significant rise in AIGC, leading to the need for specialized datasets and methodologies to assess the perceptual quality of synthetic images. In response, datasets such as AGIQA-3K [9] and AIGIQA-20K [10] have been proposed. To the best of our knowledge, CAS-3K is the first IQA dataset for medical AIGC.

3   CAS-3K Dataset

3.1   Image Collection and Scoring Rules

Image Collection. To construct CAS-3K, we curate high-quality synthetic angiographies using two publicly available datasets: XCAD [14] and CADICA [7]. After thorough manual screening, 713 high-quality **Mask-Contrast** image pairs are selected from the two datasets. Using **Mask** as input, we apply five state-of-the-art generative models[8] to synthesize angiographies (**Generated**). This

---

[8] Three GAN-based models (*i.e.,* CycleGAN [35], AttentionGAN [22], and CUT [17]) and two diffusion-based models (*i.e.,* ILVR [33] and EGSDE [34]).
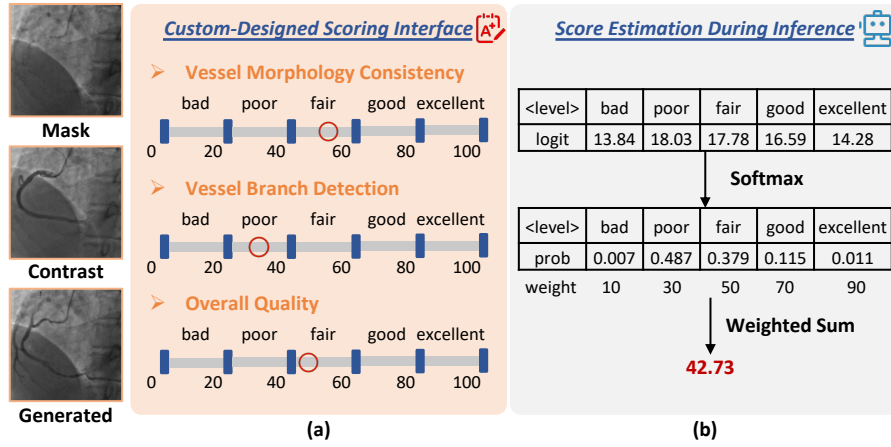
Fig. 2. An overview of the rating GUI of CAS-3K and sore estimation rules during inference. (a) illustrates how five-level scores are assigned to rate the three quality metrics of **Generated** while using **Mask** and **Contrast** as references. (b) demonstrates how the model-predicted level logits are converted into final quality scores.

resulted in a total of $3,565$ **Generated**, each paired with its corresponding **Mask** and **Contrast**.

Scoring Rules. To ensure clinically relevant quality assessment, we collaborat with two board-certified radiologists to define three task-specific evaluation metrics:

1) **Vessel Morphology Consistency (VMC):** It assesses the anatomical accuracy of major vessels in **Generated** compared to **Contrast**. Key factors include the morphological fidelity of primary vessel structures and their continuity—specifically the absence of interruptions, or breaks.
2) **Vessel Branch Detection (VBD):** It assesses the correctness of generated vascular branches. This includes evaluating the number of branches as well as the accuracy of their orientations and trajectories, as compared to those in **Contrast**.
3) **Overall Quality (OQ):** This metric provides a holistic assessment by balancing VMC and VBD, while also accounting for visual artifacts such as background noise, shadowing, or unrealistic textures. This metric reflects the overall anatomical plausibility and visual realism of **Generated**.

To support fine-grained evaluation, a five-level rating scheme is adopted inspired by the Q-Align [25]. Each image is scored on a continuous $0 \sim 100$ scale, partitioned into five semantic categories: bad, poor, fair, good, and excellent.

3.2 Subjective Quality Assessment

Following ITU-R BT.500 [19], three graduate students majoring in biomedical engineering are recruited to participate in subjective evaluation experiments.
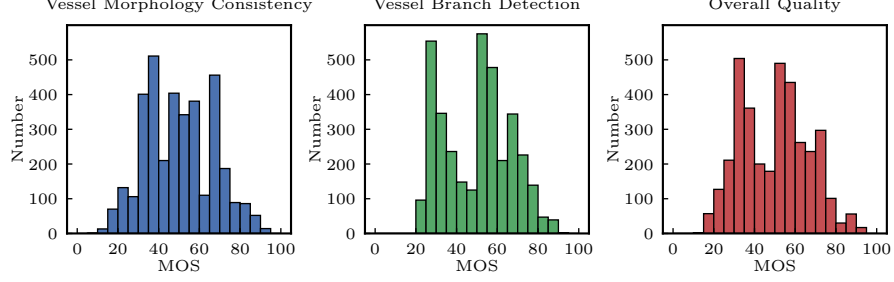
Fig. 3. Distributions of mean opinion score (MOS).

Prior to annotation, the participants undergo training using the scoring guidelines described in Section 3.1. After that, they are required to annotate 200 test samples. To assess the reliability of their evaluations, two correlation metrics are calculated: the Pearson linear correlation coefficient (PLCC) and the Spearman rank correlation coefficient (SRCC). The formal annotation phase begins only after the inter-rater correlations exceeds a threshold of 0.7.

During the formal annotation process, participants use a custom-designed scoring interface. As shown in Fig. 2 (a), each session displays three images simultaneously on a 27-inch monitor with a resolution of 1920×1080. Participants rate each image based on the three predefined quality metrics. To mitigate fatigue and maintain annotation quality, they are suggested to take breaks every 30 minutes. In total, $3,565 \times 3 \times 3 = 32,085$ subject ratings are collected, where $3,565$, 3, and 3 represent the number of synthetic images, evaluation metrics, and subjects.

3.3   Subjective Data Processing

First, the outlier rejection procedure recommended by ITU-R BT.500 [19] are adopted to eliminate unreliable annotations. No outliers were detected and excluded in our case. To mitigate inter-observer variability in perceived image quality, per-subject score normalization is performed:

$$Z_{ij} = \frac{S_{ij} - \mu_i}{\sigma_i} \tag{1}$$

where S_{ij} denotes the raw score assigned by the i-th subject to the j-th image, μ_i is the mean of scores and σ_i is the standard deviation of scores from the i-th subject. Next, Z_{ij} is linearly rescaled to $[0, 100]$ to obtain $\hat{Z}_{ij}$. Finally, the mean opinion score (MOS) is calculated following Eq. (2), and the distributions of MOS are shown in Fig. 3.

$$MOS_j = \frac{1}{N} \sum_{i=1}^{N} \hat{Z}_{ij} \tag{2}$$

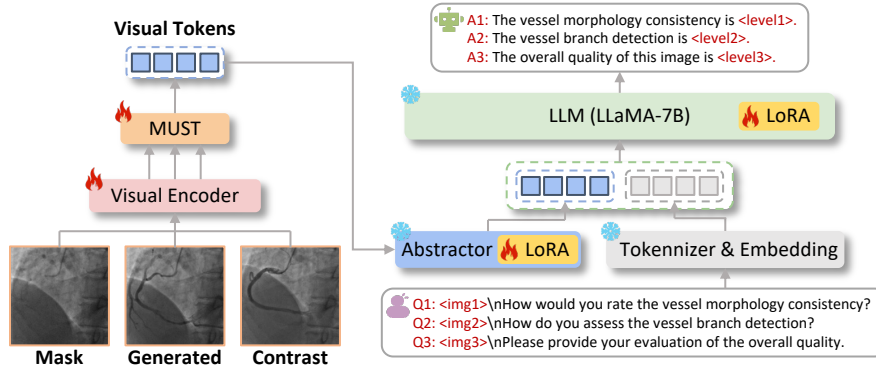where N is the number of subjects.

Fig. 4. Overall framework of CAS-IQA.

4 CAS-IQA Model

4.1 Overall Framework

As illustrated in Fig. 4, CAS-IQA takes three images as input: a generated angiography (**Generated**), a non-contrast image (**Mask**), and a real angiography (**Contrast**). These images are first processed by a vision encoder to extract corresponding visual tokens. In parallel, manually crafted text prompts are tokenized and embedded into textual tokens. The proposed Multi-path featUre fuSion and rouTing (MUST) module then fuses the visual tokens from the three input images and routes them into three specialized branches, each tailored to evaluate a specific quality metric. The fused visual tokens are subsequently processed by a vision abstractor, which compresses them using learnable tokens. Finally, the visual and textual tokens are jointly fed into a large language model (LLM) to generate quality assessment predictions.

4.2 Multi-Path Feature Fusion and Routing Module

Experienced radiologists tend to focus on different regions of **Mask**, **Generated**, and **Contrast** depending on the quality metric being assessed. For VMC and VBD, attention is primarily directed toward vessels in **Generated** and **Contrast**. Specifically, VMC emphasizes the global structural alignment of vessels, whereas VBD requires attention to finer details, such as vessel bifurcations. In contrast, OQ involves a comprehensive evaluation of the overall consistency across all three images. Motivated by these observations, a Multi-path featUre fuSion and rouTing (MUST) module is proposed, which enables fine-grained fusion and adaptive routing of visual tokens for each metric. The architecture of MUST is illustrated in Fig. 5. For the VMC and VBD branches, we first extract vessel tokens by subtracting **Contrast** and **Mask** tokens. Since **Contrast** and **Mask** are not temporally aligned, deformable convolution is employed to align
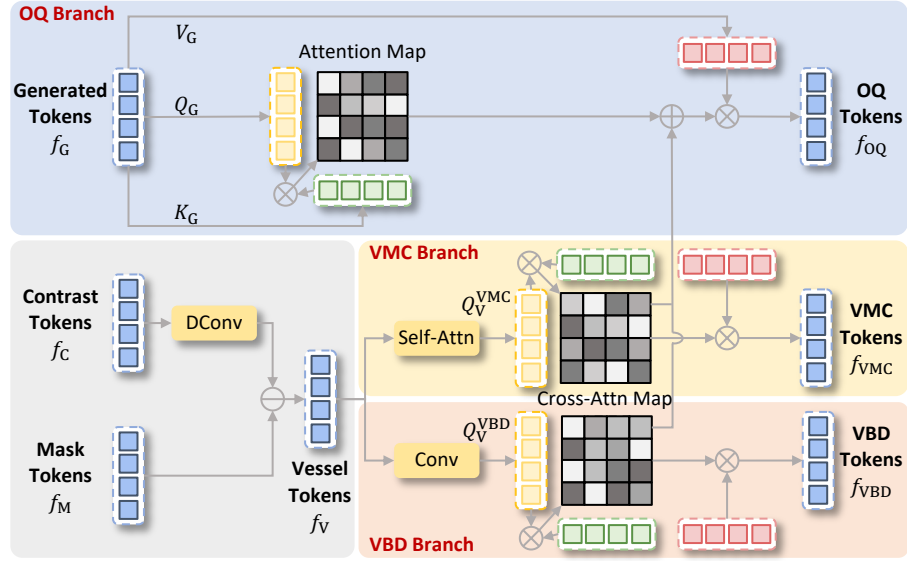
Fig. 5. Details of the MUST module.

Contrast tokens before subtraction:

$$f_{\mathrm{V}} = \mathrm{DConv}\left(f_{\mathrm{C}}\right) - f_{\mathrm{M}} \tag{3}$$

where f_{V} represents vessel tokens. f_{C} and f_{M} denote **Contrast** and **Mask** tokens, respectively. $\mathrm{DConv}\left(\cdot\right)$ means deformable convolution.

VMC Branch. To evaluate global structural consistency, a self-attention mechanism [23] is applied to vessel tokens f_{V} to obtain global vessel representations. These representations serve as queries in a cross-attention mechanism with **Generated** tokens, producing VMC-specific fused tokens f_{VMC}.

$$A_{\mathrm{VMC}} = \mathrm{Softmax}\left(\frac{Q_{\mathrm{V}}^{\mathrm{VMC}} K_{\mathrm{G}}^{T}}{\sqrt{d}}\right), f_{\mathrm{VMC}} = A_{\mathrm{VMC}} V_{\mathrm{G}} \tag{4}$$

where A_{VMC} is cross-attention maps. $Q_{\mathrm{V}}^{\mathrm{VMC}} = W_{\mathrm{Q}}^{\mathrm{VMC}} \mathrm{Self\text{-}Attn}\left(f_{\mathrm{V}}\right)$ indicate VMC queries. $W_{\mathrm{Q}}^{\mathrm{VMC}}$ and $\mathrm{Self\text{-}Attn}(\cdot)$ represent learnable projection matrices and self-attention, respectively. Keys K_{G} and values V_{G} are from **Generated** tokens.

VBD Branch. For VBD, which emphasizes local branching structures, fine-grained tokens are extracted from f_{V} using a convolutional block. These local tokens are then used to compute cross-attention with **Generated** tokens, producing VBD tokens f_{VBD}:

$$A_{\mathrm{VBD}} = \mathrm{Softmax}\left(\frac{Q_{\mathrm{V}}^{\mathrm{VBD}} K_{\mathrm{G}}^{T}}{\sqrt{d}}\right), f_{\mathrm{VBD}} = A_{\mathrm{VBD}} V_{\mathrm{G}} \tag{5}$$

where VBD queries $Q_{\mathrm{V}}^{\mathrm{VBD}} = W_{\mathrm{Q}}^{\mathrm{VBD}}\mathrm{Conv}\,(f_{\mathrm{V}})$. $\mathrm{Conv}\,(\cdot)$ is the convolutional block.

OQ Branch. Finally, to evaluate OQ, we begin by computing a self-attention map from **Generated** tokens to capture overall perceptual quality. This is then fused with the attention maps from the VMC and VBD branches using a set of learnable scalar weights α, β and γ, constrained such that $\alpha + \beta + \gamma = 1$. The final attention map is applied to **Generated** values to produce OQ tokens:

$$f_{\mathrm{OQ}} = \left\{ \alpha \left[\mathrm{Softmax}\left(\frac{Q_{\mathrm{G}}K_{\mathrm{G}}^{T}}{\sqrt{d}} \right) \right] + \beta A_{\mathrm{VMC}} + \gamma A_{\mathrm{VBD}} \right\} V_{\mathrm{G}} \tag{6}$$

4.3 Mapping Between Levels and Scores

Continuous Scores to Discrete Levels. Considering that humans tend to describe image quality in categorical levels rather than precise numerical scores [25], a five-level rating scheme is adopted. The continuous score range of $0 \sim 100$ is evenly partitioned into five intervals corresponding to the levels: {bad, poor, fair, good, excellent}. Each continuous score is mapped to its corresponding discrete level based on these intervals. These discrete labels are then used to construct the instruction–response pairs described in Section 4.4 for supervised training.

Score Estimation During Inference. As illustrated in Fig. 2 (b), a softmax function is applied over the logits associated with the five predefined level tokens to obtain a probability distribution. For each quality level, a weight equal to the midpoint of its associated score interval is assigned. The final predicted quality score is computed as the weighted sum of these probabilities.

4.4 Instruction Tuning

Instruction tuning has proven effective in unlocking the potential of VLMs for downstream tasks, and the design of high-quality user prompts is crucial for supervised fine-tuning [11]. To improve model performance on the X-ray angiography IQA task, we design a series of prompts framed as multi-turn question–answer interactions.

Vessel Morphology Consistency:
#User: `<img1>`As an experienced interventional radiologist, how would you rate the vessel morphology consistency of this image?
#Assistant: The vessel morphology consistency of this image is `<level1>`.
Vessel Branch Detection:
#User: `<img2>`From your perspective as an interventional radiologist, how do you assess the vessel branch detection in this image?
#Assistant: The vessel branch detection of this image is `<level2>`.
Overall Quality:
#User: `<img3>`With your experience in interventional radiology, please provide your evaluation of the overall quality of this image.
#Assistant: The overall quality of this image is `<level3>`.

Table 1. Comparisons with SOTA methods on the CAS-3K dataset. VMC, VBD, and OQ denote vessel morphology consistency, vessel branch detection, and overall quality, respectively. The best and second results are highlighted in red and blue, respectively.

Category	Methods	VMC		VBD		OQ	
		PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Handcrafted	NIQE [16] [SPL'12]	0.0520	0.0438	0.0289	0.0256	0.0337	0.0202
	BRISQUE [15] [TIP'12]	0.1876	0.1901	0.2359	0.2391	0.2224	0.2139
DNN-based	DBCNN [31] [TCSVT'20]	0.4990	0.4853	0.5856	0.5647	0.4745	0.4666
	HyperIQA [21] [CVPR'20]	0.6657	0.6356	0.5942	0.5873	0.6339	0.6216
	AHIQ [8] [CVPR'22]	0.6251	0.6434	0.6050	0.6010	0.6654	0.6621
	ManIQA [28] [CVPR'22]	0.6823	0.6759	0.6089	0.6060	0.6778	0.6689
	QCN [20] [CVPR'24]	0.6750	0.6651	0.6308	0.6235	0.6796	0.6765
	LoDa [27] [CVPR'24]	0.6859	0.6664	0.6353	0.6279	0.6899	0.6737
VLM-based	Q-Align [25] [ICLR'24]	0.6030	0.5907	0.5683	0.5651	0.6096	0.5999
	MA-AGIQA [24] [MM'24]	0.6892	0.6625	0.6507	0.6314	0.6696	0.6598
	CAS-IQA (Ours)	0.6925	0.6855	0.6639	0.6605	0.6985	0.6986

Note that the visual tokens `<img1>`, `<img2>`, and `<img3>` correspond to the outputs of the VMC, VBD, and OQ branches, respectively. The predicted discrete quality levels for the VMC, VBD, and OQ metrics are represented as `<level1>`, `<level2>`, and `<level3>`, respectively. Following Q-Align [25], the user queries are randomly chosen from a group of paraphrases as augmentation.

5   Experiments

5.1   Implementation Details

Model Configurations. Our CAS-IQA is built upon mPLUG-Owl2 [29], consisting of a pretrained CLIP image encoder (ViT-L), a visual abstractor (Q-Former), and a LLM (LLaMA-2-7B). Due to computational constraints and the limited size of CAS-3K, both the visual abstractor and the LLM are fine-tuned using low-rank adaptation (LoRA) [5]. The LoRA configuration uses a rank of 128 and an alpha value of 256.

Training Details. CAS-3K is partitioned into a training set of $2,850$ images and a test set of 715 images. Training is performed for 5 epochs on two NVIDIA A6000 GPUs with a batch size of 16. CAS-IQA is optimized using the AdamW optimizer with a learning rate of $2e^{-5}$, scheduled via cosine annealing.

Evaluation Metrics. In line with prior work [25,30], we evaluate model performance using the Pearson linear correlation coefficient (PLCC) and the Spearman rank-order correlation coefficient (SRCC). Higher PLCC and SRCC values indicate stronger alignment between the predicted quality scores and human subjective judgments.
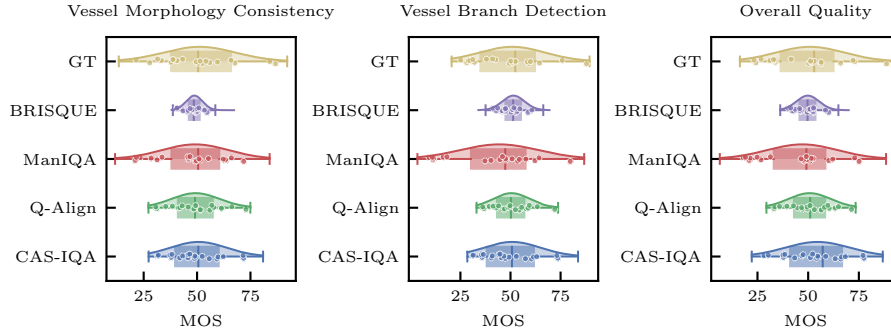
Fig. 6. MOS distributions predicted by different methods.

5.2 Comparisons with SOTA Methods

Quantitative Results. To demonstrate the effectiveness of CAS-IQA, we compare it against a comprehensive set of baselines, including two handcrafted IQA methods, six DNN-based approaches, and two VLM-based methods. Since these baseline models are designed to predict a single quality score, each of them are separately trained for the three metrics. Quantitative results is summarized in Table 1. Our main observations are as follows: **1)** Handcrafted methods (*i.e.,* NIQE [16] and BRISQUE [15]) perform poorly on the CAS-3K dataset. These methods rely on handcrafted statistical features optimized for natural image statistics. The assumptions about luminance distribution, texture, and contrast do not hold in the domain of X-ray angiographies. Such methods lack the representational capacity to capture the anatomical structures and modality-specific patterns inherent in synthetic angiographies. Consequently, the poor performance of these methods is expected; **2)** Our CAS-IQA achieves superior results compared to all baseline methods. Notably, it achieves PLCC improvements of 0.5%, 2.0%, and 4.1% for VMC, VBD, and OQ, respectively, over the second-best method, MA-AGIQA [24]. These results are consistent with our design rationale. Baseline models typically rely solely on **Generated** and therefore lack critical auxiliary information. In contrast, CAS-IQA integrates visual cues from multiple inputs and adaptively routes metric-specific visual cues to dedicated branches. This enables CAS-IQA to perform more accurate quality assessment.

Qualitative Analysis. Fig. 6 illustrates the distributions of MOS predicted by four representative methods across the three evaluation metrics. We present a detailed analysis as follows: **1)** The MOSs predicted by handcrafted methods tends to concentrate within a narrow range. This is largely attributed to the inherent domain gap between natural images. Unlike natural images, which typically contain rich textures and diverse color distributions, X-ray angiographies are structurally homogeneous and feature-sparse. As a result, these approaches fail to capture meaningful variations in image quality, mistakenly interpreting all inputs as perceptually similar; **2)** DNN-based methods, such as ManIQA [28], exhibit similar distribution patterns across three metrics, primarily due to their reliance

Table 2. Quantitative results of ablation study. The best results are highlighted in **bold**.

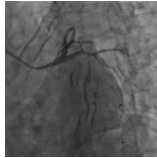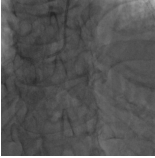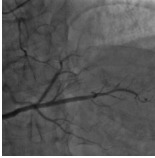| MUST	VMC		VBD		OQ	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
✗	0.6517	0.6636	0.6451	0.6491	0.6830	0.6750
✓	**0.6925**	**0.6855**	**0.6639**	**0.6605**	**0.6985**	**0.6986**
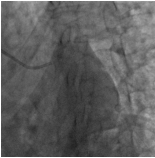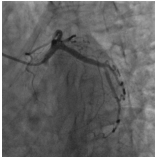
solely on human-annotated supervision without the benefit of prompt-driven guidance as seen in VLM-based approaches; **3)** Previous VLM-based methods, such as Q-Align [25], give score predictions only based on **Generated**, without incorporating auxiliary structural cues in **Mask** and **Contrast**. This lack of contextual information limits their semantic grounding and impairs their ability to assess image quality in a clinically meaningful way. In contrast, CAS-IQA, equipped with MUST module, effectively fuses and routes multi-source image features, enabling more accurate and metric-specific quality assessment.

5.3   Ablation Study

Quantitative Results. The effectiveness of the proposed MUST module is evaluated in Table 2. In the baseline configuration (w/o MUST), a simplified setup is used in which a single prompt processes visual tokens from **Mask**, **Generated**, and **Contrast** without specialized fusion or routing.

> **#User:** `<imgG>`As an experienced interventional radiologist, how would you rate the quality of this image? Contrast-free image:`<imgM>`, Contrast imgae:`<imgC>`
> **#Assistant:** The vessel morphology consistency of this image is `<level1>`. The vessel branch detection rate of this image is `<level2>`. The overall quality of this image is `<level3>`.

As shown in the results, incorporating MUST leads to consistent improvements across all three metrics. These improvements are primarily attributed to two key mechanisms within the MUST module: **1) Multi-Source Visual Fusion:** By explicitly fusing visual tokens from multiple images, MUST enables context-aware feature representation. This is particularly beneficial for metrics like OQ, which require a comprehensive assessment of visual coherence across inputs; **2) Metric-Specific Routing:** MUST introduces dedicated evaluation branches, allowing each metric to receive features aligned with its semantic requirements. For example, global vessel morphology for VMC and localized bifurcations for VBD. This targeted routing enhances metric-specific prediction accuracy. In addition to performance improvements, we also observed that CAS-IQA equipped with the MUST module achieves optimal performance within only 5 training epochs, whereas the baseline configuration (w/o MUST) requires around 15 epochs to converge. This acceleration in convergence can be attributed to several factors: First, multi-source fusion enables early-stage feature representations

Metric	w/o MUST	w/ MUST	GT
VMC	50.16	47.31	41.45
VBD	50.19	50.87	43.30
OQ	62.91	53.19	44.26

Metric	w/o MUST	w/ MUST	GT
VMC	24.75	19.98	17.89
VBD	29.79	28.34	28.93
OQ	16.48	24.02	23.30

Mask   **Contrast**   **Generated**

Fig. 7. Two examples of CAS-IQA (w/ and w/o MUST).

to be more discriminative and semantically informative, reducing the learning burden. Second, the metric-specific routing mechanism introduces task-specific inductive bias, allowing the model to optimize each evaluation metric independently and efficiently.

Qualitative Analysis. Fig. 7 presents the quality score predictions of CAS-IQA (w/ and w/o MUST) on two representative examples. In both examples, integrating the MUST module enables CAS-IQA to produce predictions that are consistently closer to human subjective judgments (*i.e.,* GT). However, we observe that the improvements brought by MUST in VBD are less pronounced compared to those in VMC and OQ. One possible explanation is that **Generated** often contain isolated or anatomically implausible vessel branches. Such artifacts are visually distinct and can be detected directly without the need for additional contextual information. This indicates that VBD, being more anomaly-sensitive, may rely more on visual saliency within **Generated** than on broader contextual consistency. In contrast, VMC and OQ require a holistic understanding of vessel morphology and global image coherence. Accurate assessment of these metrics benefits from integrating structural information across multiple images. The design of the MUST module facilitates this process by enabling multi-source feature integration and metric-specific representation learning.

6   Conclusion

This paper focuses on image quality assessment (IQA) for synthetic X-ray angiographies. To enable clinically meaningful evaluation, a curated dataset, CAS-3K, is constructed, and a unified framework, CAS-IQA, is proposed. By leveraging the powerful vision-language reasoning capabilities of VLMs, CAS-IQA can achieve strong alignment with human subjective judgments. In addition, the proposed MUST module enhances performance by adaptively fusing image tokens and routing them to metric-specific branches. Experimental results indicate that CAS-IQA achieves the state-of-the-art performance. This work unlocks the

potential of VLMs in accurate IQA for synthetic X-ray angiographies. Future work will focus on improving the interpretability of CAS-IQA to facilitate its deployment in real-world clinical applications.

Acknowledgements

References

1. Clement, O., et al.: Immediate hypersensitivity to contrast agents: The French 5-year CIRTACI study. Lancet Discovery Science **1**, 51–61 (2018)
2. Fähling, M., Seeliger, E., Patzak, A., Persson, P.B.: Understanding and preventing contrast-induced acute kidney injury. Nature Reviews Nephrology **13**(3), 169–180 (2017)
3. Fang, Y., Zhu, H., Zeng, Y., Ma, K., Wang, Z.: Perceptual quality assessment of smartphone photography. In: Proc. CVPR. pp. 3677–3686 (2020)
4. Hosu, V., Lin, H., Sziranyi, T., Saupe, D.: KonIQ-10K: An ecologically valid database for deep learning of blind image quality assessment. IEEE TIP **29**, 4041–4056 (2020)
5. Hu, E.J., et al.: LoRA: Low-rank adaptation of large language models. In: Proc. ICLR (2022)
6. Huang, D.X., et al.: CAS-GAN for contrast-free angiography synthesis. In: Proc. SSCI (2025)
7. Jiménez-Partinen, A., et al.: CADICA: A new dataset for coronary artery disease detection by using invasive coronary angiography. Expert Systems **41**(12), e13708 (2024)
8. Lao, S., et al.: Attentions help CNNs see better: Attention-based hybrid image quality assessment network. In: Proc. CVPR. pp. 1140–1149 (2022)
9. Li, C., et al.: AGIQA-3K: An open database for AI-generated image quality assessment. IEEE TCSVT **34**(8), 6833–6846 (2023)
10. Li, C., et al.: AIGIQA-20K: A large database for ai-generated image quality assessment. In: Proc. CVPR. pp. 6327–6336 (2024)
11. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Proc. NeurIPS. vol. 36, pp. 34892–34916 (2023)
12. Liu, Y., et al.: MMBench: Is your multi-modal model an all-around player? In: Proc. ECCV. pp. 216–233 (2024)
13. Lu, P., et al.: FINECAPTION: Compositional image captioning focusing on wherever you want at any granularity. In: Proc. CVPR (2025)

14. Ma, Y., et al.: Self-supervised vessel segmentation via adversarial learning. In: Proc. ICCV. pp. 7536–7545 (2021)
15. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE TIP **21**(12), 4695–4708 (2012)
16. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE SPL **20**(3), 209–212 (2012)
17. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: Proc. ECCV. pp. 319–345 (2020)
18. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proc. CVPR. pp. 10684–10695 (2022)
19. Series, B.: Methodology for the subjective assessment of the quality of television pictures. Recommendation ITU-R BT **500**(13) (2012)
20. Shin, N.H., Lee, S.H., Kim, C.S.: Blind image quality assessment based on geometric order learning. In: Proc. CVPR. pp. 12799–12808 (2024)
21. Su, S., et al.: Blindly assess image quality in the wild guided by a self-adaptive hyper network. In: Proc. CVPR. pp. 3667–3676 (2020)
22. Tang, H., Liu, H., Xu, D., Torr, P.H., Sebe, N.: AttentionGAN: Unpaired image-to-image translation using attention-guided generative adversarial networks. IEEE TNNLS **34**(4), 1972–1987 (2021)
23. Vaswani, A., et al.: Attention is all you need. Proc. NeurIPS (2017)
24. Wang, P., et al.: Large multi-modality model assisted AI-generated image quality assessment. In: Proc. MM. pp. 7803–7812 (2024)
25. Wu, H., et al.: Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In: Proc. ICML. pp. 54015–54029 (2024)
26. Wu, H., et al.: Q-Instruct: Improving low-level visual abilities for multi-modality foundation models. In: Proc. CVPR. pp. 25490–25500 (2024)
27. Xu, K., et al.: Boosting image quality assessment through efficient transformer adaptation with local feature enhancement. In: Proc. CVPR. pp. 2662–2672 (2024)
28. Yang, S., et al.: MANIQA: Multi-dimension attention network for no-reference image quality assessment. In: Proc. CVPR. pp. 1191–1200 (2022)
29. Ye, Q., et al.: mPLUG-Owl2: Revolutionizing multi-modal large language model with modality collaboration. In: Proc. CVPR. pp. 13040–13051 (2024)
30. You, Z., Cai, X., Gu, J., Xue, T., Dong, C.: Teaching large language models to regress accurate image quality scores using score distribution. In: Proc. CVPR (2025)
31. Zhang, W., Ma, K., Yan, J., Deng, D., Wang, Z.: Blind image quality assessment using a deep bilinear convolutional neural network. IEEE TCSVT **30**(1), 36–47 (2020)
32. Zhao, H., et al.: Large-scale pretrained frame generative model enables real-time low-dose DSA imaging: An AI system development and multi-center validation study. Med (2024), DOI: 10.1016/j.medj.2024.07.025
33. Zhao, M., Bao, F., Li, C., Zhu, J.: ILVR: Conditioning method for denoising diffusion probabilistic models. In: Proc. ICCV. pp. 14367–14376 (2021)
34. Zhao, M., Bao, F., Li, C., Zhu, J.: EGSDE: Unpaired image-to-image translation via energy-guided stochastic differential equations. In: Proc. NeurIPS. vol. 35, pp. 3609–3623 (2022)
35. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proc. ICCV. pp. 2223–2232 (2017)