

Enhancing Image Quality Assessment Ability of LMMs via Retrieval-Augmented Generation

Kang Fu¹, Huiyu Duan¹, Zicheng Zhang¹, Yucheng Zhu¹, Jun Zhao², Xiongkuo Min¹, Jia Wang¹, and Guangtao Zhai¹

¹ Shanghai Jiao Tong University, ²Tencent

Abstract—Large Multimodal Models (LMMs) have recently shown remarkable promise in low-level visual perception tasks, particularly in Image Quality Assessment (IQA), demonstrating strong zero-shot capability. However, achieving state-of-the-art performance often requires computationally expensive fine-tuning methods, which aim to align the distribution of quality-related token in output with image quality levels. Inspired by recent training-free works for LMM, we introduce IQARAG, a novel, training-free framework that enhances LMMs’ IQA ability. IQARAG leverages Retrieval-Augmented Generation (RAG) to retrieve some semantically similar but quality-variant reference images with corresponding Mean Opinion Scores (MOSs) for input image. These retrieved images and input image are integrated into a specific prompt. Retrieved images provide the LMM with a visual perception anchor for IQA task. IQARAG contains three key phases: Retrieval Feature Extraction, Image Retrieval, and Integration & Quality Score Generation. Extensive experiments across multiple diverse IQA datasets, including KADID, KonIQ, LIVE Challenge, and SPAQ, demonstrate that the proposed IQARAG effectively boosts the IQA performance of LMMs, offering a resource-efficient alternative to fine-tuning for quality assessment.

Index Terms—Image quality assessment, Retrieval-Augmented Generation, Large Multimodal Models, Zero-shot, Training-free

I. INTRODUCTION

Recent advancements in Large Multimodal Models (LMMs) have demonstrated exceptional capabilities across a wide spectrum of visual understanding tasks. Specifically, LMMs excel in high-level perception and reasoning tasks, such as image captioning, visual question answering, and cross-modality grounding, as validated by benchmarks like OCRBench [1]. Crucially, LMMs also exhibit strong performance in low-level visual perception and assessment, including but not limited to Image Quality Assessment (IQA) [2]–[4] and Video Quality Assessment (VQA) [5]–[7], as evidenced by Q-Bench [8]. Q-Bench [8] first demonstrated the superior zero-shot capability of LMMs in IQA. By leveraging a simple softmax-based scoring strategy on quality-related tokens, LMMs were shown to achieve performance notably surpassing both traditional hand-crafted and modern, pre-trained CLIP-based IQA methodologies. To further harness this potential, subsequent works have focused on fine-tuning LMMs for better quality perception. For instance, Q-Align [9] was inspired by subjective studies where human raters only judge discrete text-defined levels,

Email :{fuk20-20, huiyuduan, zcz1998, zyc420, minxiongkuo, jiawang, zhaiwangtao}@sjtu.edu.cn; barryjzhao@tencent.com

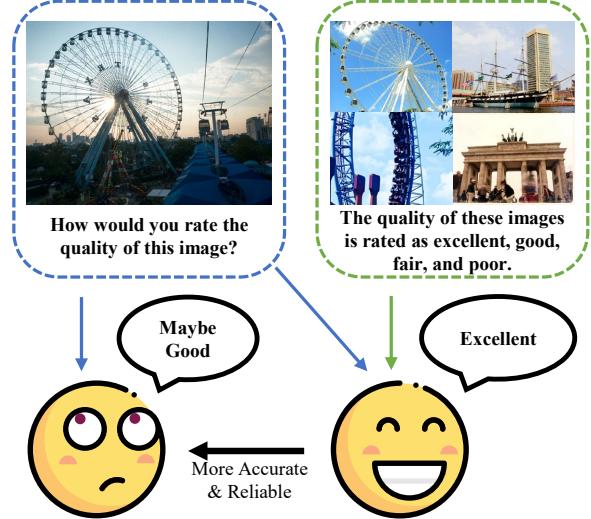


Fig. 1. Motivation of IQARAG. By referencing multiple semantically similar but quality-variant images, the people can evaluate image quality more accurate and reliable.

proposing a method to train an LMM directly with text-defined rating levels instead of continuous scores. DeQA-Score [10] proposed a distribution-based approach, discretizing the score distribution into a soft label to achieve a more precise quality score prediction. Furthermore, Q-Insight [11] introduced a reinforcement learning framework, designing a reward function to optimize LMMs for score regression and degradation perception tasks. However, these fine-tuning methods require substantial computational resources and time, making the process of adapting LMMs for specific quality assessment tasks inefficient. Fundamentally, these fine-tuning approaches aim to align the distribution of quality-related tokens in the LMM’s output more closely with the distribution of image quality scores. This insight leads us to explore alternative, resource-efficient strategies. Retrieval-Augmented Generation (RAG) [12] is a powerful and efficient technique to improve the LMM’s output accuracy and grounding by augmenting the input prompt with relevant context retrieved from an external, non-parametric knowledge base. Based on RAG, we hypothesize that providing an LMM with a set of semantically similar but quality-variant reference images can serve as a visual quality anchor for quality assessment, thereby enhancing its IQA capability without costly fine-tuning.

Therefore, we introduce **IQARAG**, a novel, training-free

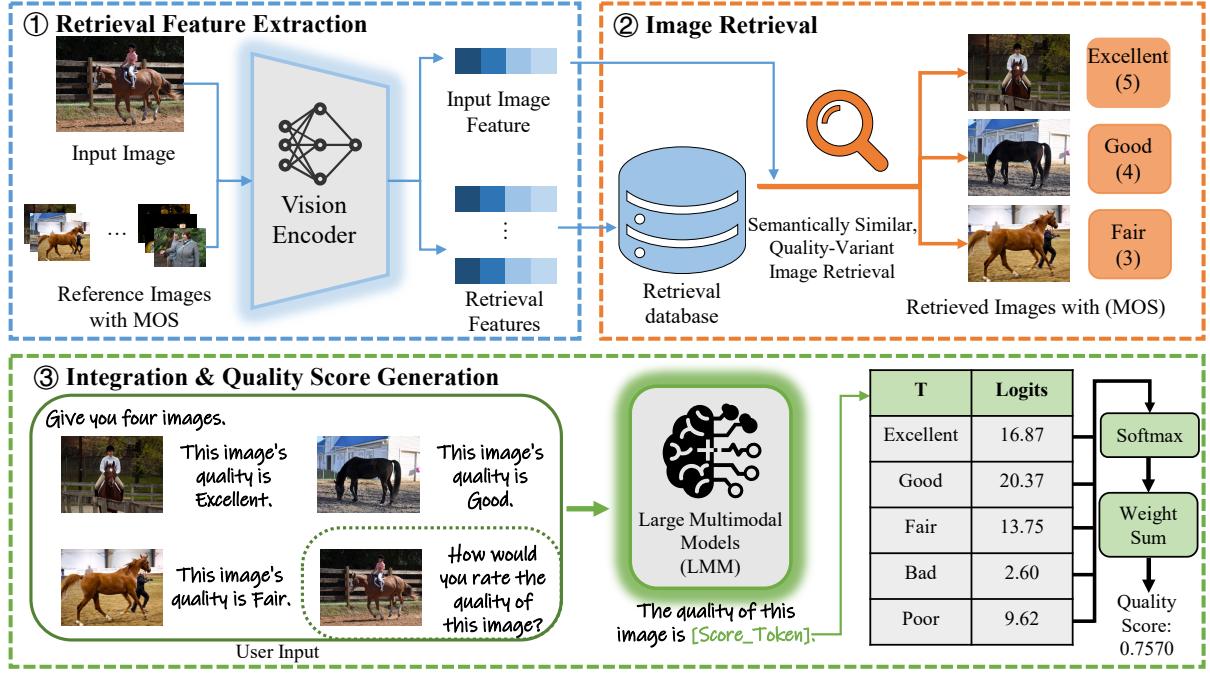


Fig. 2. **Overview of the IQARAG framework.** It comprises three phases: (1) **Retrieval Feature Extraction**, which encodes images into unified visual features; (2) **Image Retrieval**, which identifies semantically similar but quality-variant references from a curated set; and (3) **Integration & Quality Score Generation**, which feeds the input, retrieved images, and their MOSs into an LMM to derive the final quality score via quality-related tokens. In phase (3), the dotted/solid line indicates the input prompt without/with IQARAG.

methodology that leverages RAG technique to enhance the IQA performance of LMMs. The IQARAG framework consists of three key phases: (1) Retrieval Feature Extraction: Encoding the input and reference images into unified visual features via specific vision encoder. (2) Image Retrieval: Retrieving semantically similar but quality-variant reference images by calculating the similarity between the feature of the input image and a curated reference image set. (3) Integration & Quality Score Generation: Integrating the retrieved images with corresponding Mean Opinion Scores (MOSs) and the input image into a specific prompt, which is then fed into the LMM to get the final quality prediction from the output's quality-related tokens. We conducted extensive experiments on multiple representative IQA datasets, including KADID [13], KonIQ [14], LIVE Challenge [15], and SPAQ [16]. The results consistently demonstrate that the proposed IQARAG significantly enhances the IQA ability of LMMs.

Our core contributions can be summarized as follows:

- We propose the **IQARAG**, a novel training-free framework to improve IQA ability of LMMs by RAG technique. It consists of three key phases: Retrieval Feature Extraction, Image Retrieval and Integration & Quality Score Generation.
- We design a retrieval and integration pipeline that retrieves reference images and serve them as visual perception anchors to effectively guiding the LMM align its output's quality-related token distributions with realistic perception quality distributions.
- Extensive experiments on four mainstream IQA datasets demonstrate that IQARAG significant improve the IQA

performance of LMMs without fine-tuning.

II. RELATED WORK

A. Image Quality Assessment

As a fundamental task in multimedia processing, IQA systematically investigates the influence of various distortions and quality degradation factors on human perceptual experience. Early IQA methods utilize the prior knowledge of human visual system (HSV) or Natural Scene Statistics (NSS) to extract hand-crafted features to assess image quality score [17], [18]. With the advancement of deep learning, many researchers established lots of subjective IQA datasets for different applications (image compression [13], mobile photography [16], etc.). Based on these datasets, many works [19]–[24] designed different deep neural networks (DNN) to predict image quality score. The LMM shows the remarkable capabilities in high-level perception and understanding and low-level visual perception and assessment. Many works [9]–[11], [25] proposed different methods to finetune the LMMs for IQA task. However, finetuning a LMM is time-consuming and computationally expensive. So we propose a training-free pipeline to enhance the IQA ability of LMMs.

B. Retrieval Augmented Generation

RAG [12], [26]–[28] is a recent technique that utilizes the information retrieval to reduce the hallucination responses of Large Language Models (LLMs). The core mechanism is augmenting the input with relevant content retrieved from an external, non-parametric knowledge base. The principles of RAG have been extended to LMMs, where the external

knowledge base includes multimodal data such as text, images, videos, and *etc*. In the medical domain [29], RAG is critical for improving the factuality and reliability of Medical LMMs, which frequently suffer from factual inconsistencies. RAG is also being applied to long video understanding of LMM to address the challenge of limited context windows when processing long videos [29]. In our work, we aim to improving the IQA ability of LMMs by leveraging the RAG technique to retrieve the semantic similar but quality-variant reference images and serve them as visual perception anchors to reconstruct prompt.

III. PROPOSED METHOD

We propose a novel, training-free pipeline for improving IQA ability of LMMs, named IQARAG, which can be integrated into any open-source LMMs. As illustrated in Fig. 2, our pipeline comprises three key phases: (1) **Retrieval Feature Extraction**: the input image and reference images will be inputted into specific vision encoder to extract visual features. (2) **Image Retrieval**: In this phase, the semantically similar but quality-variant images are retrieved by calculating the similarity of visual feature of input image and reference images. (3) **Integration & Quality Score Generation**: This phase integrates the retrieved images and corresponding MOSs with the input image, feeding the combined prompt into the LMMs to get the final predicted quality score form the output's quality-related tokens.

A. Retrieval Feature Extraction

Give an image awaiting quality assessment I and a reference image set $\mathcal{R} = \{(R_i, \text{MOS}_i)\}_{i=1}^N$, which contains N reference images, the $\text{MOS}_i \in [0, 1]$ represents the subjective quality score for image R_i . We can first use the vision encoder to extract their corresponding visual features:

$$F_I = \mathcal{V}(I; \theta) : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^D, \quad (1)$$

$$\mathcal{F}_{\mathcal{R}} = \mathcal{V}(\mathcal{R}; \theta) : \mathbb{R}^{N \times H \times W \times 3} \rightarrow \mathbb{R}^{N \times D}, \quad (2)$$

where $\mathcal{V}(\cdot; \theta)$ denotes the vision encoder with parameters θ . H and W are the height and width of the image. D is the feature dimension. F_I and $\mathcal{F}_{\mathcal{R}}$ are the extracted image visual feature and reference image visual feature set respectively. The extracted visual features will be used for subsequent retrieval.

B. Image Retrieval

In this phase, we can retrieve semantically similar but quality-variant reference images by calculate the similarity between the input image visual feature and reference image visual features:

$$\mathcal{S} = \{\text{sim}(F_I, F_{R_i})\}_i^N, \quad \forall i \in \{1, \dots, N\}, \quad (3)$$

where $\text{sim}(\cdot, \cdot)$ denote the similarity metric, which can be implemented using various measures such as cosine similarity or L-norm distances. In our methods, we employ the L_2 norm for this computation:

$$\text{sim}(a, b) = \|a - b\|_2. \quad (4)$$

The top- K most relevant reference images are retrieved via:

$$\mathcal{R}^* = \{(R_{\tau(k)}, \text{MOS}_{\tau(k)})\}_{k=1}^K, \quad (5)$$

where τ is a permutation of indices $\{1, \dots, N\}$ that sorts the similarity scores \mathcal{S} in ascending order, i.e., $\mathcal{S}_{\tau(1)} \leq \mathcal{S}_{\tau(2)} \leq \dots \leq \mathcal{S}_{\tau(N)}$. The current retrieval images are semantically similar images. To ensure diversity in quality levels, we partition the MOS range $[0, 1]$ into 5 uniform bins and sample from each:

$$\mathcal{B}_j = \left[\frac{j-1}{5}, \frac{j}{5} \right), \quad j \in \{1, \dots, 5\}. \quad (6)$$

For each quality bin \mathcal{B}_j , we select the first image from \mathcal{R}^* (top- K retrieved) falling in that bin:

$$\mathcal{R}' = \bigcup_{j=1}^5 \left\{ (R_k, \text{MOS}_k) \in \mathcal{R}^* \mid \text{MOS}_k \in \mathcal{B}_j \right\}_{\text{first}}, \quad (7)$$

where $\{\cdot\}_{\text{first}}$ denotes the first occurrence in \mathcal{R}^* per bin. The final semantically similar but quality-variant reference image set \mathcal{R}' contains $P \leq 5$ images:

$$\mathcal{R}' = \{(R_{k_p}, \text{MOS}_{k_p})\}_{p=1}^P. \quad (8)$$

It is worth noting that if \mathcal{R}^* does not contain an image in the bin \mathcal{B}_j , then ultimately \mathcal{R}' will not contain an image with quality in the bin \mathcal{B}_j .

C. Integration & Quality Score Generation

After retrieving the semantically similar but quality-variant reference image set \mathcal{R}' , we organize it with input image and specific text T to construct prompt and then fed it into LMM to get the log probabilities (*logits*) of quality-related tokens in “[SCORE_TOKEN]”. The overall process can be formulated as:

$$\mathcal{P}_{\log}(w) = \text{LMM}(\{\mathcal{R}', I, T\}) \quad w \in \mathcal{W}, \quad (9)$$

where $\mathcal{W} = \{\text{excellent, good, fair, poor, bad}\}$ is the predefined quality-related words set. $\mathcal{P}_{\log}(w)$ is the *logits* of specific word w . Then, we can conduct a close-set softmax on the *logits* $\mathcal{P}_{\log}(\mathcal{W})$ to get probabilities:

$$P(w) = \frac{\exp(\mathcal{P}_{\log}(w))}{\sum_{w' \in \mathcal{W}} \exp(\mathcal{P}_{\log}(w'))}, \quad w \in \mathcal{W} \quad (10)$$

The final predicted quality score S is obtained through weighted sum:

$$S = \sum_{w \in \mathcal{W}} P(w) \cdot l(w). \quad (11)$$

where $l(w)$ is the weight of specific word s . For instance, For instance, the weight of *excellent* to *poor* can be set 1 to 0.

TABLE I

RESULTS ON FOUR MAINSTREAM IQA DATASETS. THE "RATIO" COLUMN DENOTES THE PROPORTION OF REFERENCE IMAGES TO TEST IMAGES. THE GRAY ROW HIGHLIGHTS RESULTS USING IQARAG WITH THE SPECIFIED LMM. **AVG.** REPRESENTS THE MEAN PERFORMANCE ACROSS ALL DATASETS, WHILE **COM.** INDICATES THE PERFORMANCE ON THE COMBINED DATASET.

Ratio	LMM	KADID-10k		KonIQ-10k		LIVEC		SPAQ		AVG.		COM.	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
1:9	Qwen3-VL	0.7280	0.7171	0.7712	0.8090	0.7897	0.7735	0.8427	0.7920	0.7829	0.7729	0.7378	0.7132
		0.7682	0.7252	0.8071	0.8036	0.7958	0.8189	0.8169	0.8282	0.7970	0.7940	0.7612	0.7672
	InternVL3.5	0.5174	0.5432	0.6046	0.6518	0.4756	0.5099	0.6479	0.6711	0.5614	0.5940	0.6009	0.6341
		0.5186	0.5247	0.6268	0.6656	0.4955	0.5158	0.7631	0.7672	0.6010	0.6183	0.6326	0.6429
	Kimi-VL	0.7342	0.7017	0.9094	0.8938	0.8129	0.8455	0.8283	0.8622	0.8212	0.8258	0.7738	0.7620
		0.7254	0.7068	0.9110	0.9121	0.8397	0.8380	0.8249	0.8552	0.8253	0.8280	0.8089	0.8137
1:4	Qwen3-VL	0.7290	0.7168	0.7692	0.8081	0.8011	0.7818	0.8415	0.7900	0.7852	0.7742	0.7377	0.7131
		0.7686	0.7213	0.7900	0.7963	0.8068	0.8054	0.8183	0.8307	0.7959	0.7884	0.7530	0.7629
	InternVL3.5	0.5236	0.5487	0.6056	0.6506	0.5087	0.5360	0.6452	0.6691	0.5708	0.6011	0.6046	0.6358
		0.5279	0.5317	0.6850	0.7381	0.4830	0.5297	0.7841	0.8019	0.6200	0.6503	0.6250	0.6435
	Kimi-VL	0.7356	0.7011	0.9098	0.8943	0.8253	0.8519	0.8281	0.8619	0.8247	0.8273	0.7738	0.7618
		0.7287	0.7096	0.9084	0.9055	0.8480	0.8541	0.8177	0.8509	0.8257	0.8300	0.8120	0.8123
3:7	Qwen3-VL	0.7320	0.7192	0.7671	0.8048	0.7906	0.7705	0.8427	0.7944	0.7831	0.7722	0.7396	0.7158
		0.7707	0.7418	0.7985	0.7976	0.8121	0.8110	0.8188	0.8273	0.8000	0.7944	0.7694	0.7764
	InternVL3.5	0.5210	0.5491	0.5982	0.6378	0.5015	0.5306	0.6469	0.6709	0.5669	0.5971	0.6014	0.6358
		0.5206	0.5191	0.6517	0.6840	0.5374	0.5764	0.7932	0.8006	0.6257	0.6450	0.6219	0.6410
	Kimi-VL	0.7367	0.7035	0.9098	0.8920	0.8253	0.8569	0.8266	0.8619	0.8246	0.8286	0.7758	0.7636
		0.7357	0.7160	0.9077	0.9058	0.8471	0.8571	0.8427	0.8687	0.8333	0.8369	0.8092	0.8135

IV. EXPERIMENT

A. Datasets

In order to prove the efficiency of IQARAG, we conduct experiments on four mainstream IQA datasets: KADID-10k [13], KonIQ-10k [14], LIVE Challenge [15], SPAQ [16]. The KADID-10k dataset is a large-scale, synthetically distorted IQA database. It consists of 10,125 distorted images generated from 81 high-quality reference images. The KonIQ-10k dataset is a large-scale, in-the-wild IQA database. It comprises 10,073 diverse images sourced from real-world photography. The LIVE Challenge dataset consists of 1,162 in-the-wild images with subjective MOS. The SPAQ dataset is a large-scale IQA database specifically focused on the perceived quality and attributes of smartphone photography. It includes 11,125 images captured by 66 different smartphone models under various real-world conditions.

B. Implementation Details

1) *Evaluation Criteria:* Following previous work [7], [8], we employ the common consistency evaluation criteria to judge the correlation between the predicted scores and the quality annotations. These criteria include the Spearman Rank Correlation Coefficient (SRCC) and the Pearson Linear Correlation Coefficient (PLCC). An effective quality assessment model should aim for SRCC and PLCC values close to 1.

2) *Experimental Setup:* We select three mainstream LMMs for our experiments: Qwen3-VL [30], InternVL3.5 [31], and Kimi-VL [32]. Specifically, we use the following versions: Qwen3-VL-8B-Instruct, InternVL3.5-8B, and Kimi-VL-A3B-Instruct. To ensure a comprehensive analysis of performance, we investigate the impact of the reference image set and the test set size distribution by splitting the dataset using three different ratios: 1:9, 1:4, and 3:7. We evaluate the performance

of the selected LMMs both without and with the integration of IQARAG. For the vision encoder used to extract visual features, we employ the respective vision encoder built into each LMM. For the cross dataset experiment, we use the LIVE Challenge as the reference image set and evaluated our methods on other three datasets. In the vision encoder experiment, we also test classic image encoders such as ResNet [33], Swin-b (Swin Transformer Base) [34], DINoV2 [35], and CLIP [36] for comparative analysis. Furthermore, we conduct an ablation study on the number of reference images, P . For the cross dataset, feature experiments and the ablation study, we use Qwen3-VL as the experimental LMM. All experiments were performed on a server equipped with four NVIDIA 4090 GPUs. We utilized the Python package faiss [37] to accelerate the image retrieval process.

3) *Prompt setting:* We detail the prompt templates used for the LMMs both without and with the integration of IQARAG. We denote the input image and reference image as $\langle \text{img} \rangle$ and $\langle \text{rimg} \rangle$, and $\langle \text{level} \rangle$ represents the corresponding quality level annotation of the reference image.

Without IQARAG:

User: $\langle \text{img} \rangle$ How would you rate the quality of this image?
Assistant: The quality of this image is [SCORE_TOKEN]

With IQARAG:

User: Give you P images.
User: $\langle \text{rimg} \rangle$ This image's quality is $\langle \text{level} \rangle$. (This sentence used for each reference image in \mathcal{R}')
User: $\langle \text{img} \rangle$ How would you rate the quality of this image?
Assistant: The quality of this image is [SCORE_TOKEN]

C. Experiment results

1) *IQARAG Performance:* For the results shown in the Table I. We can find that: (1) The integration of IQARAG significantly improves the IQA performance of LMMs, demonstrat-

																							
E	G	F	P	B	S	E	G	F	P	B	S	E	G	F	P	B	S	E	G	F	P	B	S
14.81	21.75	27.00	12.25	23.63	0.48	14.75	19.38	28.38	14.32	27.88	0.31	14.69	17.00	28.25	15.50	29.00	0.16	20.00	14.88	22.00	21.25	29.25	0.00
24.38	35.50	33.00	19.00	30.88	0.72	22.88	35.25	36.00	20.88	33.00	0.56	23.75	32.75	38.25	25.50	37.50	0.34	23.12	24.88	28.13	42.75	42.75	0.15
					0.75						0.51					0.33							0.14

Fig. 3. Example *logits* of quality-related tokens (Excellent, Good, Fair, Poor, Bad). The **S** means quality score. The green, blue and red data indicates without/with IQARAG and ground-truth respectively. It indicates that employing the IQARA framework aligns the *logits* distribution of quality-related tokens more closely with the ground-truth quality score distribution.

TABLE II
RESULTS OF CROSS DATASET EXPERIMENT. METRICS ARE (PLCC + SRCC)/2. THE GRAY ROW REPRESENTS THE RESULT OF USING IQARAG.

LMM	KADID	KonIQ	SPAQ	AVG.
Qwen3-VL	0.7236	0.7896	0.8168	0.7776
	0.6352	0.7912	0.8278	0.7514
Intern3.5VL	0.5315	0.6286	0.6574	0.6058
	0.5141	0.6278	0.7933	0.6450
Kimi-VL	0.7192	0.9024	0.8451	0.8222
	0.7018	0.8952	0.8500	0.8157

ing this mechanism’s universal effectiveness. While IQARAG benefits all models, its effectiveness is partially dependent on the LMM’s underlying capability, as seen by InternVL3.5’s lower performance compared to the others. (2) IQARAG still has great performance when the reference set is small (1:9 ratio). In the same time, the performance gain is maintained even as the reference set size increases (up to the 3:7 ratio), confirming that IQARAG provides a robust performance boost under normal dataset split. (3) Kimi-VL exhibits the strongest baseline performance across all datasets and splitting ratios, consistently achieving the highest overall average scores among the tested LMMs, which may because it has the most parameters. Fig 3 shows the examples without/with IQARAG, we can find that the *logits* distribution is more close to realistic quality distribution and the predict score is more accuracy when applying IQARAG.

TABLE III
RESULTS OF VISION ENCODER EXPERIMENT. METRICS ARE (PLCC + SRCC)/2. ‘-’ INDICATES THAT THE IQARAG METHOD IS NOT USED. ‘VE’ REPRESENTS VISION ENCODER BUILT INTO LMM. BEST IN RED, SECOND IN BLUE.

Features	KADID	KonIQ	LIVEC	SPAQ	AVG.
-	0.7256	0.7859	0.7805	0.8185	0.7777
ResNet	0.7467	0.7985	0.7986	0.8150	0.7897
Swin-b	0.7503	0.8007	0.8049	0.8199	0.7939
DINOv2	0.7525	0.8038	0.8052	0.8235	0.7962
CLIP	0.7536	0.7956	0.8076	0.8260	0.7957
VE	0.7563	0.7935	0.8115	0.8231	0.7972

2) *Cross dataset experiment*: The Table II shows the results of cross dataset experiment. We can find that: after applying IQARAG, the results showed a decline on the KADID-10k dataset but improved on other datasets. This discrepancy may arise because KADID-10k is a synthetic dataset, whereas the others are realistic datasets (including LIVEC). This indicates that the effectiveness of IQARAG depends on the similarity between the reference image dataset and the test dataset.

3) *Vision Encoder Selection*: The Table III shows the results when using different vision encoders to extract visual features. We can find that the vision encoder built into the LMM achieves the highest average performance, which suggests that the vision encoder built into the LMM is the most effective feature extractor among those tested encoders for IQARAG method. Among the tested standalone encoders (ResNet, Swin-b, DINOv2, CLIP), DINOv2 and CLIP show the strongest overall results.

4) *Ablation study*: The Table IV presents the results with different numbers of reference images *P*. Ablation study. Results are (PLCC + SRCC)/2. ‘-’ indicates that the IQARAG method is not used. Best in Red, second in Blue.

TABLE IV
THE NUMBER OF REFERENCE IMAGES *P* ABLATION STUDY. RESULTS ARE (PLCC + SRCC)/2. ‘-’ INDICATES THAT THE IQARAG METHOD IS NOT USED. BEST IN RED, SECOND IN BLUE.

P	KADID	KonIQ	LIVEC	SPAQ	AVG.
-	0.7256	0.7859	0.7805	0.8185	0.7777
1	0.7221	0.7853	0.7842	0.8069	0.7746
2	0.7377	0.7885	0.7912	0.8127	0.7825
3	0.7496	0.7998	0.8020	0.8149	0.7916
4	0.7518	0.7909	0.8041	0.8232	0.7925
5	0.7563	0.7980	0.8115	0.8231	0.7972

V. CONCLUSION

In this paper, we propose **IQARAG**, a novel training-free RAG framework designed to enhance the IQA capability of LMMs. The framework consists of three key phases: Retrieval

Feature Extraction, where visual features of input and reference images are extracted using a specific vision encoder; Image Retrieval, which retrieves semantically similar but quality-variant reference images with MOSs by calculating similarity of input image and reference image features; and Integration & Quality Score Generation, where the input and retrieved images are combined to construct a multimodal prompt for the LMM to predict the final quality score. Experimental results demonstrate that LMMs equipped with **IQARAG** achieve significantly improved performance on IQA tasks, confirming the effectiveness of our proposed approach.

REFERENCES

- [1] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai, “Ocrbench: on the hidden mystery of ocr in large multimodal models,” *Science China Information Sciences*, vol. 67, no. 12, pp. 220102, 2024.
- [2] Jiarui Wang, Huiyu Duan, Guangtao Zhai, Junlong Wang, and Xiongkuo Min, “Aigv-assessor: Benchmarking and evaluating the perceptual quality of text-to-video generation with lmm,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025.
- [3] Zicheng Zhang, Yingjie Zhou, Chunyi Li, Baixuan Zhao, Xiaohong Liu, and Guangtao Zhai, “Quality assessment in the era of large models: A survey,” *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 21, no. 7, pp. 1–31, 2025.
- [4] Jiarui Wang, Huiyu Duan, Guangtao Zhai, and Xiongkuo Min, “Quality assessment for ai generated images with instruction tuning,” *IEEE Transactions on Multimedia (TMM)*, 2025.
- [5] Xiongkuo Min, Huiyu Duan, Wei Sun, Yucheng Zhu, and Guangtao Zhai, “Perceptual video quality assessment: A survey,” *Science China Information Sciences*, vol. 67, no. 11, pp. 211301, 2024.
- [6] Zicheng Zhang, Ziheng Jia, Haoning Wu, Chunyi Li, Zijian Chen, Yingjie Zhou, Wei Sun, Xiaohong Liu, Xiongkuo Min, Weisi Lin, et al., “Q-bench-video: Benchmark the video quality understanding of lmms,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 3229–3239.
- [7] Huiyu Duan, Qiang Hu, Jiarui Wang, Liu Yang, Zitong Xu, Lu Liu, Xiongkuo Min, Chunlei Cai, Tianxiao Ye, Xiaoyun Zhang, and Guangtao Zhai, “Fineqv: Fine-grained user generated content video quality assessment,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025.
- [8] Zicheng Zhang, Haoning Wu, Erli Zhang, Guangtao Zhai, and Weisi Lin, “Q-bench: A benchmark for multi-modal foundation models on low-level vision from single images to pairs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [9] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al., “Q-align: Teaching lmms for visual scoring via discrete text-defined levels,” *arXiv preprint arXiv:2312.17090*, 2023.
- [10] Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong, “Teaching large language models to regress accurate image quality scores using score distribution,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025, pp. 14483–14494.
- [11] Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang, “Q-insight: Understanding image quality via visual reinforcement learning,” *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2025.
- [12] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al., “Retrieval-augmented generation for knowledge-intensive nlp tasks,” *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 9459–9474, 2020.
- [13] Hanhe Lin, Vlad Hosu, and Dietmar Saupe, “Kadid-10k: A large-scale artificially distorted iqa database,” in *QoMEX*, 2019, pp. 1–3.
- [14] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe, “Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment,” *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 4041–4056, 2020.
- [15] Deepthi Ghadiyaram and Alan C Bovik, “Massive online crowdsourced study of subjective and objective picture quality,” *IEEE Transactions on Image Processing (TIP)*, vol. 25, no. 1, pp. 372–387, 2015.
- [16] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang, “Perceptual quality assessment of smartphone photography,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [17] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik, “Making a “completely blind” image quality analyzer,” *IEEE Signal Processing Letters (SPL)*, 2012.
- [18] Lin Zhang, Lei Zhang, and Alan C Bovik, “A feature-enriched completely blind image quality evaluator,” *IEEE Transactions on Image Processing (TIP)*, 2015.
- [19] Hossein Talebi and Peyman Milanfar, “Nima: Neural image assessment,” *IEEE Transactions on Image Processing (TIP)*, 2018.
- [20] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang, “Blind image quality assessment using a deep bilinear convolutional neural network,” *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 30, no. 1, pp. 36–47, 2020.
- [21] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinjiu Sun, and Yanning Zhang, “Blindly assess image quality in the wild guided by a self-adaptive hyper network,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [22] Huiyu Duan, Xiongkuo Min, Yucheng Zhu, Guangtao Zhai, Xiaokang Yang, and Patrick Le Callet, “Confusing image quality assessment: Toward better augmented reality experience,” *IEEE Transactions on Image Processing (TIP)*, vol. 31, pp. 7206–7221, 2022.
- [23] Xilei Zhu, Liu Yang, Huiyu Duan, Xiongkuo Min, Guangtao Zhai, and Patrick Le Callet, “Esiqa: Perceptual quality assessment of vision-probe-based egocentric spatial images,” *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2025.
- [24] Huiyu Duan, Xiongkuo Min, Wei Sun, Yucheng Zhu, Xiao-Ping Zhang, and Guangtao Zhai, “Attentive deep image quality assessment for omnidirectional stitching,” *IEEE Journal of Selected Topics in Signal Processing (JSTSP)*, vol. 17, no. 6, pp. 1150–1164, 2023.
- [25] Zicheng Zhang, Haoning Wu, Zhongpeng Ji, Chunyi Li, Erli Zhang, Wei Sun, Xiaohong Liu, Xiongkuo Min, Fengyu Sun, Shangling Jui, et al., “Q-boost: On visual quality assessment ability of low-level multimodality foundation models,” in *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE, 2024, pp. 1–6.
- [26] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang, “Retrieval-augmented generation for large language models: A survey,” *arXiv preprint arXiv:2312.10997*, vol. 2, no. 1, 2023.
- [27] Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu, “Evaluation of retrieval-augmented generation: A survey,” in *CCF Conference on Big Data*. Springer, 2024, pp. 102–120.
- [28] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun, “Benchmarking large language models in retrieval-augmented generation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 17754–17762.
- [29] Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao, “Rule: Reliable multimodal rag for factuality in medical vision language models,” *arXiv preprint arXiv:2407.05131*, 2024.
- [30] Yuxuan Cai Shuai Bai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al., “Qwen3-vl technical report,” *arXiv preprint arXiv:2511.21631*, 2025.
- [31] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xinguang Wei, et al., “Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency,” *arXiv preprint arXiv:2508.18265*, 2025.
- [32] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, et al., “Kimi-VL technical report,” 2025.
- [33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

- [34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10012–10022.
- [35] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, et al., “Dinov2: Learning robust visual features without supervision,” *arXiv:2304.07193*, 2023.
- [36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, et al., “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [37] Jeff Johnson, Matthijs Douze, and Hervé Jégou, “Billion-scale similarity search with gpus,” *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.