Depicting Beyond Scores: Advancing Image Quality Assessment through Multi-modal Language Models

Zhiyuan You[12*], Zheyuan Li[26*], Jinjin Gu[34*], Zhenfei Yin[34], Tianfan Xue[1†], and Chao Dong[235†]

[1] The Chinese University of Hong Kong
[2] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[3] Shanghai AI Laboratory [4] University of Sydney
[5] Shenzhen University of Advanced Technology [6] University of Macau
zhiyuanyou@foxmail.com zheyuanli884886@gmail.com jinjin.gu@sydney.edu.au
zyin7056@uni.sydney.edu.au tfxue@ie.cuhk.edu.hk chao.dong@siat.ac.cn
[*] Contribute Equally [†] Corresponding Author
Project Page: https://depictqa.github.io

Abstract. We introduce a **Depict**ed image **Q**uality **A**ssessment method (DepictQA), overcoming the constraints of traditional score-based methods. DepictQA allows for detailed, language-based, human-like evaluation of image quality by leveraging Multi-modal Large Language Models (MLLMs). Unlike conventional Image Quality Assessment (IQA) methods relying on scores, DepictQA interprets image content and distortions descriptively and comparatively, aligning closely with humans' reasoning process. To build the DepictQA model, we establish a hierarchical task framework, and collect a multi-modal IQA training dataset. To tackle the challenges of limited training data and multi-image processing, we propose to use multi-source training data and specialized image tags. These designs result in a better performance of DepictQA than score-based approaches on multiple benchmarks. Moreover, compared with general MLLMs, DepictQA can generate more accurate reasoning descriptive languages. We also demonstrate that our full-reference dataset can be extended to non-reference applications. These results showcase the research potential of multi-modal IQA methods.

Keywords: Image Quality Assessment · Multi-modal Language Models

1 Introduction

Image Quality Assessment (IQA) is an important topic in low-level vision research [18, 24, 38, 75], and it is widely applied in image generation and processing [8, 10, 11, 49]. IQA aims to measure and compare the quality of images, expecting the final results to be aligned with human judgments. Existing IQA methods [18,47,61,75] mainly output the quality or similarity scores, which have apparent shortcomings. First, image quality is affected by different factors that

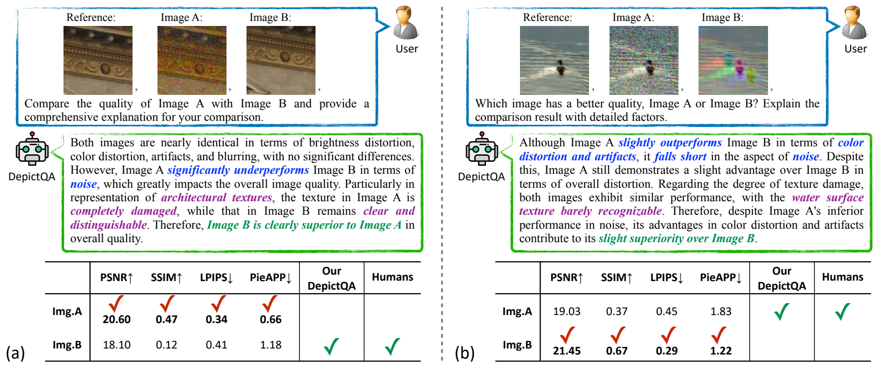

Fig. 1: Comparison between our DepictQA and score-based IQA methods, including PSNR, SSIM [61], LPIPS [75], and PieAPP [47]. Score-based IQA methods only provide numerical scores devoid of reasoning and justification. Thus they disagree with human judgments in complex scenarios when **(a)** images are misaligned and **(b)** both images suffer from severe distortions. In contrast, DepictQA first identifies the distortions of images, then weighs the influences of different distortions to the texture damages, and finally obtains the comparison results that are better aligned with human judgments.

cannot be effectively expressed by a simple score, *e.g.*, noise, color distortion, and artifacts in Fig. 1. Second, the reasoning process by humans cannot be well modeled by current IQA methods. For example, in Fig. 1b, humans may first identify the distortions (*i.e.*, noise in Image A, color distortion and artifacts in Image B), then weigh the impacts of these distortions on overall visual quality (color distortion and artifacts in Image B are worse than noise in Image A), and finally conclude that Image A is better than Image B. On the contrary, existing IQA methods simply compare the quality scores of these two images.

To better align with humans, we explore a new paradigm for IQA, named **Depict**ed image **Q**uality **A**ssessment (DepictQA). Inspired by recent Large Language Models (LLMs) [41, 58] and multi-modal technologies [30, 82], we believe that language is the key to solving the above problems. As shown in Fig. 1, DepictQA takes both images and a question as inputs, then outputs a paragraph that describes the quality of images from multiple aspects. Furthermore, empowered by the reasoning capability of LLMs, DepictQA can weigh the importance of each distortion and make the final judgment. For instance, in Fig. 1a, DepictQA finds that "the texture in Image A is completely damaged" while "Image B remains clear and distinguishable", thus concludes "Image B is superior to Image A". Learning this kind of reasoning makes DepictQA better at aligning human judgments than existing methods in complex scenarios like misalignment (Fig. 1a) and multiple distortions (Fig. 1b). Meanwhile, these descriptive outputs can be naturally understood by humans, greatly improving the interpretability.

To integrate language into IQA, we establish a hierarchy of tasks, inspired by human evaluation. Humans first perceive the distortions of the image, then use this information to determine the image quality. Also, it is easier for humans to compare the difference between two images in a single dimension (*e.g.*, color distortion) than quantitatively evaluate the overall quality of an image or the similarity between two images, as verified by [45–47]. Based on this intuition, DepictQA does not produce scores, but describes the image quality and compares two images. Specifically, we break DepictQA task into a hierarchy of 3 tasks (detailed in Fig. 2): (1) Quality Description, (2) Quality Comparison, and (3) Comparison Reasoning. These designs follow the process of human evaluation.

To train the proposed DepictQA, we further construct a multi-modal IQA dataset, named M-BAPPS, by collecting text descriptions based on the existing BAPPS IQA dataset [75]. Our M-BAPPS dataset contains 5,104 detailed high-quality text descriptions and 115,646 brief descriptions. For high-quality texts, we first collect the quality-related information through a carefully designed questionnaire (details shown in Fig. 2a and 2c), the results of which are further converted into a descriptive paragraph using GPT-4 [41]. To further increase the size of the training set, we also augment the dataset with brief descriptions. Specifically, we convert the existing quality comparison label in BAPPS into a brief description using pre-generated templated texts, such as "Image A maintains a distinct advantage in terms of image quality over Image B".

With the dataset mentioned above, we resort to Multi-modal Large Language Models (MLLMs) [21, 30, 82] to bridge the gap between images and descriptive texts. However, directly applying existing MLLMs to our DepictQA faces two challenges. First, there are only limited images with high-quality descriptions, preventing the model from robustly correlating images and text descriptions. In this aspect, we present a multi-source training approach to increase the size of training data. Specifically, two additional sources are used. One is images with only brief templated texts, as mentioned above. The other one is external quality-unrelated content description data, the Detailed Description dataset in [69], which contains 48,734 image-text pairs. Although these two datasets are not directly designed for the descriptive reasoning ability, we find that the former one can help bridge images and texts in quality-related tasks, while the latter one can serve as a regularization. Second, many MLLMs have difficulty in distinguishing multiple images, but our setup requires two or more images. We solve this problem by employing specialized tags for different images, instead of a unified tag for all images. Empirical results demonstrate that these approaches effectively mitigate the two challenges and bring a better DepictQA model.

Finally, we conduct extensive experiments to prove the effectiveness of our DepictQA. First, DepictQA achieves state-of-the-art performance on multiple existing IQA benchmark, well aligned with human judgments. Also, DepictQA can describe the distortions and texture damages in images and explain the reasoning process when comparing two images, thus generating more accurate descriptions compared with general-purpose MLLMs. Even compared with notably GPT-4V [42], DepictQA has significantly better comparison ability and comparable reasoning ability. Moreover, we demonstrate the utility of our full-reference dataset in non-reference applications. These results attest to the superiority of our DepictQA and the research potential of multi-modal IQA tasks.

2 Related Works

Score-based IQA methods. Most existing IQA methods rely on scores to assess image quality. They can be categorized into *full-reference* and *non-reference* methods. (1) Full-reference methods assess image quality by computing the similarity score between a distorted image and a high-quality reference image. Traditional methods rely on human-designed metrics like structural similarity [61], image information [52], phase congruency with gradient magnitude [72], *etc.* Learning-based methods aim to align with human assessment through data-driven training. LPIPS [75] shows that the learned features can effectively function as a perceptual metric, exhibiting high consistency with human judgments. In alignment with advancements in the deep-learning community, data-driven approaches [5, 7, 14–16, 47, 68, 81] have similarly spurred innovations in IQA. (2) Non-reference methods evaluate the quality of a distorted image without a reference image. Traditional methods [34, 37–40, 50, 57] primarily calculate quality scores based on human-designed natural image statistics. Deep-learning-based methods [23, 24, 31, 44, 54, 80, 83] replace hand-crafted statistics by learning quality priors from extensive data. Recent works further enhance the performance by introducing graph representation [55], CLIP pre-training [60], continual learning [76], multitask learning [77], and so on. However, score-based IQA methods exhibit inherent limitations, particularly the inability to reflect the intricate analyses and weights of multiple aspects, as discussed in Sec. 1.

MLLMs incorporate the vision modality into large language models [12, 41, 58], aiming to leverage their emergent ability to achieve general vision ability. These MLLMs [2, 13, 27, 30, 42, 67, 69, 73, 74, 82] have demonstrated a general visual ability and can tackle various multi-modality tasks, including image captioning [1, 9, 71], visual question answering [17, 32, 33], document understanding [35, 36, 53], *etc.* Although proficient in these high-level perception tasks, we demonstrate in Sec. 5.3 that general MLLMs are still not good at IQA tasks.

MLLM-based IQA methods aim to achieve better alignment with human perception leveraging languages [66]. Q-Bench [62, 78] constructs a benchmark to assess existing MLLMs in low-level perception tasks. Q-Instruct [63] and Co-Instruct [65] further promote the low-level perception ability of MLLMs by introducing large-scale datasets. Q-Align [64] utilizes the text-guided instruction tuning for more accurate quality score regression. Our work distinguishes itself from existing works. Our focus lies on quality comparison regarding distortions and texture damages across multiple images, whereas existing works primarily center on low-level perception and score regression within individual images.

3 DepictQA Task and Dataset

3.1 Task Description

Before introducing our method, we need to rethink the paradigm of IQA. To reflect the human process of assessing image quality, we intend to apply language as a powerful interactive tool. Intuitively, DepictQA needs the following abilities.

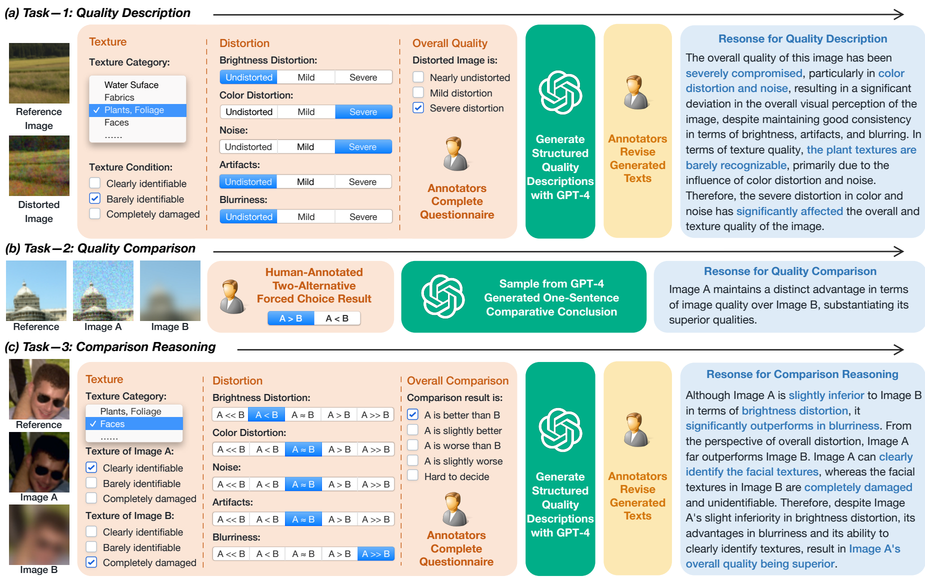

Fig. 2: Collection of the responses in our M-BAPPS dataset. We first carefully design a questionnaire to collect quality-related information. We then employ the GPT-4 [41] to convert our annotated questionnaire results into natural language. Finally, the outputs of GPT-4 are modified and improved by the annotators to correct errors, eliminate ambiguities, and supplement important information.

First, DepictQA needs to **identify the distortions and texture damages** (Fig. 2a). Humans begin assessing image quality by identifying distortions and texture damages, as this is the basis for any subsequent assessment.

Second, DepictQA is required to **compare distorted images** like Fig. 2b rather than just calculate scores for individual images. Quantifying image quality has drawbacks, as the information from a single score is quite limited. It has also been verified that humans tend to make a biased quality assessment on a single image, but are more consistent and reliable in comparing two images [18,19,45].

Third, DepictQA should **weigh and judge multiple aspects** that affect image quality. Humans consider many factors when comparing image quality. For example, when comparing an underexposed image and a blurry image in Fig. 2c, one may need to consider the real impacts of these two distortions on the texture representation, and weigh among these considerations. DepictQA should mimic this weighing ability, which distinguishes it from previous IQA methods.

Based on the above discussion, as described in Fig. 2, we design a hierarchical task paradigm, progressively guiding DepictQA to obtain the above abilities:

– *Task–1: Quality Description*. Based on the provided reference image and a distorted image, describe the distortions and texture damages in the distorted image, and comment on the overall quality of the distorted image.

- *Task–2: Quality Comparison.* Based on the provided reference image and its two distorted versions, determine which distorted image has a better quality.
- *Task–3: Comparison Reasoning.* Based on the comparison result, describe the distortions and texture damages of the two distorted images, make inferences, and weigh the pros and cons to justify its judgment on image quality.

3.2 Dataset Construction

Data serves as the key factor for training MLLMs. We follow the scheme of supervised fine-tuning [30,69] to train DepictQA. Supervised fine-tuning requires collecting {images, question, response} data pairs, where "images" are the ones to be evaluated, "question" describes the task, and "response" is the reference answer. In this section, we describe in detail our selection of images and the collection of questions and responses for the three tasks.

Image collection for the IQA dataset consists of two aspects, one is the selection of reference images, and the other is the collection of distorted images. Existing works usually collect a large number of distorted images on a small number of reference images [18, 19, 45, 47]. However, the semantic richness of the training images is also very important. In our work, we collect reference images and their corresponding distorted images from the BAPPS dataset [75]. BAPPS dataset contains 197k highly diverse samples, meeting the diversity requirements of DepictQA training. For each sample, there is a reference image and its two distorted versions, as given in Fig. 2b and 2c. BAPPS dataset also provides human annotations, indicating which distorted image is more similar to the reference one, *i.e.*, has better perceptual quality. These annotations can be used to build our dataset and validate our newly annotated data.

Question collection. Users may express similar questions in different sentences, like the two questions in Fig. 1. To encourage the robustness to users' questions, for each task, we first leverage GPT-4 to generate 30 questions. We then manually remove ambiguous and duplicate ones and modify inaccurate ones to form a set of 10 questions (see *Appendix*). During training and testing, we randomly sample a question from the question set to construct the data pair.

Response collection. A straightforward way to collect high-quality training texts is manually writing [43,59]. However, when annotators are inexperienced or tired, human-written texts can lead to biases and uneven quality. In this work, we use advanced LLMs to assist annotators in generating structured training texts, as shown in Fig. 2. We first collect the information that we want the texts to describe through a carefully designed questionnaire. Answering questions greatly reduces the possibility of ambiguity among annotators and ensures that the information is structured. We then use GPT-4 to convert our annotated questionnaire results into natural language. Finally, the outputs of GPT-4 are modified and improved by the annotators to correct errors, eliminate ambiguities, and add important information. This process greatly reduces the difficulty of collecting training texts and improves the quality of the training texts.

Next, we introduce the details of the questionnaire for different tasks.

Table 1: Statistics of our constructed M-BAPPS dataset with respect to different tasks and dataset splits.

	# Task–1 Quality Description	# Task–2 Quality Comparison	# Task–3 Comparison Reasoning
Training / Validation	1,115 / 50	115,646 / 9,440	3,739 / 200

Task–1: Quality Description. A distorted image and its reference image are shown to annotators. Fig. 2a shows our questionnaire with three parts: texture, distortion, and overall quality. For the texture part, annotators are asked to select the one that best matches the image from a list containing 11 typical texture types, including object edges, bricks, fabrics, plants or foliage, architectures, artificial strips, hairs or furs, faces, sky or clouds, stones or ground, and water surface. These 11 types are selected based on the existing IQA [19] and texture recognition [3] research. Additionally, annotators are asked to indicate whether the texture is "clearly identifiable", "barely identifiable", or "completely damaged". Regarding the distortion part, we ask annotators to summarize with the following five aspects: "brightness", "color", "noise", "artifacts", and "blurriness". For each distortion, we use three levels for evaluation: "undistorted", "mild", and "severe". This can express most distortions that appear in images. Finally, annotators need to comment on the overall quality of the image into three levels: "nearly undistorted", "mild distortion", and "severe distortion".

Task–2: Quality Comparison. BAPPS dataset already includes binary comparison labels (*i.e.*, Image A or Image B is better) for all image pairs. To convert these comparison labels into textural responses, we first build a response pool using GPT-4, including 20 generated sentences for "Image A is better" and another 20 for "Image B is better". Then, for each comparison label, we randomly sample one response from the pool, as depicted in Fig. 2b. However, the diversity of language output poses a challenge to evaluation. For the convenience of evaluation, inspired by LLaVA-1.5 [29], we randomly sample half of the questions and add the following short answer prompt: "Answer the question using a single word or phrase". Correspondingly, the response will be a single phrase like "Image A" or "Image B" indicating the less distorted image.

Task–3: Comparison Reasoning. As shown in Fig. 2c, annotators are given two distorted images and the reference image. The annotation pipeline is similar to Task–1. Annotators compare two distorted images from the five kinds of distortions and the overall distortion using five options: "superior" ($>>$), "slightly superior" ($>$), "roughly equal" ($\approx$), "slightly inferior" ($<$), and 'inferior" ($<<$).

Dataset statistics. The statistics of our dataset are presented in Tab. 1 (more in *Appendix*). Our dataset comprises 5,104 detailed high-quality samples (Task–1 and Task–3), along with 115,646 brief templated samples (training set of Task–2). The validation set of Task–2 is the same as the "Traditional" and "CNN" categories (two sets of distortions) in BAPPS's validation set [75]. Each training sample is individually annotated by one annotator. In the validation set, samples are annotated by two annotators only if they reach a consensus.

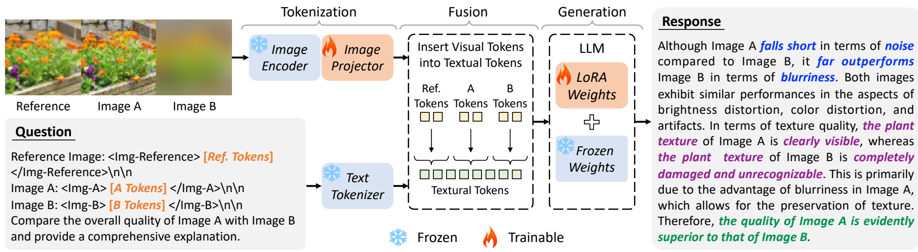

Fig. 3: Framework of DepictQA. A frozen pre-trained image encoder is employed to encode images to visual tokens, followed by a trainable image projector to project visual tokens to textual space. The question texts are tokenized by a text tokenizer. Visual tokens and textual tokens are then fused and jointly processed by an LLM, fine-tuned through the LoRA technique [20]. Our model is capable of producing comprehensive and informative explanations for image quality comparisons.

As a means of verification, the annotated "Overall Comparison" judgments in Task–3 exhibit a quite high consistency rate of 84.3% with the ground-truth judgments in the BAPPS dataset. For brevity, the three tasks will be shortened to *description*, *comparison*, and *reasoning* in the following.

4 DepictQA Framework

4.1 Model Architecture

Fig. 3 shows the workflow of our DepictQA. DepictQA takes images and a quality-related question as inputs, and generates a descriptive text as the response. In *comparison* or *reasoning* tasks, there are three input images: a reference image and Image A and B. In *description* task, there are two input images: a reference image and a distorted image. The input images and the question are first tokenized, then fused, finally processed by the LLM for response generation.

Tokenizing input images and question. As shown in Fig. 3, we employ a frozen CLIP pre-trained ViT-L/14 [48] as the image encoder to represent the input images as visual tokens. Then, the question texts are tokenized to textual tokens by SentencePiece tokenizer [25]. The visual tokens cannot be directly processed by the LLM due to different embedding spaces, so we use a trainable image projector to map visual tokens into the textual space as in [30, 82].

Token fusion. We insert the visual tokens into pre-defined positions within the textual tokens for token fusion. As show in Fig. 3, [*Ref. Tokens*], [*A Tokens*], [*B Tokens*] in the input question are these pre-defined positions.

Response generation using LLM. The fused tokens are fed into LLM to generate the final response. Our DepictQA is compatible with different LLMs (Vicuna-v0-7B [12], Vicuna-v1.5-7B [12], and LLaMA-2-chat-7B [59]), as shown in Tab. 5. Pre-trained LLMs do not work well on IQA tasks, and need to be fine-tuned on our dataset. However, complete LLM fine-tuning is resource-intensive and data-hungry, requiring tens of thousands of high-quality human-written texts [43]. To get around of data shortage issue, we resort to LoRA [20], an efficient LLM fine-tuning technique, which selectively adjusts only a small set of parameters in the LLM. Following [20, 69], we apply LoRA to fine-tune the projection layers in all self-attention modules.

Unique tag to distinguish multiple images. Existing MLLMs [30, 82] are primarily designed to handle a single input image. They insert the visual tokens between the start (*e.g.*, `<Img>`) and end (*e.g.*, `</Img>`) tags. A simple extension to multi-image input is using textual cues to distinguish images, *e.g.*, adding "Reference Image:", "Image A:", and "Image B:" before visual tokens. However, this simple extension sometimes fails to distinguish images, probably because the proportion of these short textual cues in the full texts is too small. To mitigate this, motivated by [79], we adopt the unique tag for each image. In *comparison* and *reasoning* tasks, we select `<Img-Reference>`, `<Img-A>`, and `<Img-B>` as start tags, adhering to the HTML rule by adding a trailing slash ("/") in the end tags. In *description* task, the reference image retains the same tags, while the distorted image employs the original tags, `<Img></Img>`.

4.2 Training Scheme

Multi-source training data. As stated in Sec. 1, our training images come from three sets: (1) 115,646 brief templated data of Task–2 (*comparison*), (2) 4,854 high-quality data of Task–1 (*description*) and Task–3 (*reasoning*), which are duplicated by 20 times to increase the training weights, and (3) 48,734 content description data (Detailed Description dataset in [69], duplicated 2 times during training), which are unrelated to IQA. The abundant templated data principally bridge images and descriptive texts in quality-related tasks. The limited yet high-quality data foster the model's descriptive and reasoning skills. The IQA-unrelated content description data can serve as regularization, given the limited text diversity of our IQA data for an MLLM. Experimental results in Tab. 4 prove the effectiveness of the three sources of training data.

Training objective. Following existing works [6, 30, 69], the training objective of DepictQA is the next token prediction loss: a cross-entropy loss between predicted and ground-truth next tokens. Only the tokens in the responses contribute to the loss computation. Also, only the image projection layer and LoRA parameters are optimized, comprising a mere 0.25% of the total parameters (7B).

5 Experiments

This section discusses experimental setups and results. In LoRA, the rank and scale factor are both set as 16. In each attention layer of LLM, the projection weights of "query", "key", "value", and "output" are adjusted using two LoRA delta parameters. DepictQA is trained for 1 epoch with batch size 64 on 8 GPUs (NVIDIA RTX A6000 48G). Adam optimizer with $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.001, and learning rate $5e^{-4}$ is used. The training is completed in around 12 hours. See *Appendix* for details and the training and inference costs.

5.1 Metrics

Unlike traditional score-based IQA methods, evaluating the diverse, descriptive, and textual results of multi-modal IQA methods is not trivial. Therefore, we adopt three different kinds of metrics for a comprehensive evaluation.

Accuracy, SRCC, and PLCC. For *comparison* task, we employ the accuracy metric. MLLMs usually produce diverse textual outputs, necessitating transformation to bi-classification results (*i.e.*, Image A or Image B is better) for accuracy calculation. As described in the *Task–2* part in Sec. 3.2, we add the short answer prompt "Answer the question using a single word or phrase" during training, guiding DepictQA to directly output bi-classification results. To further evaluate the alignment between the comparison results and human perception, we conduct pair-wise comparison and adopt voting method to convert the bi-classification results to quality scores. The quality scores are assessed using Spearman Rank Correlation Coefficient (SRCC) and Pearson Linear Correlation Coefficient (PLCC) following score-based IQA methods [60, 77].

GPT-4 score. For *description* and *reasoning* tasks, following [12, 30], we utilize the GPT-4 score for assessment. Specifically, we provide GPT-4 with all information in the human-labeled questionnaire as context. Then, we give GPT-4 both the model-generated response and the corresponding ground truth response. Based on the context, GPT-4 evaluates the helpfulness, relevance, accuracy, and level of detail of these two responses, and gives an overall score on a scale of 0 to 10, where a higher score indicates better quality. Finally, the relative score with regard to the ground truth response is reported as the GPT-4 score.

Reasonable rate by human evaluators. We observe that GPT-4 score exhibits excessive confidence in some low-quality responses where a wrong response even receives a GPT-4 score over 60% (see *Appendix*). Thus, for a comprehensive evaluation, given both the images and responses, human evaluators label each response as reasonable or not. A reasonable response should adhere to three criteria: indicating the major one distortion, no severe comparison mistakes, and self-consistency (see *Appendix*). The reasonable rate serves as another metric.

5.2 Comparison with Score-based IQA Methods

To demonstrate the effectiveness of DepictQA, we compare it with score-based IQA methods on the *comparison* task (Task–2). We evaluate four traditional IQA methods including PSNR, SSIM [61], VIF [52], and FSIM [72], and four deep-learning-based IQA methods including DeepIQA [4], PieAPP [47], LPIPS [75], and DISTS [15]. Here our base LLM is selected as LLaMA-2-chat-7B [59].

Quantitative results of quality comparison on BAPPS are depicted in Tab. 2a. Our DepictQA surpasses the best traditional method, FSIM, by a large margin ($\sim$11%). We also stably outperform the best deep-learning-based competitors, LPIPS and DISTS (>2.6%), even near the human performance. We show in Tab. 5 that the accuracy could be further enhanced with pre-training initialization (but with near 3$\times$ training time). Unlike high-level perception tasks where multi-modal approaches usually lag behind single-modal methods [69], we

Table 2: Quantitative results of quality comparison task on traditional IQA benchmarks. Our DepictQA achieves the state-of-the-art performance.

(a) Accuracy in Traditional / CNN distortion categories on BAPPS benchmark [75].

Type	Method	Comparison
Oracle	Human	80.8 / 84.4
Traditional	PSNR	59.9 / 77.8
	SSIM [61]	60.3 / 79.1
	VIF [52]	55.6 / 74.4
	FSIM [72]	62.7 / 79.4
Learning	DeepIQA [4]	70.3 / 79.4
	PieAPP [47]	72.7 / 77.0
	LPIPS [75]	76.0 / 82.8
	DISTS [15]	77.2 / 82.2
DepictQA (Ours)		**80.3** / **84.2**

(b) SRCC / PLCC metrics on benchmark datasets including KADID [28], CSIQ [26], TID2013 [45], LIVE [51], LIVE-MD [22], MDID2013 [56].

Datasets	Methods		
	FSIM [72]	LPIPS [75]	DepictQA (Ours)
KADID	0.855/0.857	0.799/0.803	**0.939**/**0.944**
CSIQ	0.937/0.937	0.905/0.926	**0.940**/**0.954**
TID2013	0.841/0.875	0.798/0.851	**0.853**/**0.886**
LIVE	0.894/0.908	0.906/**0.921**	**0.913**/0.914
LIVE-MD	0.877/0.910	0.897/0.913	**0.905**/**0.928**
MDID2013	0.717/0.770	0.731/0.764	**0.820**/**0.835**

show that multi-modal IQA methods can surpass score-based counterparts in the quality comparison task. Nevertheless, the quantitative comparison is not the key issue, we pay more attention to the description and reasoning abilities.

Quantitative results of quality comparison on multiple IQA datasets are shown in Tab. 2b. In this experiment, we include additional training datasets including PIPAL [19], KADID [28], and CSIQ [26]. These datasets consist of reference images and corresponding distorted versions. To adapt them to *comparison* task, we reorganize the datasets by randomly selecting a reference image and its two distortions. The ground truth is established by comparing the MOS or DMOS of these two distortions. The questions and responses are constructed with pre-defined templates. During evaluation, besides these datasets' non-overlapped validation split, we also evaluate on unseen benchmarks including TID2013 [45], LIVE [51], LIVE-MD [22], MDID2013 [56]. We conduct pairwise comparison and adopt voting method to transform the bi-classification results to quality scores. SRCC and PLCC metrics are reported. DepictQA stably outperforms two baselines, validating its effectiveness across multiple datasets.

5.3 Comparison with General Multi-modal LLMs

We also compare DepictQA with general MLLMs on *description* (Task–1) and *reasoning* (Task–3) tasks. We also transform the reasoning responses to bi-classification results to calculate the comparison accuracy. We choose four MLLMs as baselines: LLaVA [30], LLaVA-1.5 [29], MiniGPT4 [82], and LAMM [69]. We provide explicit instructions to inform the MLLMs of the task definition.

Quantitative results are illustrated in Tab. 3. LLaVA and MiniGPT4 (LLaMA-2-chat) fail in IQA tasks, yielding either nearly identical results or irrelevant results across most samples. Other general MLLMs exhibit poor performance, indicating their inadequacy for IQA tasks. However, after fine-tuning on our M-BAPPS dataset, DepictQA achieves significantly improved performance.

Table 3: Comparison with general MLLMs. All LLMs have 7B parameters. "LLaMA-2c" means LLaMA-2-chat. Metric for comparison: accuracy. Since general MLLMs can produce responses without explicit comparison results, accuracy is reported with these responses included / excluded. Metric for description and reasoning: reasonable rate / GPT-4 score. General MLLMs are not capable of IQA tasks.

Method	LLaVA [30]	LLaVA-1.5 [29]	MiniGPT4 [82]	MiniGPT4 [82]	LAMM [69]	LAMM [69]	DepictQA (Ours)
LLM (7B)	LLaMA-2c	Vicuna-v1.5	Vicuna-v0	LLaMA-2c	Vicuna-v0	LLaMA-2c	Vicuna-v1.5
Description	fail	18.0 / 65.8	16.0 / 49.9	fail	12.0 / 62.5	8.0 / 57.4	**64.0** / **76.2**
Comparison	fail	43.0 / 50.6	38.0 / 46.3	fail	53.0 / 55.8	44.0 / 48.9	**82.0**
Reasoning	fail	7.0 / 63.9	1.0 / 42.4	fail	4.0 / 58.1	5.0 / 52.4	**53.0** / **76.4**

One possible reason for general MLLMs' inadequacy is that they are trained on single images, while our tasks require multi-image input. Therefore, we also test LLaVA-1.5 on *description* task without a reference (*i.e.*, single-image input). We then employ GPT-4 to verify whether ground-truth distortions are mentioned in the responses. The results (24%) indicate some improvements over the full-reference setting (18%). However, the overall performance remains sub-optimal. Qualitative results in *Appendix* show the unsatisfying results of general MLLMs.

Qualitative results. Three qualitative results of *reasoning* task are depicted in Fig. 1 and Fig. 3. More qualitative results and failure cases of *description*, *comparison*, and *reasoning* tasks are illustrated in *Appendix*.

5.4 Ablation Studies

We conduct extensive ablation studies to verify the effectiveness of our methodologies. The LLM used in ablation studies is Vicuna-v1.5-7B [12] if not specified.

Effects of multi-source training data are detailed in Tab. 4. (1) Task-2 data (*comparison*) remarkably improves comparison accuracy (#2 *vs.* #4). Though these texts are brief and templated, the abundant samples still help bridge the images and texts in quality-related tasks. (2) As evidenced by #1 *vs.* #4 and #3 *vs.* #5, Task-3 data (*reasoning*) is necessary for a robust reasoning ability. Task-1 data (*description*) boosts reasoning metric (#4 *vs.* #5), mainly by helping identify distortions. Therefore, the high-quality texts of Task–1 and Task–3 are necessary to depict the image quality with language. (3) However, as shown by #1 *vs.* #3 and #4 *vs.* #5, Task-1 data (*description*) harms the comparison performance because of the gap between the two tasks in prompts and the number of input images. (4) Fortunately, these negative effects can be eliminated by content description data (#5 *vs.* #6), which serves as an regularization in light of M-BAPPS's limited text variety for a huge MLLM. Additionally, the content description data helps enrich the text diversity, as shown in *Appendix*.

Unique tag effectively mitigates the confusion problem, detailed in Fig. 4. Confusion occurs when distortions in one image are mistakenly attributed to another. To quantify this, we manually review 50 responses and compute the confusion rate. With the unified tag, the model needs to distinguish images

Table 4: Ablation studies of multi-source training data. "Cont." means the content description data. Metric for comparison: accuracy within Traditional / CNN categories. Metric for reasoning: reasonable rate / GPT-4 score. Multi-source training data stably improves the performance.

#	Training Data Task2	Task1	Task3	Cont.	Comparison	Reasoning
1	✓				78.6/82.5	N/A
2			✓		66.4/65.4	31.0/74.2
3	✓	✓			76.3/80.5	N/A
4	✓		✓		79.7/83.3	41.0/74.3
5	✓	✓	✓		78.1/82.9	45.0/**77.2**
6	✓	✓	✓	✓	**80.0**/**83.8**	**53.0**/76.4

Table 5: Ablation studies of LLMs and initialization. "High-level" means the MLLM is pre-trained on high-level perception tasks. Metric for comparison: accuracy within Traditional / CNN categories. DepictQA is compatible with different LLMs. Pre-training on high-level perceptual tasks brings improvement.

#	LLM	Init.	Comparison
1	Vicuna-v0-7B	Random	79.5/83.7
2	Vicuna-v0-7B	High-level	**79.8**/**84.2**
3	Vicuna-v1.5-7B	Random	80.0/83.8
4	Vicuna-v1.5-7B	High-level	**81.2**/**85.3**
5	LLaMA-2-chat-7B	Random	**80.3**/84.2
6	LLaMA-2-chat-7B	High-level	80.1/**85.1**

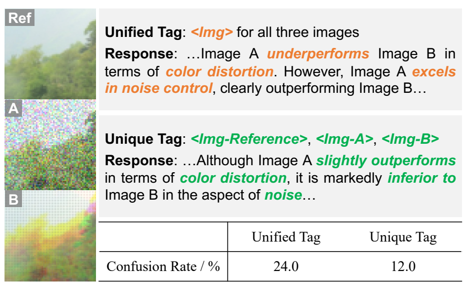

Fig. 4: Unique tag alleviates the confusion problem using clearer instructions. The confusion rate drops dramatically.

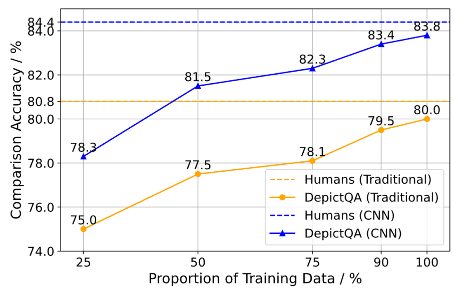

Fig. 5: The comparison performance gradually increases with the size of training data increasing.

through textual hinds, as stated in Sec. 4.1, leading to a 24% confusion rate. One unique tag for each image significantly reduces the confusion rate to 12%. Two alternative methods to distinguish multiple images (*i.e.*, image embedding, unique projector) are studied in *Appendix*, showing the advantage of unique tag.

The size of training data is studied in Fig. 5. As the size of training data increases, the comparison performance gradually increases. Thus the quantity of the training data still plays a key role in the MLLM-based method.

LLMs and the initialization methods are studied in Tab. 5. For high-level initialization, trainable parameters are pre-trained on high-level perception tasks (*e.g.*, description and captioning). (1) Overall, our DepictQA is compatible with different LLMs, stably yielding high performance. Also, the performance is slightly improved with more advanced LLMs (*e.g.*, Vicuna-v1.5 and LLaMA-2-chat). (2) With the same LLM, pre-training on high-level perceptual tasks further enhances the overall performance. Nevertheless, pre-training on high-level perceptual tasks will also increase the training time by nearly three times.

Table 6: Full-reference / non-reference performance. Metric for comparison: accuracy. Metric for description and reasoning: GPT-4 score.

Task	Performance
Description	76.4 / 84.5
Comparison	82.3 / 78.5
Reasoning	77.2 / 78.8

Table 7: Comparison with the proprietary GPT-4V [42]. Metric for comparison: accuracy. Metric for reasoning: reasonable rate / GPT-4 score.

Method	Comparison	Reasoning
GPT-4V [42]	65.0	**52.0** / **106.4**
DepictQA (ours)	**90.0**	**53.0** / 77.6

5.5 Extensions

Non-reference applications. Although our dataset is initially gathered in a full-reference setting, we demonstrate its utility in non-reference applications. We generate non-reference training data by simply removing the reference image in our original dataset. This non-reference data is combined to co-trained our model. The results are shown in Tab. 6. For the *description* task, non-reference performance stably surpasses full-reference, possibly because single-image tasks are easier than multi-image tasks. Our model keeps its high performance in the non-reference setting, proving the feasibility of the non-reference extension.

Comparison with GPT-4V [42] is conducted on the first 100 samples in the validation set of Task–3 (*reasoning*)[1]. We request GPT-4V to complete both *comparison* and *reasoning* tasks. GPT-4V is given detailed instructions and two examples for task description. As shown in Tab. 7, DepictQA and GPT-4V have mutual advantages and disadvantages. The GPT-4 score of GPT-4V is quite high, even higher than human-annotated responses (>100%), because of the linguistic fluency and detailed description of contents. However, we achieve comparable performance at the human-evaluated reasonable rate. Moreover, DepictQA is significantly better than GPT-4V in quality comparison. Note that GPT-4V is close-source and expensive to access, thus developing an effective MLLM for IQA is worthwhile. See *Appendix* for qualitative results and failure cases of GPT-4V.

6 Conclusions and Limitations

In this preliminary attempt, we propose an MLLM-based IQA model, DepictQA, demonstrating the possibility of depicting image quality with languages. There is still a long way to go for real-world application. (1) The amount and coverage of data are not sufficient, limiting the generalization. (2) The performance in *description* and *reasoning* tasks is not satisfying. (3) The distortion types can be more than five, even not pre-defined. Also, fine-grained comparisons on local details are preferred. (4) The voting method with pair-wise comparison is time-consuming. (5) Finally, whether MLLM-based IQA methods can take the place of score-based ones is still an open question. These belong to our future works.

[1] The results were tested on the preview web version released in Nov. 2023. A comprehensive evaluation of the stable API version, `gpt-4-turbo`, is given in [70].

Acknowledgement. This work was sponsored by CUHK Direct Grants (RCFUS) No. 4055189, National Natural Science Foundation of China (Grant No. 62276251), and the Joint Lab of CAS-HK.

References

1. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: Nocaps: Novel object captioning at scale. In: ICCV (2019)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. In: NeurIPS (2022)
3. Bell, S., Upchurch, P., Snavely, N., Bala, K.: Material recognition in the wild with the materials in context database. In: CVPR (2015)
4. Bosse, S., Maniry, D., Müller, K.R., Wiegand, T., Samek, W.: Deep neural networks for no-reference and full-reference image quality assessment. IEEE TIP (2017)
5. Bosse, S., Maniry, D., Müller, K.R., Wiegand, T., Samek, W.: Deep neural networks for no-reference and full-reference image quality assessment. IEEE TIP (2018)
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: NeurIPS (2020)
7. Cao, Y., Wan, Z., Ren, D., Yan, Z., Zuo, W.: Incorporating semi-supervised and positive-unlabeled learning for boosting full reference image quality assessment. In: CVPR (2022)
8. Chen, X., Li, Z., Pu, Y., Liu, Y., Zhou, J., Qiao, Y., Dong, C.: A comparative study of image restoration networks for general backbone network design. arXiv preprint arXiv:2310.11881 (2023)
9. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
10. Chen, Z., Zhang, Z., Li, H., Li, M., Chen, Y., Li, Q., Feng, H., Xu, Z., Chen, S.: Deep linear array pushbroom image restoration: A degradation pipeline and jitter-aware restoration network. In: AAAI (2024)
11. Chen, Z., He, Z., Lu, Z., Lu, Z.M.: Prompt-based test-time real image dehazing: a novel pipeline. arXiv preprint arXiv:2309.17389 (2023)
12. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://vicuna.lmsys.org (2023)
13. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In: NeurIPS (2023)
14. Ding, K., Liu, Y., Zou, X., Wang, S., Ma, K.: Locally adaptive structure and texture similarity for image quality assessment. In: ACM MM (2021)
15. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE TPAMI (2020)
16. Ghildyal, A., Liu, F.: Shift-tolerant perceptual similarity metric. In: ECCV (2022)
17. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In: CVPR (2017)

18. Gu, J., Cai, H., Chen, H., Ye, X., Ren, J., Dong, C.: Image quality assessment for perceptual image restoration: A new dataset, benchmark and metric. arXiv preprint arXiv:2011.15002 (2020)
19. Gu, J., Cai, H., Chen, H., Ye, X., Ren, J., Dong, C.: Pipal: a large-scale image quality assessment dataset for perceptual image restoration. arXiv preprint arXiv:2007.12142 (2020)
20. Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: ICLR (2021)
21. Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Liu, Q., et al.: Language is not all you need: Aligning perception with language models. arXiv preprint arXiv:2302.14045 (2023)
22. Jayaraman, D., Mittal, A., Moorthy, A.K., Bovik, A.C.: Objective quality assessment of multiply distorted images. In: Asilomar Conference on Signals, Systems and Computers (ASILOMAR) (2012)
23. Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no-reference image quality assessment. In: CVPR (2014)
24. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: CVPR (2021)
25. Kudo, T., Richardson, J.: Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: EMNLP (2018)
26. Larson, E.C., Chandler, D.M.: Most apparent distortion: full-reference image quality assessment and the role of strategy. Journal of Electronic Imaging (2010)
27. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
28. Lin, H., Hosu, V., Saupe, D.: Kadid-10k: A large-scale artificially distorted iqa database. In: International Conference on Quality of Multimedia Experience (QoMEX) (2019)
29. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
30. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
31. Liu, X., van de Weijer, J., Bagdanov, A.D.: RankIQA: Learning from rankings for no-reference image quality assessment. In: ICCV (2017)
32. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281 (2023)
33. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. In: NeurIPS (2022)
34. Ma, C., Yang, C.Y., Yang, X., Yang, M.H.: Learning a no-reference quality metric for single-image super-resolution. Computer Vision and Image Understanding (2017)
35. Masry, A., Do, X.L., Tan, J.Q., Joty, S., Hoque, E.: ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In: ACL Findings (2022)
36. Mathew, M., Karatzas, D., Jawahar, C.: DocVQA: A dataset for vqa on document images. In: WACV (2021)
37. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE TIP (2012)

38. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Sign. Process. Letters (2013)
39. Moorthy, A.K., Bovik, A.C.: A two-step framework for constructing blind image quality indices. IEEE Sign. Process. Letters (2010)
40. Moorthy, A.K., Bovik, A.C.: Blind image quality assessment: From natural scene statistics to perceptual quality. IEEE TIP (2011)
41. OpenAI: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
42. OpenAI: GPT-4V(ision) system card. https://openai.com/research/gpt-4v-system-card (2023)
43. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. In: NeurIPS (2022)
44. Pan, D., Shi, P., Hou, M., Ying, Z., Fu, S., Zhang, Y.: Blind predicting similar quality map for image quality assessment. In: CVPR (2018)
45. Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., et al.: Image database tid2013: Peculiarities, results and perspectives. Signal Processing: Image Communication (2015)
46. Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., Battisti, F.: Tid2008-a database for evaluation of full-reference visual quality assessment metrics. Advances of Modern Radioelectronics (2009)
47. Prashnani, E., Cai, H., Mostofi, Y., Sen, P.: Pieapp: Perceptual image-error assessment through pairwise preference. In: CVPR (2018)
48. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
49. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
50. Saad, M.A., Bovik, A.C., Charrier, C.: Blind image quality assessment: A natural scene statistics approach in the dct domain. IEEE TIP (2012)
51. Sheikh, H.: Live image quality assessment database release 2. http://live.ece.utexas.edu/research/quality (2005)
52. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE TIP (2006)
53. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards VQA models that can read. In: CVPR (2019)
54. Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., Zhang, Y.: Blindly assess image quality in the wild guided by a self-adaptive hyper network. In: CVPR (2020)
55. Sun, S., Yu, T., Xu, J., Zhou, W., Chen, Z.: GraphIQA: Learning distortion graph representations for blind image quality assessment. IEEE TMM (2022)
56. Sun, W., Zhou, F., Liao, Q.: Mdid: A multiply distorted image database for image quality assessment. PR (2017)
57. Tang, H., Joshi, N., Kapoor, A.: Learning a blind measure of perceptual image quality. In: CVPR (2011)
58. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
59. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
60. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: AAAI (2023)

61. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP (2004)
62. Wu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Wang, A., Li, C., Sun, W., Yan, Q., Zhai, G., et al.: Q-bench: A benchmark for general-purpose foundation models on low-level vision. arXiv preprint arXiv:2309.14181 (2023)
63. Wu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Wang, A., Xu, K., Li, C., Hou, J., Zhai, G., et al.: Q-instruct: Improving low-level visual abilities for multi-modality foundation models. arXiv preprint arXiv:2311.06783 (2023)
64. Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., et al.: Q-align: Teaching lmms for visual scoring via discrete text-defined levels. arXiv preprint arXiv:2312.17090 (2023)
65. Wu, H., Zhu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Li, C., Wang, A., Sun, W., Yan, Q., et al.: Towards open-ended visual quality comparison. arXiv preprint arXiv:2402.16641 (2024)
66. Wu, T., Ma, K., Liang, J., Yang, Y., Zhang, L.: A comprehensive study of multimodal large language models for image quality assessment. arXiv preprint arXiv:2403.10854 (2024)
67. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023)
68. Yin, G., Wang, W., Yuan, Z., Han, C., Ji, W., Sun, S., Wang, C.: Content-variant reference image quality assessment via knowledge distillation. In: AAAI (2022)
69. Yin, Z., Wang, J., Cao, J., Shi, Z., Liu, D., Li, M., Sheng, L., Bai, L., Huang, X., Wang, Z., et al.: Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. In: NeurIPS (2023)
70. You, Z., Gu, J., Li, Z., Cai, X., Zhu, K., Dong, C., Xue, T.: Descriptive image quality assessment in the wild. arXiv preprint arXiv:2405.18842 (2024)
71. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics (2014)
72. Zhang, L., Zhang, L., Mou, X., Zhang, D.: FSIM: A feature similarity index for image quality assessment. IEEE TIP (2011)
73. Zhang, P., Dong, X., Wang, B., Cao, Y., Xu, C., Ouyang, L., Zhao, Z., Ding, S., Zhang, S., Duan, H., Zhang, W., Yan, H., Zhang, X., Li, W., Li, J., Chen, K., He, C., Zhang, X., Qiao, Y., Lin, D., Wang, J.: Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. arXiv preprint arXiv:2309.15112 (2023)
74. Zhang, R., Han, J., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Gao, P., Qiao, Y.: Llama-adapter: Efficient fine-tuning of language models with zero-init attention. arXiv preprint arXiv:2303.16199 (2023)
75. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
76. Zhang, W., Li, D., Ma, C., Zhai, G., Yang, X., Ma, K.: Continual learning for blind image quality assessment. IEEE TPAMI (2022)
77. Zhang, W., Zhai, G., Wei, Y., Yang, X., Ma, K.: Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In: CVPR (2023)
78. Zhang, Z., Wu, H., Zhang, E., Zhai, G., Lin, W.: A benchmark for multi-modal foundation models on low-level vision: from single images to pairs. arXiv preprint arXiv:2402.07116 (2024)

79. Zhao, H., Cai, Z., Si, S., Ma, X., An, K., Chen, L., Liu, Z., Wang, S., Han, W., Chang, B.: Mmicl: Empowering vision-language model with multi-modal in-context learning. arXiv preprint arXiv:2309.07915 (2023)
80. Zheng, H., Fu, J., Zeng, Y., Zha, Z.J., Luo, J.: Learning conditional knowledge distillation for degraded-reference image quality assessment. ICCV (2021)
81. Zhou, W., Wang, Z.: Quality assessment of image super-resolution: Balancing deterministic and statistical fidelity. In: ACM MM (2022)
82. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
83. Zhu, H., Li, L., Wu, J., Dong, W., Shi, G.: MetaIQA: deep meta-learning for no-reference image quality assessment. In: CVPR (2020)